

**HANDBOOK  
OF  
AUTOMATION  
COMPUTATION  
AND  
CONTROL**

**VOLUME 1**

**HANDBOOK OF AUTOMATION,  
COMPUTATION, AND CONTROL**

**Volume 1**

---

**CONTROL FUNDAMENTALS**

**NEW YORK • JOHN WILEY & SONS, INC.**

London • Chapman & Hall, Limited

# HANDBOOK OF AUTOMATION, COMPUTATION, AND CONTROL

Volume 1

---

CONTROL FUNDAMENTALS

Prepared by a Staff of Specialists

Edited by

**EUGENE M. GRABBE**

**SIMON RAMO**

**DEAN E. WOOLDRIDGE**

The Ramo-Wooldridge Corporation

Los Angeles, California

Copyright © 1958, by John Wiley & Sons, Inc.

---

All Rights Reserved. This book or any part thereof must not be reproduced in any form without the written permission of the publisher.

Library of Congress Catalog Card Number: 58-10800  
Printed in the United States of America

# CONTRIBUTORS

- E. L. ARNOFF**, *Case Institute of Technology, Cleveland, Ohio (Chapter 15)*
- J. E. BARNES, Jr.**, *General Electric Company, Schenectady, New York (Chapter 26)*
- C. E. BRADFORD**, *General Electric Company, Pittsfield, Massachusetts (Chapter 22)*
- J. M. CAMERON**, *National Bureau of Standards, Washington, D. C. (Chapter 14)*
- R. F. CLIPPINGER**, *Datamatic, A Division of Minneapolis-Honeywell Regulator Company, Newton Highlands, Massachusetts (Co-Editor, Chapter 14)*
- A. B. CLARKE**, *University of Michigan, Ann Arbor, Michigan (Chapter 13)*
- A. H. COPELAND, SR.**, *University of Michigan, Ann Arbor, Michigan (Chapters 11 and 12)*
- P. G. CUSHMAN**, *General Electric Company, Pittsfield, Massachusetts (Chapter 23)*
- M. W. DE MERIT**, *General Electric Company, Schenectady, New York (Chapter 22)*
- J. B. DIAZ**, *University of Maryland, College Park, Maryland (Chapter 14)*
- B. DIMSDALE**, *Service Bureau Corporation, Los Angeles, California (Chapter 14)*
- P. ELIAS**, *Massachusetts Institute of Technology, Cambridge, Massachusetts (Chapter 16)*
- B. FRIEDMAN**, *University of California, Berkeley, California (Chapter 14)*
- W. M. GAINES**, *General Electric Company, Tempe, Arizona (Editor, Part E; Chapters 19 and 25)*
- G. E. HAY**, *University of Michigan, Ann Arbor, Michigan (Chapters 4 and 5)*

- E. ISAACSON**, *New York University, New York City, New York (Chapter 14)*
- S. J. JENNINGS**, *General Electric Company, Evendale, Ohio (Chapter 20)*
- W. KAPLAN**, *University of Michigan, Ann Arbor, Michigan (Co-Editor, Part A; Chapters 5, 7, 8, 9, and 10)*
- J. H. LEVIN**, *Datamatic, A Division of Minneapolis-Honeywell Regulator Company, Newton Highlands, Massachusetts (Co-Editor, Chapter 14)*
- D. L. LIPPITT**, *General Electric Company, Schenectady, New York (Chapter 24)*
- R. C. LYNDON**, *University of Michigan, Ann Arbor, Michigan (Chapters 2 and 3)*
- M. MANNOS**, *Datamatic, A Division of Minneapolis-Honeywell Regulator Company, Newton Highlands, Massachusetts (Chapter 14)*
- P. MERTZ**, *Bell Telephone Laboratories, New York City, New York (Chapters 17 and 18)*
- S. G. REQUE**, *General Electric Company, Tempe, Arizona (Chapter 21)*
- R. RICHTMEYER**, *New York University, New York City, New York (Chapter 14)*
- E. H. ROTHE**, *University of Michigan, Ann Arbor, Michigan (Chapter 6)*
- W. E. SOLLECITO**, *General Electric Company, Schenectady, New York (Chapter 21)*
- R. M. THRALL**, *University of Michigan, Ann Arbor, Michigan (Co-Editor, Part A; Chapter 1)*
- A. A. WINKELJOHANN**, *General Electric Company, Evendale, Ohio (Chapter 20)*

## FOREWORD

The proliferation of knowledge now makes it most difficult for scientists or engineers to keep ahead of change even in their own fields, let alone in contiguous fields. One of the fields where recent change has been most noticeable, and in fact exponential, has been automatic control. This three-volume Handbook will aid individuals in almost every branch of technology who must constantly refresh their memories or refurbish their knowledge about many aspects of their work.

Automation, computation, and control, as we know them, have been evolving for centuries, but within the last generation their impact has been felt in nearly every segment of human endeavor. Feedback principles were exploited by Leonardo da Vinci and applied by James Watt. Some of the early theoretical work of importance was contributed by Lord Kelvin, who also, together with Charles Babbage, pointed the way to the development of today's giant computational aids. Since about the turn of the present century, the works of men like Minorsky, Nyquist, Wiener, Bush, Hazen, and Von Neuman gave quantum jumps to computation and control. But it was during and immediately following World War II that quantum jumps occurred in abundance. This was the period when theories of control, new concepts of computation, new areas of application, and a host of new devices appeared with great rapidity. Technologists now find these fields charged with challenge, but at the same time hard to encompass. From the activities of World War II such terms as servomechanism, feedback control, digital and analog computer, transducer, and system engineering reached maturity. More recently the word automation has become deeply entrenched as meaning something about the field on which no two people agree.

Philosophically minded technologists do not accept automation merely as a third Industrial Revolution. They see it, as they stand about where the editors of this *Handbook* stood when they projected this work, as a manifestation of one of the greatest *Intellectual Revolutions in Thinking* that has occurred for a long time. They see in automation the natural consequence of man's urge to exploit modern science on a wide front to perform useful tasks in, for example, manufacturing, transportation, business, physical science, social science, medicine, the military, and government. They see that it has brought great change to our conventional way of thinking about the human use of human beings, to quote Norbert Wiener, and in turn about how our engineers will be trained to solve tomorrow's engineering problems. They even see that it has precipitated some deep thinking on the part of our indus-

trial and union leadership about the organization of workers in order not to hold captive bodies of workmen for jobs that automation, computation, and control have swept or will soon sweep away.

Perhaps the important new face on today's technological scene is the degree to which the broad field needs codification and unification in order that technologists can optimize their role to exploit it for the general good. One of the early instances of organized academic instruction in the field was at The Massachusetts Institute of Technology in the Electrical Engineering Department in September 1939, as a course entitled Theory and Application of Servomechanisms. I can well recollect discussions around 1940 with the late Dr. Donald P. Campbell and Dr. Harold L. Hazen, which led temporarily to renaming the course Dynamic Analysis of Automatic Control Systems because so few students knew what "servomechanisms" were. But when the GI's returned from war everybody knew, and everyone wanted instruction. Since that time engineering colleges throughout the land have elected to offer organized instruction in a multitude of topics ranging from the most abstract mathematical fundamentals to the most specific applications of hardware. Textbooks are available on every subject along this broad spectrum. But still the practicing control or computer technologist experiences great difficulty keeping abreast of what he needs to know.

As organized instruction appeared in educational institutions, and as industrial activity increased, professional societies organized groups in the areas of control and computation to meet the needs of their members to tell one another about technical advances. Within the past five years several trade journals have undertaken to report regularly on developments in theory, components, and systems. The net effect of all this is that the technologist is overwhelmed with fragmentary, sometimes contradictory, redundant information that comes at him at random and in many languages. The problem of assessing and codifying even a portion of this avalanche of knowledge is beyond the capabilities of even the most able technologist.

The editors of the *Handbook* have rightly concluded that what each technologist needs for his long term professional growth is *to have a body of knowledge that is negotiable at par in any one of a number of related fields for many years to come*. It would be ideal, of course, if a college education could give a prospective technologist this kind of knowledge. It is in the hope of doing this that engineering curricula are becoming more broadly based in science and engineering science. But it is unlikely that even this kind of college training will be adequate to cope with the consequences of the rapid proliferation of technology as is manifest in the area of automation, computation, and control. Hence, handbooks are an essential component of the technical literature when they provide the unity and continuity that are requisite.

I can think of no better way to describe this *Handbook* than to say that the editors, in both their organization of material and selection of substance, have given technologists a unified work of lasting value. It truly represents today's optimum package of that body of knowledge that will be negotiable at par by technologists for many years to come in a wide range of disciplines.

GORDON S. BROWN  
*Massachusetts Institute of Technology*

## PREFACE

Accelerated advances in technology have brought a steady stream of automatic machines to our factories, offices, and homes. The earliest automation forms were concerned with doing work, followed by the controlling function, and recently the big surge in automation has been directed toward data handling functions. New devices ranging from digital computers to satellites have resulted from military and other government research and development programs. Such activity will continue to have an important impact on automation progress.

One of the pressures for the development of automation has been the growing complexity and speed of business and industrial operations. But automation in turn accelerates the tempo of whatever it touches, so that we can expect future systems to be even larger, faster, and more complex. While a segment of engineering will continue to mastermind, by rule of thumb procedures, the design and construction of automatic equipment and systems, a growing percentage of engineering effort will be devoted to activities that may be classified as *problem solving*. The activities of the problem solver involve analysis of previous behavior of systems and equipment, simulation of present situations, and predictions about the future. In the past, problem solving has largely been practiced by engineers and scientists, using slide rules and hand calculators, but with the advent of large-scale data processing systems, the range of applications has been broadened considerably to include economic, government, and social activities. Air traffic control, traffic simulation, library searching, and language translation, are typical of the problems that have been attacked.

This *Handbook* is directed toward the problem solvers—the engineers, scientists, technicians, managers, and others from all walks of life who are concerned with applying technology to the mushrooming developments in automatic equipment and systems. It is our purpose to gather together in one place the available theory and information on general mathematics,

feedback control, computers, data processing, and systems design. The emphasis has been on practical methods of applying theory, new techniques and components, and the ever broadening role of the electronic computer. Each chapter starts with definitions and descriptions aimed at providing perspective and moves on to more complicated theory, analysis, and applications. In general, the *Handbook* assumes some engineering training and will serve as an information source and refresher for practicing engineers. For management, it will provide a frame of reference and background material for understanding modern techniques of importance to business and industry. To others engaged in various ramifications of automation systems, the *Handbook* will provide a source of definitions and descriptive material about new areas of technology.

It would be difficult for any one individual or small group of individuals to prepare a handbook of this type. A large number of contributors, each with a field of specialty, is required to provide the engineer with the desired coverage. With such a broad field, it is difficult to treat all material in a homogeneous manner. Topics in new fields are given in more detail than the older, established ones since there is a need for more background information on these new subjects. The organization of the material is in three volumes as shown on the inside cover of the *Handbook*. Volume 1 is on *Control Fundamentals*, Volume 2 is concerned with *Computers and Data Processing*, and Volume 3 with *Systems and Components*.

In keeping with the purpose of this *Handbook*, Volume 1 has a strong treatment of general mathematics which includes chapters on subjects not ordinarily found in engineering handbooks. These include sets and relations, Boolean algebra, probability, and statistics. Additional chapters are devoted to numerical analysis, operations research, and information theory. Finally, the present status of feedback control theory is summarized in eight chapters. Components have been placed with systems in Volume 3 rather than with control theory in Volume 1, although any discussion of feedback control must, of necessity, be concerned with components.

The importance of computing in research, development, production, real time process control, and business applications, has steadily increased. Hence, Volume 2 is devoted entirely to the design and use of analog and digital computers and data processors. In addition to covering the status of knowledge today in these fields, there are chapters on unusual computer systems, magnetic core and transistor circuits, and an advanced treatment of programming. Volume 3 emphasizes systems engineering. A part of the volume covers techniques used in important industrial applications by examining typical systems. The treatment of components is largely concerned with how to select components among the various alternates, their mathematical description and their integration into systems. There is also

a treatment of the design of components of considerable importance today. These include magnetic amplifiers, semiconductors, and gyroscopes.

We consider this *Handbook* a pioneering effort in a field that is steadily pushing back frontiers. It is our hope that these volumes will not only provide basic information on new fields, but will also inspire work and further research and development in the fields of automatic control. The editors are pleased to acknowledge the advice and assistance of Professor Gordon S. Brown and Professor Jerome S. Wiesner of the Massachusetts Institute of Technology, and Dr. Brockway McMillan of the Bell Telephone Laboratories, in organizing the subject matter. To the contributors goes the major credit for providing clear, thorough treatments of their subjects. The editors are deeply indebted to the large number of specialists in the control field who have aided and encouraged this undertaking by reviewing manuscripts and making valuable suggestions. Many members of the technical staff and secretarial staff of The Ramo-Wooldridge Corporation have been especially helpful in speeding the progress of the *Handbook*.

EUGENE M. GRABBE  
SIMON RAMO  
DEAN E. WOOLDRIDGE

*August 1958*



# CONTENTS

## A. GENERAL MATHEMATICS

<b>Chapter 1. Sets and Relations</b>	1-01
1. Sets	1-01
2. Relations	1-05
3. Functions	1-06
4. Binary Relations on a Set	1-07
5. Equivalence Relations	1-07
6. Operations	1-08
7. Order Relations	1-09
8. Sets of Points	1-10
References	1-11
<b>Chapter 2. Algebraic Equations</b>	2-01
1. Polynomials	2-01
2. Real Roots	2-03
3. Complex Roots	2-04
References	2-06
<b>Chapter 3. Matrix Theory</b>	3-01
1. Vector Spaces	3-01
2. Linear Transformations	3-03
3. Coordinates	3-04
4. Echelon Form	3-05
5. Rank, Inverses	3-07
6. Determinants, Adjoint	3-08
7. Equivalence	3-09
8. Similarity	3-10
9. Orthogonal and Symmetric Matrices	3-13
10. Systems of Linear Inequalities	3-14
References	3-17

<b>Chapter 4.</b>	<b>Finite Difference Equations</b>	4-01
	1. Definitions	4-01
	2. Linear Difference Equations	4-03
	3. Homogeneous Linear Equations with Constant Coefficients	4-04
	4. Nonhomogeneous Linear Equations with Constant Coefficients	4-05
	5. Linear Equations with Variable Coefficients	4-07
	References	4-08
<b>Chapter 5.</b>	<b>Differential Equations</b>	5-01
	1. Basic Concepts	5-01
	2. Equations of First Order and First Degree	5-02
	3. Linear Differential Equations	5-04
	4. Equations of First Order but not of First Degree	5-07
	5. Special Methods for Equations of Higher than First Order	5-09
	6. Solutions in Form of Power Series	5-10
	7. Simultaneous Linear Differential Equations	5-12
	8. Numerical Methods	5-14
	9. Graphical Methods—Phase Plane Analysis	5-15
	10. Partial Differential Equations	5-20
	References	5-22
<b>Chapter 6.</b>	<b>Integral Equations</b>	6-01
	1. Definitions and Main Problems	6-01
	2. Relation to Boundary Value Problems	6-03
	3. General Theorems	6-05
	4. Theorems on Eigenvalues	6-06
	5. The Expansion Theorem and Some of Its Consequences	6-07
	6. Variational Interpretation of Eigenvalue Problem	6-08
	7. Approximation Methods	6-10
	References	6-17
<b>Chapter 7.</b>	<b>Complex Variables</b>	7-01
	1. Functions of a Complex Variable	7-01

	2. Analytic Functions. Harmonic Functions	
	7-04	
	3. Integral Theorems	7-05
	4. Power Series. Laurent Series	7-08
	5. Zeros. Singularities. Residues. Argument Principle	7-11
	6. Analytic Continuation	7-16
	7. Riemann Surfaces	7-17
	8. Elliptic Functions	7-18
	9. Functions Defined by Linear Differential Equations	7-21
	10. Other Transcendental Functions	7-25
	References	7-28
<b>Chapter 8.</b>	<b>Operational Mathematics</b>	<b>8-01</b>
	1. Heaviside Operators	8-01
	2. Application to Differential Equations	8-05
	3. Superposition Principle. Response to Unit Function and Delta Function	8-06
	4. Appraisal of the Heaviside Calculus	8-07
	5. Operational Calculus Based on Integral Transforms	8-07
	6. Fourier Series. Finite Fourier Transform	8-10
	7. Fourier Integral. Fourier Transforms	8-15
	8. Laplace Transforms	8-17
	9. Other Transforms	8-18
	References	8-19
<b>Chapter 9.</b>	<b>Laplace Transforms</b>	<b>9-01</b>
	1. Fundamental Properties	9-01
	2. Transforms of Derivatives and Integrals	9-03
	3. Translation. Transform of Unit Function, Step Functions, Impulse Function (Delta Function)	9-06
	4. Convolution	9-08
	5. Inversion	9-09
	6. Application to Differential Equations	9-10
	7. Response to Impulse Functions	9-15
	8. Equations Containing Integrals	9-18
	9. Weighting Function	9-18
	10. Difference-Differential Equations	9-20

	11. Asymptotic Behavior of Transforms	9-21
	References	9-21
<b>Chapter 10.</b>	<b>Conformal Mapping</b>	10-01
	1. Definition of Conformal Mapping.	
	General Properties	10-01
	2. Linear Fractional Transformations	10-05
	3. Mapping by Elementary Functions	10-06
	4. Schwarz-Christoffel Mappings	10-08
	5. Application of Conformal Mapping to	
	Boundary Value Problems	10-09
	References	10-11
<b>Chapter 11.</b>	<b>Boolean Algebra</b>	11-01
	1. Table of Notations	11-01
	2. Definitions of Boolean Algebra	11-01
	3. Boolean Algebra and Logic	11-05
	4. Canonical Form of Boolean Functions	11-08
	5. Stone Representation	11-09
	6. Sheffer Stroke Operation	11-10
	References	11-11
<b>Chapter 12.</b>	<b>Probability</b>	12-01
	1. Fundamental Concepts and Related	
	Probabilities	12-01
	2. Random Variables and Distribution	
	Functions	12-04
	3. Expected Value	12-06
	4. Variance	12-11
	5. Central Limit Theorem	12-13
	6. Random Processes	12-18
	References	12-20
<b>Chapter 13.</b>	<b>Statistics</b>	13-01
	1. Nature of Statistics	13-01
	2. Probability Background	13-02
	3. Important Probability Distributions	13-04
	4. Sampling	13-06
	5. Bivariate Distributions	13-13
	6. Tests for Goodness of Fit	13-16
	7. Sequential Analysis	13-16
	8. Monte Carlo Method	13-17
	9. Statistical Tables	13-18
	References	13-21

**B. NUMERICAL ANALYSIS**

<b>Chapter 14. Numerical Analysis . . . . .</b>	<b>14-01</b>
1. Interpolation, Curve Fitting, Differentiation, and Integration	14-01
2. Matrix Inversion and Simultaneous Linear Equations	14-13
3. Eigenvalues and Eigenvectors	14-28
4. Digital Techniques in Statistical Analysis of Experiments	14-48
5. Ordinary Differential Equations	14-55
6. Partial Differential Equations	14-64
References	14-88

**C. OPERATIONS RESEARCH**

<b>Chapter 15. Operations Research . . . . .</b>	<b>15-01</b>
1. Operations Research and Mathematical Models	15-02
2. Solution of the Model	15-10
3. Inventory Models	15-21
4. Allocation Models	15-31
5. Waiting Time Models	15-73
6. Replacement Models	15-86
7. Competitive Problems	15-99
8. Data for Model Testing	15-115
9. Controlling the Solution	15-120
10. Implementation	15-123
References	15-124

**D. INFORMATION THEORY AND TRANSMISSION**

<b>Chapter 16. Information Theory . . . . .</b>	<b>16-01</b>
1. Introduction	16-01
2. General Definitions	16-02
3. Simple Discrete Sources	16-08
4. More Complicated Discrete Sources	16-19
5. Discrete Noiseless Channels	16-24
6. Discrete Noisy Channels I. Distribution of Information	16-26
7. Discrete Noisy Channels II. Channel Capacity and Interpretations	16-32
8. The Continuous Case	16-39
References	16-46

<b>Chapter 17. Smoothing and Filtering</b>	17-01
1. Definitions: Smoothing and Prediction. Symbols	17-01
2. Definitions: Correlation	17-05
3. Relationship between Correlation and Signal Structure	17-09
4. Design of Optimum Filter	17-13
5. Extensions of Procedure	17-19
6. Network Synthesis	17-25
References	17-32
<b>Chapter 18. Data Transmission</b>	18-01
1. Introduction and Symbols	18-01
2. Formation and Use of the Electrical Signal	18-07
3. Transmission Impairment	18-18
References	18-30
<b>E. FEEDBACK CONTROL</b>	
<b>Chapter 19. Methodology of Feedback Control</b>	19-01
1. Symbols for Feedback Control	19-01
2. General Feedback Control System Definitions	19-04
3. Feedback Control System Design Considerations	19-12
4. Selection of Method of Synthesis for Feedback Controls	19-19
References	19-21
<b>Chapter 20. Fundamentals of System Analysis</b>	20-01
1. Representation of Physical Systems	20-01
2. Classical Methods of Analysis	20-28
3. Block Diagrams	20-56
4. System Types	20-66
5. Error Coefficients	20-70
6. Analysis of A-C Servos: Carrier Systems	20-79
References	20-84
<b>Chapter 21. Stability</b>	21-01
1. Introduction	21-01
2. Classical Solution Approach	21-02

- 3. Routh's Criterion 21-05
- 4. Nyquist Stability Criterion 21-09
- 5. Bode Attenuation Diagram Approach 21-29
- 6. Root Locus Method 21-46
- 7. Miscellaneous Stability Criteria 21-71
- 8. Closed Loop Response from Open Loop Response 21-72
- References 21-81

**Chapter 22. Relation between Transient and Frequency Response . . . . . 22-01**

- 1. Introduction 22-01
- 2. Response Characteristics Defined 22-02
- 3. Relation between Transient Response and Location of Roots of Characteristic Equation 22-03
- 4. Relation between Closed Loop and Open Loop Roots 22-15
- 5. Design Charts Relating Open Loop Frequency Response and Transient Response 22-18
- 6. Approximate Relations—Rules of Thumb 22-43
- 7. Numerical and Graphical Techniques of Relating Transient and Frequency Response 22-43
- References 22-61

**Chapter 23. Feedback System Compensation . . . . . 23-01**

- 1. Design Criteria and Techniques 23-01
- 2. Compensating Components: D-C Systems 23-18
- 3. Compensating Networks: A-C Systems 23-48
- 4. Open-Closed Loop Control 23-54
- References 23-56

**Chapter 24. Noise, Random Inputs, and Extraneous Signals . . . . . 24-01**

- 1. Introduction 24-01
- 2. Mathematical Description of Noise 24-02
- 3. Measurement of Noise 24-06
- 4. System Response to Noise 24-11

## CONTENTS

	5. System Design in the Presence of Noise	
	24-15	
	References 24-19	
<b>Chapter 25.</b>	<b>Nonlinear Systems</b>	<b>25-01</b>
	1. Definitions	25-01
	2. General Nonlinear System Problem	25-03
	3. Methods of Analysis: Linearization	25-07
	4. Methods of Analysis: Describing Function	25-13
	5. Methods of Analysis: Phase Plane, Graphical	
	Solution of System Equations	25-36
	6. Other Methods of Analysis	25-43
	7. Nonlinear System Compensation	25-48
	References	25-66
<b>Chapter 26.</b>	<b>Sampled-Data Systems and</b>	
	<b>Periodic Controllers</b>	<b>26-01</b>
	1. Description and Definition of Sampled-Data	
	System	26-01
	2. Methods of Transient Analysis	26-06
	3. Sampled-Data System Stability	26-15
	4. Sampled-Data System Synthesis	26-20
	References	26-32

## INDEX

# MATHEMATICS

## A. GENERAL MATHEMATICS

R. M. Thrall and  
W. Kaplan, Editors

1. Sets and Relations, by R. M. Thrall
2. Algebraic Equations, by R. C. Lyndon
3. Matrix Theory, by R. C. Lyndon
4. Finite Difference Equations, by G. E. Hay
5. Differential Equations, by G. E. Hay and W. Kaplan
6. Integral Equations, by E. H. Rothe
7. Complex Variables, by W. Kaplan
8. Operational Mathematics, by W. Kaplan
9. Laplace Transforms, by W. Kaplan
10. Conformal Mapping, by W. Kaplan
11. Boolean Algebra, by A. H. Copeland, Sr.
12. Probability, by A. H. Copeland, Sr.
13. Statistics, by A. B. Clarke



## Sets and Relations

R. M. Thrall

1. Sets	1-01
2. Relations	1-05
3. Functions	1-06
4. Binary Relations on a Set	1-07
5. Equivalence Relations	1-07
6. Operations	1-08
7. Order Relations	1-09
8. Sets of Points	1-10
References	1-11

### 1. SETS

A *set* is a collection of objects of any sort. The words *class*, *family*, *ensemble*, *aggregate* are synonyms for the term *set*.

Each object in a set is called an *element* (*member*) of the set. If  $S$  denotes the set and  $b$  an element of  $S$ , one writes:

$$b \in S;$$

this is read: " $b$  belongs to  $S$ ." If  $b$  does not belong to  $S$ , one writes:  $b \notin S$ .

Sets will generally be designated by capital letters, elements by lower case letters.

**IMPORTANT EXAMPLES OF SETS.**  $Z$ , the set of *positive integers*  $z$ ;  $Z$  consists of the numbers 1, 2, 3,  $\dots$ ;

$J$ , the set of all *integers*  $j$  (including 0 and the negative integers);

$Q$ , the set of *rational numbers*  $q$  (fractions  $a/b$ , where  $a$  is an integer and  $b$  is a positive integer);

$R$ , the set of all *real numbers*  $r$  (numbers which are expressible as unending decimals);

$C$ , the set of all *complex numbers*  $c$  (numbers of form  $x + y\sqrt{-1}$ , where  $x$  and  $y$  are real).

In geometry one employs *sets of points*; for example, all points on a specified line or all points inside a circle.

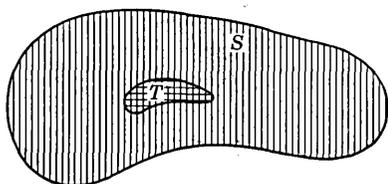


FIG. 1. Set and subset.  $T$  is a subset of  $S$ ,  $T \subset S$ .

Geometric diagrams can be helpful in reasoning about sets which may have no reference to geometry (Fig. 1).

A set can be designated by listing its elements between braces. Thus

$$S = \{1, -3, 7\}$$

is the set whose elements are the numbers 1, -3, and 7. For infinite sets

one still uses braces, but instead of writing all the elements one gives a rule for set membership. For *example*,  $Z = \{z | z \text{ is a positive integer}\}$  is an abbreviation for "Z is the set of all  $z$  for which  $z$  is a positive integer." This set is also sometimes designated (less precisely) by

$$Z = \{1, 2, 3, \dots, n, \dots\}.$$

Two sets are said to be equal if they have exactly the same members. For *example*, if

$$A = \{1, 3, 5, 7\}, \quad B = \{3, 1, 5, 1, 7\},$$

then  $A = B$ . Neither the order in which the elements are written down nor the number of times that an element is repeated within the braces is significant. If  $A$  is not equal to  $B$ , one writes:  $A \neq B$ .

**Subsets.** The set  $T$  is said to be a *subset* of the set  $S$  (Fig. 1) if every member of  $T$  is also a member of  $S$ ; in symbols,

$$T \subset S \quad \text{or} \quad S \supset T.$$

If both  $T \subset S$  and  $S \subset T$ ,  $S$  and  $T$  are equal.

If  $S \supset T$  and  $S \neq T$ , i.e., if  $S$  contains every element of  $T$  and at least one element not in  $T$ , one says that  $S$  *contains*  $T$  *properly* or that  $T$  is a *proper subset* of  $S$ ; in symbols,

$$S > T \quad \text{or} \quad T < S.$$

The symbol

$$T = \{t | t \in S \text{ and } t \text{ has properties } P, Q\}$$

indicates that  $T$  is the subset of  $S$  consisting of precisely those elements  $t$  of  $S$  which have the properties  $P$  and  $Q$ .

**EXAMPLE.** If  $S$  is a set of people, one might consider the subset  $T$  of  $S$  consisting of those members of  $S$  who are college graduates and are married:

$$T = \{t | t \in S \text{ and } t \text{ is a college graduate and } t \text{ is married}\}.$$

Another *example*. If  $R$  is the set of all real numbers, then

$$T = \{t | t \in R \text{ and } 1 \leq t \leq 3\} \text{ is called a } \textit{closed interval}.$$

The symbol  $\emptyset$  is used to designate the *empty* (void) set, which has no elements;  $\emptyset$  is a subset of every set.

The collection of all subsets of a given set  $S$  is called the *power set* of  $S$  and is denoted by  $2^S$ . Thus

$$2^S = \{U | U \text{ is a subset of } S\}.$$

The notation is suggested by the fact that if  $S$  is a finite set consisting of  $n$  elements, then  $2^S$  has  $2^n$  elements.

**Difference, Complement, Union and Intersection.** If  $S$  and  $T$  are sets, the *difference*  $S - T$  is defined to be the set consisting of those elements of  $S$  which are not in  $T$ , thus

$$S - T = \{s | s \in S \text{ and } s \notin T\}.$$

If  $T$  is a subset of  $S$ , then  $S - T$  is called the *complement* of  $T$  in  $S$  and is denoted by  $C_s T$  or even by  $CT$  if there can be no doubt about what  $S$  is.

The *union* of  $S$  and  $T$ , denoted by  $S \cup T$ , is the set consisting of those elements in  $S$  or in  $T$  (or in both).

The set  $(S - T) \cup (T - S)$  is termed the *symmetric difference* of  $S$  and  $T$  and is commonly denoted by  $S \oplus T$ . Thus

$$S \oplus T = \{s | s \in S \text{ and } s \notin T\} \cup \{s | s \in T \text{ and } s \notin S\}.$$

The *intersection* of  $S$  and  $T$ , denoted as  $S \cap T$ , is the set consisting of those elements common to  $S$  and  $T$ . Thus

$$S \cap T = \{y | y \in S \text{ and } y \in T\}.$$

Difference, symmetric difference, complement, union, and intersection are illustrated in Fig. 2.

The *union* of a finite number of sets  $S_1, S_2, \dots, S_n$  is defined as the set consisting of all elements belonging to at least one of the sets  $S_1, \dots, S_n$ . In symbols,

$$\bigcup_{\alpha=1}^n S_\alpha = \{t | \text{there is at least one } S_\alpha \text{ for which } t \in S_\alpha\}.$$

If one denotes by  $A = \{1, 2, \dots, n\}$  the set over which the index  $\alpha$  varies, one can also denote this set by  $\bigcup_{\alpha \in A} S_\alpha$ . The same notation can be used when  $A$  is an arbitrary index set, not necessarily finite.

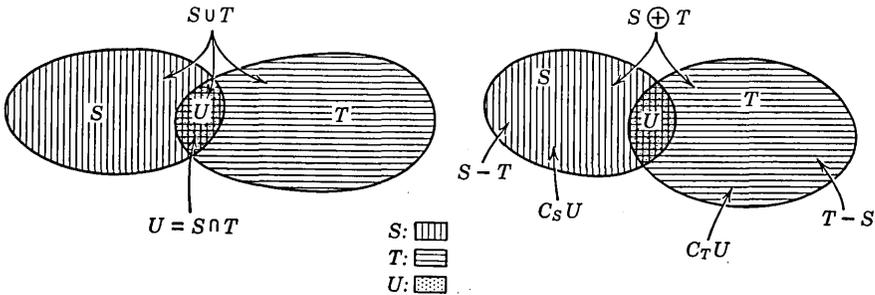


FIG. 2. Set operations.

The *intersection* of  $S_1, S_2, \dots, S_n$  is defined as the set consisting of all elements common to all of  $S_1, S_2, \dots, S_n$ . In symbols,

$$\bigcap_{\alpha=1}^n S_\alpha = \{t | t \in S_\alpha \text{ for } \alpha = 1, 2, \dots, n\} = \bigcap_{\alpha \in A} S_\alpha.$$

If one restricts attention to subsets of one fixed set  $U$ , and complements are taken with respect to  $U$ , then there is a *duality* between union and intersection:

$$C(S \cup T) = CS \cap CT,$$

$$C(S \cap T) = CS \cup CT.$$

Accordingly, if all sets are replaced by their complements, all unions are replaced by intersections, all intersections by unions.

Union, intersection, and complement satisfy several other laws, of which the following are typical:

$$A \cup B = B \cup A, \quad A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

These come under the theory of *Boolean algebra* (Sec. 7 and Chap. 11).

**Cartesian Product.** The *Cartesian product*  $S \times T$  of two sets  $S$  and  $T$  is the set of all ordered pairs  $(s, t)$ , where  $s \in S$  and  $t \in T$ . Thus, if

$$S = \{a, b\}, \quad T = \{1, 2, 3\},$$

then  $S \times T$  consists of the 6 pairs

$$(a, 1), (a, 2), (a, 3), (b, 1), (b, 2), (b, 3).$$

If  $R$  denotes the set of all real numbers, then  $R \times R$  consists of all ordered pairs  $(x, y)$ , where  $x$  and  $y$  are real; thus  $R \times R$  can be represented by the points  $(x, y)$  of the  $xy$ -plane of analytic geometry.

The Cartesian product of sets  $S_1, \dots, S_n$  in the order given is the set

$$\prod_{i=1}^n S_i$$

of all ordered  $n$ -tuples (vectors)  $(S_1, \dots, S_n)$ , where  $s_i \in S_i$  ( $i = 1, 2, \dots, n$ ). If  $S_1, \dots, S_n, \dots$  is an infinite sequence of sets, one denotes the Cartesian product by

$$\prod_{i=1}^{\infty} S_i;$$

this is the set of all infinite sequences  $(s_1, \dots, s_n, \dots)$ , where  $s_i \in S_i$  ( $i = 1, \dots, n, \dots$ ).

## 2. RELATIONS

Let two sets  $A, B$  be given. A *binary relation*  $R$  is said to be given between elements  $a$  of  $A$  and  $b$  of  $B$  if, for certain pairs  $(a, b)$ , the relation  $R$  is valid, while for all others it is invalid. For example, let  $A = \{1, 2, 3\}$ ,  $B = \{4, 6, 8\}$  and let the relation  $R$  between  $a$  and  $b$  be the condition that  $a$  equal  $\frac{1}{2}b$ . The relation is valid for the pairs

$$(2, 4), (3, 6)$$

and for no others.

In general, *specification* of a relation  $R$  is equivalent to selection of the pairs  $(a, b)$  for which  $R$  is to be valid. This set of pairs is a subset of the Cartesian product  $A \times B$ . Hence a relation  $R$  is equivalent to selection of a subset of  $A \times B$ . This subset can also be denoted by  $R$ .

If  $(a, b)$  is a pair for which the relation  $R$  is valid, one can write:  $aRb$ . In particular cases, other symbols can be used. In the example given above, one can write:  $a = \frac{1}{2}b$ .

When  $A = B$ , the relation  $R$  holds for certain pairs  $(a, b)$  of elements of  $A$ . One says:  $R$  is a *binary relation in*  $A$ . If  $A$  is the set of positive integers,  $R$  can be chosen as the relation of "being less than"; in symbols,  $aRb$  becomes  $a < b$ . One can then interpret  $R$  as the set of all pairs  $(a, b)$  of positive integers for which  $a < b$ . Thus, instead of writing  $2 < 3$ , one can write:  $(2, 3) \in R$ , to indicate that the relation is valid; similarly,  $(4, 3) \notin R$ , since  $4 < 3$  is false.

Let  $R$  be a relation between elements  $a$  of  $A$  and  $b$  of  $B$ , so that  $aRb$  for certain pairs  $(a, b)$ . Whenever  $aRb$ , one writes  $bR^T a$  and calls  $R^T$  the

transpose of the relation  $R$ . For example, the transpose of the relation  $<$  is the relation  $>$ ; the transpose of the relation "is the wife of" is the relation "is the husband of."

**Range and Domain.** Let  $R$  be a relation between elements  $a$  of  $A$  and  $b$  of  $B$ . For each  $a$  in  $A$  one denotes by  $R(a)$  the set of all  $b$  in  $B$  for which  $aRb$ ;  $R(a)$  is called the *image* of  $a$  under  $R$ . For each subset  $S$  of  $A$  one denotes by  $R(S)$  the set of all  $b$  in  $B$  for which  $aRb$  for at least one  $a$  in  $S$ ;  $R(S)$  is called the *image* of  $S$  under  $R$ . The image of  $A$  under  $R$ , the set  $R(A)$ , is called the *range* of  $R$ .

The *counterimage* of an element  $b$  under  $R$  is the set of all elements  $a$  for which  $aRb$ ; this is the same as the set  $R^T(b)$ . The *counterimage* of a subset  $U$  of  $B$  is the set of all  $a$  in  $A$  for which  $aRb$  for at least one  $b$  in  $U$ ; this is the same as the set  $R^T(U)$ . The set  $R^T(B)$  is called the *domain* of  $R$ .

A relation is sometimes called a *correspondence* between its domain and its range.

**Product of Relations.** If  $R$  is a relation between elements  $a$  of  $A$  and  $b$  of  $B$  and  $S$  is a relation between elements  $b$  of  $B$  and elements  $c$  of  $C$ , the *product* relation (or composition)  $RS$  is defined as a relation between elements  $a$  of  $A$  and  $c$  of  $C$ , as follows:  $aRSc$  whenever for some  $b$  in  $B$  one has  $aRb$  and  $bSc$ .

EXAMPLE. The product can be illustrated by a communications network. Let  $A, B, C$  be three sets of stations, let  $aRb$  mean that  $a$  can communicate with  $b$  and let  $bSc$  mean that  $b$  can communicate with  $c$ . Then  $aRSc$  means that there is a two-stage communication link from  $a$  through some intermediate station  $b$  to  $c$ . For products of three relations one has the associative law  $A(BC) = (AB)C$ .

### 3. FUNCTIONS

A relation  $F$  between elements  $a$  of  $A$  and  $b$  of  $B$  is said to be a *function* if, for every  $a$  in  $A$ ,  $F(a)$  is either empty or contains just one element. If in addition  $F$  has domain  $A$ , one says that  $F$  is a *function on  $A$  into  $B$* . In this case one identifies  $F(a)$  with its unique element  $b$  and writes:  $b = F(a)$   
or

$$a \xrightarrow{F} b$$

The terms *mapping* and *transformation* are synonyms for function. A function  $F$  on  $A$  into  $B$  can be defined directly as a correspondence which assigns to each element  $a$  in  $A$  a unique image  $b = F(a)$  in  $B$ . The set  $B$  is called a *codomain* of  $F$ . The range  $F(A)$  is a subset of  $B$ . If  $F(A) = B$ , one says that  $F$  is a *function on  $A$  onto  $B$* .

If both  $F$  and  $F^T$  are functions, one says that  $F$  is a *one-to-one* function.

The transpose  $F$  is then termed the *inverse* of  $F$  and is denoted by  $F^{-1}$ . A one-to-one function on  $A$  onto  $B$  is called a *one-to-one correspondence between  $A$  and  $B$* . In this case one has

$$(FF^T)a = F^T(F(a)) = a \text{ for all } a \text{ in } A;$$

$FF^T$  is the *identity function*  $E_A$  on  $A$  onto  $A$ :  $E_A(a) = a$ ; similarly,  $F^TF = E_B$ .

In classical analysis, a function  $F$  is often denoted by  $F(x)$ . The symbol  $F(x)$  then has two meanings: the value of the function for a particular  $x$ , and the function as a whole. Similarly,  $F(x, y)$  denotes a function of two variables and also the value of the function for given  $x, y$ .

#### 4. BINARY RELATIONS ON A SET

A relation  $R$  on a set  $A$  is said to be

*Identical* if  $R = E_A$ , i.e., if  $aRb$  is equivalent to  $a = b$ ;

*Reflexive* if  $R \supset E_A$ , i.e., if  $aRa$  for all  $a$  in  $A$ ;

*Irreflexive* if  $R \cap E_A = \emptyset$ , i.e., if  $aRa$  for no  $a$  in  $A$ ;

*Transitive* if  $R^2 \subset R$ , i.e., if  $aRb$  and  $bRc$  imply  $aRc$ ;

*Symmetric* if  $R = R^T$ , i.e., if  $aRb$  implies  $bRa$ ;

*Antisymmetric* if  $R \cap R^T \subset E_A$ , i.e., if  $aRb$  and  $bRa$  imply  $a = b$ ;

*Asymmetric* if  $R \cap R^T = \emptyset$ , i.e., if for no  $a, b$  is  $aRb$  and  $bRa$ ;

*Acyclic* if  $R^n \cap E_A = \emptyset$  for all  $n$ , i.e., if  $a_1Ra_2, a_2Ra_3, \dots, a_{n-1}Ra_n$  imply  $a_1 \neq a_n$ ;

*Complete* if  $R \cup R^T = A \times A$ , i.e., if for each pair  $(a, b)$  either  $aRb$  or  $bRa$ ;

*Trichotomous* if  $R \cup R^T \cup E_A = A \times A$  and  $R \cap R^T = \emptyset$ , i.e., if for each pair  $(a, b)$  exactly one of the relations  $aRb, bRa, a = b$  holds.

Note that a relation is asymmetric if and only if it is antisymmetric and irreflexive.

**EXAMPLES.** The parallel relation on lines in a plane is symmetric, irreflexive but not transitive (unless a line is defined to be parallel to itself). The relation  $<$  on the real numbers is asymmetric, transitive, trichotomous, and acyclic; whereas the relation  $\leq$  is antisymmetric, reflexive, transitive, and complete. The relation "is at least as good as" is reflexive, transitive, but not antisymmetric if there are two objects judged equally good.

#### 5. EQUIVALENCE RELATIONS

By a *partition* of a set  $A$  is meant a subdivision of  $A$  into subsets, no two of which have an element in common. By an *equivalence relation*  $R$  in  $A$  is meant a relation  $R$  in  $A$  which is reflexive, symmetric, and transitive.

Each partition of  $A$  determines an equivalence relation in  $A$ ;  $aRb$  holds when  $a$  and  $b$  are in the same subset of the partition. Conversely, each equivalence relation determines a partition of  $A$ ; the subsets of the partition are the sets  $R(a)$ , i.e., the sets of form  $\{b | bRa\}$ . The sets  $R(a)$  are called *equivalence classes*.

Equivalence is the basis of classification; the equivalence classes contain elements which, although not identical, can be regarded as alike or interchangeable for some purpose. *Example.* The sorting of nuts and bolts is based on the equivalence relation "has the same size and shape as." A property shared by all elements of each equivalence class is called an *invariant*. More formally, let  $R$  be an equivalence relation on a set  $A$ . A function  $F$  on  $A$  is said to be an *invariant relative to  $R$*  if  $aRb$  implies  $F(a) = F(b)$ . For example, if  $R$  is the relation of congruence on a set  $A$  of triangles, then the function  $F(a) = \text{area of triangle } a$  is an invariant.

A set of invariants  $F, G, \dots$  relative to a relation  $R$  is said to be *complete* if

$$F(a) = F(b), G(a) = G(b), \dots \text{ together imply } aRb.$$

The language "a necessary and sufficient for  $K$  is that  $P_1, P_2, \dots$  all hold" frequently states that  $P_1, P_2, \dots$  are a complete set of invariants for an equivalence relation associated with  $K$ . Many of the theorems of elementary geometry fall in this class.

One sometimes is interested in choosing from each equivalence class a representative from which one or more invariants can be easily calculated. Such representatives are said to be in *normal form* or *standard form*. More technically, let  $R$  be an equivalence relation on a set  $A$ . A function which assigns to each equivalence class  $R(a)$  one of its members is called a *canonical form* relative to  $R$ . Thus in matrix theory (Chap. 3) one has canonical forms for row equivalence, equivalence, congruence, orthogonal congruence, and similarity. It is customary to select a representative which displays a complete set of invariants.

## 6. OPERATIONS

A function  $F$  assigning to each ordered pair  $(a, b)$ , with  $a$  in  $A$  and  $b$  in  $B$ , an element  $c$  of set  $C$  is called a (binary) *operation* on  $A \times B$ . If  $F(a, b) = c$ , one also writes  $aFb = c$ . If  $A = B$ ,  $F$  is called an *operation on  $A$* . If also  $C = A$ ,  $F$  is called an *interior* operation; otherwise  $F$  is called an *exterior* operation.

**EXAMPLES.** Addition and multiplication of numbers are interior operations. The scalar product of two vectors is an exterior operation.

Let  $F$  be an interior operation on  $A$  and let  $R$  be an equivalence relation

on  $A$ . One says that  $F$  has the *substitution property* relative to  $R$ , if  $aRa'$  and  $bRb'$  imply  $(aFb)R(a'Fb')$ . *Example.* Let  $A$  be the set of all integers; let  $aRb$  mean that  $a$  and  $b$  have the same parity (both even or both odd). Then  $aRa'$  and  $bRb'$  imply that  $(a + b)R(a' + b')$ . Thus addition has the substitution property relative to  $R$ .

An exterior operation is said to have the substitution property relative to  $R$  if  $aRa'$  and  $bRb'$  imply  $(aFb) = (a'Fb')$ .

## 7. ORDER RELATIONS

A relation  $\leq$  on a set  $A$  is said to be a *partial order* if

- (i)  $a \leq a$  (*reflexivity*),
- (ii)  $a \leq b$  and  $b \leq c$  imply  $a \leq c$  (*transitivity*),
- (iii)  $a \leq b$  and  $b \leq a$  imply  $a = b$  (*antisymmetry*).

If  $a \leq b$  and  $a \neq b$ , one writes:  $a < b$ . The relation  $<$  is then asymmetric:

- (iv) for no  $a, b$  is  $a < b$  and  $b < a$ ;

it is also transitive. If  $a \leq b$  and  $b < c$ , one writes:  $b \geq a$  and  $c > b$  (transposition). An element  $a$  of  $A$  is said to be an *upper bound* for the subset  $B$  of  $A$  if  $b \leq a$  for all  $b$  in  $B$ ; if also  $a \leq c$  for every upper bound  $c$  of  $B$ , one says that  $a$  is a *least upper bound* (l.u.b.) for  $B$ . An upper bound for  $B$  which belongs to  $B$  is called a *maximal element* of  $B$ .

If in these definitions one replaces  $\leq$  by  $\geq$ , the resulting concepts are called *lower bound*, *greatest lower bound* (g.l.b.) and *minimal element*, respectively.

The least upper bound (greatest lower bound) of a set, if it exists, is unique.

The partial order is said to be a *linear order* or *chain order*, if it is complete:

- (v) for every  $a, b$  either  $a \leq b$  or  $b \leq a$ .

**EXAMPLES.** The relation  $a \leq b$  between real numbers is a linear order; there is no maximal or minimal element. The complex numbers  $x + yi$  can be partially ordered by the definition:  $a + bi \leq c + di$  if  $a < c$  or if  $a = c$  and  $b = d$ . Numbers with the same real part are not compared.

A partially ordered set  $A$  is said to be a *lattice* if each subset containing two elements has a least upper bound and a greatest lower bound.

**EXAMPLE.** Let  $A$  be the class of all subsets of a given set  $B$  and let  $S \leq T$  if  $S$  is a subset of  $T$ , i.e., if  $S \subset T$ . Then  $A$  is a lattice and l.u.b.  $\{S, T\} = S \cup T$ , g.l.b.  $\{S, T\} = S \cap T$ . One extends this notation to lattices generally and uses  $a \cup b$  (a cup  $b$ ) for l.u.b.  $\{a, b\}$ ,  $a \cap b$  (a cap  $b$ ) for g.l.b.  $\{a, b\}$ . If every subset of  $A$  has a g.l.b. and a l.u.b., then  $A$  is called a *complete lattice*. The *example* given of a class of all subsets of a given set is a complete lattice.

In a lattice the operations  $\cup$  and  $\cap$  have the following properties:  
For all  $a, b, c$  in  $A$

- $a \cup b = b \cup a$  and  $a \cap b = b \cap a$  (commutative laws);  
 $(a \cup b) \cup c = a \cup (b \cup c)$  and  $(a \cap b) \cap c = a \cap (b \cap c)$  (associative laws);  
 $a \cup a = a, a \cap a = a$  (idempotent laws);  
 $a \cap (a \cup b) = a, a \cup (a \cap b) = a$  (absorptive laws).

A lattice is said to be *distributive* if for all  $a, b, c$  in  $A$

$$a \cup (b \cap c) = (a \cup b) \cap (a \cup c)$$

or, equivalently, if  $a \cap (b \cup c) = (a \cap b) \cup (a \cap c)$  for all  $a, b, c$ .

If  $A$  has a minimal element and a maximal element, one ordinarily denotes them by  $0, 1$  respectively. Two elements  $a$  and  $b$  are said to be complements of each other if  $a \cap b = 0$  and  $a \cup b = 1$ . A lattice is said to be *complemented* if each of its elements has a complement. In a distributive lattice, no element can have more than one complement. A *Boolean algebra* is a complemented, distributive lattice. (See Chap. 11.)

**EXAMPLE.** The partially ordered set formed of the class of all subsets of a given set  $B$  forms a Boolean algebra. The minimal and maximal elements are the empty set  $\emptyset$ , and the set  $B$  the complement is the same as that defined in Sect. 1.

**REMARK.** A relation  $\lesssim$  which satisfies only conditions (i) and (ii) is called a *preorder* or *quasi-order*. An example is the relation "is at least as good as" between automobiles.

## 8. SETS OF POINTS

Sets of real numbers can be interpreted as sets of points on the *real line*, or *number axis*. For fixed  $a, b, a < b$ ,

- $\{x | a < x < b\}$  is an *open interval*,  
 $\{x | a \leq x \leq b\}$  is a *closed interval*,  
 $\{x | a \leq x < b\}$  or  $\{x | a < x \leq b\}$  is a *half-open interval*.

For fixed  $a, \epsilon, \epsilon > 0$ , the set

$$\{x | a - \epsilon < x < a + \epsilon\}$$

is the  $\epsilon$ -neighborhood of  $a$ . An arbitrary set of real numbers is *open* if each element of the set has an  $\epsilon$ -neighborhood contained in the set. A set is *closed* if its complement is open. A number  $a$  is a *limit point* of a set  $A$  if every  $\epsilon$ -neighborhood of  $a$  contains at least one element of  $A$  differing from  $a$ .

Sets of ordered number pairs  $(x, y)$  can be interpreted as sets of points in the  $xy$ -plane. For fixed  $(a, b)$  and  $\epsilon > 0$  the set

$$\{(x, y) \mid (x - a)^2 + (y - b)^2 < \epsilon\}$$

is the  $\epsilon$ -neighborhood of  $(a, b)$ . A set of points in the  $xy$ -plane is *open* if each point  $(a, b)$  in the set has an  $\epsilon$ -neighborhood contained in the set. A set is *closed* if its complement is open. A point  $(a, b)$  is a *limit point* of a set  $A$  if every  $\epsilon$ -neighborhood of  $(a, b)$  contains at least one element of  $A$  differing from  $(a, b)$ . An open set is called an *open region* or *domain* if each two points of the set can be joined by a broken line within the set. A point  $(a, b)$  is a *boundary point* of set  $A$  if every  $\epsilon$ -neighborhood of  $(a, b)$  contains at least one point of  $A$  and at least one point not in  $A$ . The boundary of  $A$  is the set of all boundary points of  $A$ . A *closed region* is a set formed of the union of an open region and its boundary. A point  $(a, b)$  of a set  $A$  is called an *isolated point* of  $A$  if some  $\epsilon$ -neighborhood of  $(a, b)$  contains no element of  $A$  other than  $(a, b)$ .

## REFERENCES

The references for this chapter fall into several levels. The most elementary discussions of foundations and set theory are found in Refs. 1 and 4. References 5 and 7 are basic graduate level texts in the foundation of mathematics; Ref. 6 is at the same level for general set theory. Reference 3 (Chap. 11) gives a simple introduction to lattice theory and Boolean algebra. Reference 2 is a treatise on all phases of lattice theory, including preorder, partial order, and Boolean algebra.

1. C. B. Allendoerfer and C. O. Oakley, *Principles of Mathematics*. McGraw-Hill, New York, 1955.
2. Garrett Birkhoff, *Lattice Theory*, American Mathematical Society, New York, 1948.
3. Garrett Birkhoff and Saunders MacLane, *A Survey of Modern Algebra*, Macmillan, New York, 1941.
4. J. G. Kemeny, J. L. Snell, and G. L. Thompson, *Introduction to Finite Mathematics*, Prentice-Hall, Englewood Cliffs, New Jersey, 1957.
5. R. B. Kershner and L. R. Wilcox, *The Anatomy of Mathematics*, Ronald, New York, 1950.
6. Erich Kamke, *Theory of Sets*, Dover, New York, 1950.
7. R. L. Wilder, *Introduction to the Foundations of Mathematics*, Wiley, New York, 1952.



## Algebraic Equations

R. C. Lyndon

1. Polynomials	2-01
2. Real Roots	2-03
3. Complex Roots	2-04
References	2-06

### 1. POLYNOMIALS

A *polynomial* may be defined as a function  $f = f(x)$  defined by an equation

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

where the coefficients  $a_0, a_1, \dots, a_n$  are constants (real or complex) and  $x$  is variable (real or complex). The *leading coefficient*  $a_n$  will be assumed  $\neq 0$ . The *degree* of  $f$  is  $n$ . A polynomial  $a_2 x^2 + a_1 x + a_0$  of degree 2 is *quadratic*; a polynomial  $a_1 x + a_0$ , of degree 1 is *linear*; we accept the constant polynomials:  $f(x) = a_0$ , although the zero polynomial  $f(x) \equiv 0$  must be tacitly excluded from certain contexts.

An algebraic equation of degree  $n$  is an equation of form: polynomial of degree  $n$  in  $x = 0$ ; that is, of form

$$f(x) = a_n x^n + \cdots + a_0 = 0 \quad (a_n \neq 0).$$

A *root* of such an equation is a value of  $x$  which satisfies it; a root of the equation is called a *root* of  $f(x)$  or a *zero* of  $f(x)$ . Thus  $r$  is a root of  $f(x)$  if and only if  $f(r) = 0$ .

The *fundamental theorem of algebra* asserts that an algebraic equation of degree  $n$  ( $n = 1, 2, \dots$ ) has at least one root (real or complex) (Refs. 1, 2). From this it follows that an algebraic equation of degree  $n$  has exactly  $n$  roots (some of which may be repeated, see below).

The operations of addition, subtraction, multiplication, and division of polynomials will be assumed to be familiar.

*Synthetic division* is an abbreviation of division by a linear polynomial,  $x - c$ . As an *illustration*, the division of  $3x^2 - 7x + 11$  by  $x - 2$  is carried out in long form and by synthetic division.

$$\begin{array}{r}
 3x - 1 \\
 x - 2 \overline{) 3x^2 - 7x + 11} \\
 \underline{3x^2 - 6x} \phantom{+ 11} \\
 -x + 11 \\
 \phantom{-x + } \underline{x + 2} \\
 \phantom{-x + } \phantom{x + } 9
 \end{array}
 \qquad
 2 \begin{array}{|c|c|c|}
 \hline
 3 & -7 & 11 \\
 \hline
 0 & 6 & -2 \\
 \hline
 3 & -1 & 9 \\
 \hline
 \end{array}$$

Either method yields the *quotient*  $3x - 1$  and *remainder* 9, so that

$$3x^2 - 7x + 11 = (3x - 1)(x - 2) + 9.$$

In the synthetic process, on the first line one replaces  $x - 2$  by  $+2$ ,  $3x^2 - 7x + 11$  by the numbers 3,  $-7$ , 11. A zero is placed below the 3 and added to yield 3; the result is multiplied by 2 to yield 6; the 6 is added to  $-7$  to yield  $-1$ ; the  $-1$  is multiplied by 2 to yield  $-2$ ; the  $-2$  is added to 11 to yield 9. On the third line the coefficients of the quotient,  $3x - 1$ , and the remainder, 9, appear in order.

**REMAINDER THEOREM.** *If a polynomial  $f(x)$  is divided by  $x - c$ , then the remainder is  $f(c)$ .*

**FACTOR THEOREM.**  *$c$  is a root of  $f(x)$  if and only if  $x - c$  is a factor of  $f(x)$  (Ref. 2).*

**APPLICATION.** If one root,  $c$ , of  $f(x)$  has been found, the remaining roots of  $f(x)$  will be roots of the quotient polynomial  $f(x) \div (x - c)$ , which is of degree  $n - 1$ . Repetition of this reasoning leads to a representation of  $f(x)$  as a constant times a product of linear factors  $(x - c_1), (x - c_2), \dots$ . Since  $f(x)$  is of degree  $n$  there must be exactly  $n$  such factors:

$$f(x) = a_n(x - c_1)(x - c_2) \cdots (x - c_n).$$

Thus  $f(x)$  has  $n$  roots  $c_1, c_2, \dots, c_n$ , some of which may be equal. If  $c_1$  is repeated  $m$  times, so that  $(x - c_1)^m$  is a factor of  $f(x)$  (and  $(x - c_1)^{m+1}$  is not a factor), then  $c_1$  is a root of *multiplicity*  $m$ .

**Repeated Roots.** If  $c$  is a repeated root of  $f(x)$  (a root of multiplicity 2 or more), then  $c$  will also be a root of  $f'(x)$ , the *derivative* of  $f(x)$ ,

$$f'(x) = na_nx^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + a_1.$$

To find the repeated roots, one can proceed as follows. Let

$$f_0(x) = f(x), \quad f_1(x) = f'(x),$$

and by division obtain

$$f_0(x) = g_1(x)f_1(x) + f_2(x),$$

where  $f_2(x)$  is of degree lower than that of  $f_1(x)$ . Continue, taking

$$f_{i-1}(x) = g_i(x)f_i(x) + f_{i+1}(x),$$

until  $f_{i+1}(x) = 0$ . Then the repeated roots of  $f(x)$  are the roots of  $f_i(x)$ . If  $f_i(x)$  is a (non-zero) constant,  $f(x)$  has no repeated roots. Otherwise all repeated roots of  $f(x)$  can be found as the roots of  $f_i(x)$ , which has degree lower than that of  $f$  (Ref. 2).

## 2. REAL ROOTS

In this section  $f(x)$  denotes a polynomial with *real* coefficients. If  $f(x)$  is of odd degree,  $f(x)$  has at least one real root, whereas  $x^2 + 1$ , for example, has no real roots. Two problems will be considered: (1) establishing existence of real roots, perhaps within prescribed intervals; (2) computing to a satisfactory accuracy the value of a root that has been approximately located.

**Graphical Methods.** One plots the graph of  $y = f(x)$ . The roots of odd multiplicity are the values of  $x$  at which the curve crosses the  $x$ -axis, while at roots of even multiplicity the curve is tangent to the  $x$ -axis. If  $f(x_1)$  and  $f(x_2)$  have opposite signs, there is a root between  $x_1$  and  $x_2$ . In practice, one could use synthetic division to compute the values of  $f(x)$  for a number of values of  $x$  within some interval  $a \leq x \leq b$ . The values  $a$  and  $b$  can be chosen so that all roots lie between  $a$  and  $b$ ; in particular, all real roots lie in the interval

$$-\left(\frac{M}{|a_n|} + 1\right) \leq x \leq \frac{M}{|a_n|} + 1,$$

where  $M$  is the largest of the numbers  $|a_0|, |a_1|, \dots, |a_{n-1}|$ . Narrower bounds can often be found by inspection. If in computation of  $f(b)$  by synthetic division, the third row consists of non-negative numbers, then no real root exceeds  $b$ . An alternative criterion is Newton's rule: if the values  $f(b), f'(b), \dots, f^{(n)}(b)$  of the successive derivatives are all non-negative, then no root exceeds  $b$ . These last two rules can be applied to the equa-

tion obtained by replacing  $x$  by  $-x$ , in order to obtain a lower bound  $a$ . The following rule is sometimes useful: if  $g(x) = x^n f(1/x)$  and if  $g$  has all of its real roots between  $-b$  and  $+b$ , then  $f$  has no real roots between  $-(1/b)$  and  $(1/b)$ .

**Derivative.** The value  $f'(c)$  of the derivative at  $c$  gives the rate of increase (decrease, if  $f'(c) < 0$ ) of  $f(x)$  at  $x = c$ . At an extremum (relative maximum or minimum) of  $f(x)$ ,  $f'(x)$  is zero; there can be at most  $n - 1$  such values of  $x$  (*critical points* of  $f(x)$ ).

**ROLLE'S THEOREM.** *Between each two real roots of  $f(x)$  there is at least one critical point.*

**Descartes and Sturm Tests.** Zero is a root of  $f(x)$  only if  $a_0 = 0$ . By division by  $x$  or some power of  $x$ , all zero roots can be removed. Information about the number of positive roots is given by:

**DESCARTES'S RULE.** *The signs of the coefficients  $a_n, a_{n-1}, \dots, a_0$  in order, omitting possible zeros, form a string of '+'s and '-'s. The number  $v$  of alternations in sign is defined as the number of consecutive pairs  $+-$  or  $-+$ . The number  $p$  of positive roots is no greater than  $v$  and  $v - p$  is even. (Negative roots of  $f(x)$  are the positive roots of  $f(-x)$ .)*

**EXAMPLES.**  $x^2 + x + 1 = 0, v = 0$ , no positive roots;  $x^2 - 2x + 3 = 0, v = 2$ , 0 or 2 positive roots;  $x^2 + 2x - 3 = 0, v = 1$ , 1 positive root. A more precise criterion is given by:

**STURM'S THEOREM.** *Write  $f_0(x) = f(x), f_1(x) = f'(x)$  and, stepwise,  $f_{i-1}(x) = q_i(x)f_i(x) - f_{i+1}(x)$ , where  $f_{i+1}(x)$  is of lower degree than  $f_i(x)$ . Continue until some  $f_{m+1}(x) = 0$ . Now suppose  $a < b, f(a) \neq 0, f(b) \neq 0$ . Let  $v(a)$  be the number of alternations in sign in the sequence of values  $f_1(c), f_2(c), \dots, f_m(c)$  (zeros omitted). Then  $v(a) - v(b)$  is the exact number of distinct real roots between  $a$  and  $b$  (Ref. 2).*

**Newton's Method.** If  $x_1$  is an approximate value of a root of  $f(x)$  then one sets

$$x_2 = x_1 - [f(x_1)/f'(x_1)],$$

$$x_3 = x_2 - [f(x_2)/f'(x_2)], \dots$$

The sequence of numbers thus defined converges to a real root of  $f(x)$ , provided  $f(x_1)f''(x_1) > 0$  and it is known that  $x_1$  lies in an interval containing a root of  $f(x)$  but none of  $f'(x)$  or of  $f''(x)$  (Ref. 2).

### 3. COMPLEX ROOTS

Let  $f(z) = a_n z^n + \dots + a_0$  be a polynomial in the *complex* variable  $z, z = x + iy, i = \sqrt{-1}$ . The coefficients are allowed to be real or complex. If they are real, complex roots of  $f(z)$  come in conjugate pairs,  $x \pm yi$ , so that the total number of nonreal complex roots is even.

If  $f(z)$  is of degree 2, 3, or 4, explicit algebraic formulas for all roots are available (Ref. 1).

It is proved in Galois theory that similar formulas for equations of higher degree do not exist (Ref. 1).

**Equations for Real and Imaginary Parts.** Replacement of  $z$  by  $x + iy$  in the equation  $f(z) = 0$  and equating real and imaginary parts separately to zero leads to two simultaneous equations in the real variables  $x, y$ . These can be solved by elimination.

**EXAMPLE.**  $z^3 - z + 1 = 0$ . Replacement of  $z$  by  $x + iy$  leads to the equations  $x^3 - 3xy^2 - x + 1 = 0, 3x^2y - y^3 - y = 0$ . To find nonreal roots, one assumes  $y \neq 0$  and is led to the equations  $8x^3 - 2x - 1 = 0, y^2 = 3x^2 - 1$ . The first has one real root  $x = 0.66$ . Hence  $0.66 \pm 0.55i$  are the nonreal roots of the equation.

**Application of Argument Principle.** The argument principle, when applied to the polynomial  $f(z)$ , states that the total change in the argument (polar angle) of the complex number  $w = f(z)$ , as  $z$  traces out a simple closed path (circuit)  $C$ , equals  $2\pi$  times the number of zeros of  $f(z)$  inside  $C$  (provided  $f(z) \neq 0$  on  $C$ ). (See Chap. 7, Sect. 5.) The path  $C$  can be chosen as a circle, semicircle, square, or other convenient shape, and the variation of the argument of  $w$  can be evaluated graphically. One can pass to the limit from a semicircle in order to find the number of roots in a half-plane. This is the basis of the *Nyquist criterion* (Chap. 21).

In general, no root can lie outside the circle with center at  $z = 0$  and radius  $1 + (M/|a_n|)$ , where  $M$  is the largest of  $|a_0|, |a_1|, \dots, |a_{n-1}|$  (Ref. 2).

**Hurwitz-Routh Criterion.** This is a rule for determining whether all roots of  $f(z)$  lie in the left half-plane (i.e., have negative real parts). For a given sequence  $c_0, c_1, \dots, c_n, \dots$ , one denotes by  $\Delta_k$  the determinant

$$\Delta_k = \begin{vmatrix} c_1 & c_0 & 0 & 0 & \cdot & 0 \\ c_3 & c_2 & c_1 & c_0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{2k-1} & c_{2k-2} & \cdot & \cdot & \cdot & c_k \end{vmatrix},$$

so that  $\Delta_1 = c_1$ ,

$$\Delta_2 = \begin{vmatrix} c_1 & c_0 \\ c_3 & c_2 \end{vmatrix}, \quad \Delta_3 = \begin{vmatrix} c_1 & c_0 & 0 \\ c_3 & c_2 & c_1 \\ c_5 & c_4 & c_3 \end{vmatrix}.$$

For a given polynomial  $f(z) = c_0z^n + c_1z^{n-1} + \dots + c_n$  with real coefficients and  $c_0 > 0$ , one forms  $\Delta_1, \dots, \Delta_{n-1}$ , with  $c_k$  replaced by 0 for

$k > n$ . All roots of  $f(z)$  lie in the left half-plane if and only if  $\Delta_1 > 0$ ,  $\Delta_2 > 0$ ,  $\dots$ ,  $\Delta_{n-1} > 0$  (Ref. 3).

**Gräffe's Method.** Gräffe's method is efficient for finding a complex root, or successively all roots, of a polynomial  $f(z)$ . For simplicity, suppose that  $f(z)$  has no repeated roots, as can always be arranged by the methods indicated above (Sect. 1). One must further suppose that  $f(z)$  has a single root  $r_0$  of maximum absolute value; if this fails for  $f(z)$  it will hold for the new polynomial  $g(z) = f(z + c)$  for all but certain special values of  $c$ . It is necessary to have some rough idea of the argument of the root  $r_0$ ; for example, if  $r_0$  is real, to know whether it is positive or negative.

Starting with the polynomial  $f(z) = f_1(z) = z^n + a_1z^{n-1} + \dots$ , one forms  $f_1(-z)$ . The product  $f_1(z)f_1(-z)$  contains only even powers of  $z$ , hence is of the form  $f_1(z)f_1(-z) = f_2(z^2)$ . Similarly,  $f_3(z)$  is formed from  $f_2(z)$ :  $f_3(z^2) = f_2(z) \cdot f_2(-z)$ , and the process is continued to form a sequence of polynomials  $f_k(z) = z^n + a_kz^{n-1} + \dots$ . (As justification note that  $f_k$  has roots which are the  $2^k$ th powers of the roots of  $f$ ; that  $-a_k$  is the sum of the roots of  $f_k$ , and hence that the ratio of  $-a_k$  to  $r_0^{2^k}$  approaches 1 as  $k \rightarrow \infty$ ). One chooses a value  $z_k$  of the  $2^k$ th root of  $-a_k$ ; the choice of  $z_k$  is made to agree as closely as possible in argument with the initial estimate for the argument of  $r_0$ . The successive values  $z_1, z_2, \dots$  can be expected to approach  $r_0$  rapidly.

After the root of largest absolute value has been found, one could divide out the corresponding factor and proceed to find the root of next largest absolute value. In practice, it is generally more efficient to use an elaboration of Gräffe's method (Ref. 7).

## REFERENCES

1. Garrett Birkhoff and Saunders MacLane, *A Survey of Modern Algebra* (Revised edition), Macmillan, New York, 1953.
2. L. E. Dickson, *First Course in the Theory of Equations*, Wiley, New York, 1922.
3. E. A. Guillemin, *The Mathematics of Circuit Analysis*, Wiley, New York, 1949.
4. C. C. MacDuffee, *Theory of Equations*, Wiley, New York, 1954.
5. J. V. Uspensky, *Theory of Equations*, McGraw-Hill, New York, 1948.
6. L. Weisner, *Introduction to the Theory of Equations*, Macmillan, New York, 1938.
7. F. A. Willers, *Practical Analysis*, Dover, New York, 1948.

## Matrix Theory

R. C. Lyndon

1. Vector Spaces	3-01
2. Linear Transformations	3-03
3. Coordinates	3-04
4. Echelon Form	3-05
5. Rank, Inverses	3-07
6. Determinants, Adjoint	3-08
7. Equivalence	3-09
8. Similarity	3-10
9. Orthogonal and Symmetric Matrices	3-13
10. Systems of Linear Inequalities	3-14
References	3-17

### 1. VECTOR SPACES

Let  $F$  denote the rational number system, or the real number system, or the complex number system; in the following, elements of  $F$  are termed *scalars* and are denoted by small Roman letters  $a, b, c, \dots$ . A *vector space*  $V$  over  $F$  is defined (Ref. 9) as a set of elements called *vectors*, denoted by small Greek letters  $\alpha, \beta, \gamma, \dots$ , for which the operations of addition:  $\alpha + \beta$  and multiplication by scalars:  $\alpha a$  are defined and satisfy the following rules:

(i) For each pair  $\alpha, \beta$  in  $V$ ,  $\alpha + \beta$  is an element of  $V$  and  $\alpha + \beta = \beta + \alpha$ ,  $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma$ ;

(ii) For each  $\alpha$  in  $V$  and each  $a$  in  $F$ ,  $a\alpha$  is an element of  $V$  and, for arbitrary  $b$  in  $F$  and  $\beta$  in  $V$

$$\begin{aligned} a(\alpha + \beta) &= a\alpha + a\beta, & (a + b)\alpha &= a\alpha + b\alpha, \\ a(b\alpha) &= (ab)\alpha, & 1\alpha &= \alpha; \end{aligned}$$

(iii) For given  $\alpha, \beta$  in  $V$ , there is a unique vector  $\gamma$  in  $V$  such that  $\alpha + \gamma = \beta$ . In particular, there is a unique vector denoted by  $0$  such that  $\alpha + 0 = \alpha$  for all  $\alpha$  in  $V$ .

When  $F$  is the real number system,  $V$  is called a *real* vector space; when  $F$  is the complex number system,  $V$  is a *complex* vector space. The system  $F$  can be chosen more generally as a *field* (Ref. 9). The vectors of mechanics in 3-dimensional space form a real vector space  $V$ . In terms of a coordinate system, the elements of  $V$  are ordered triples  $(x, y, z)$  of real numbers; addition and multiplication by real scalars are defined as follows:

$$\begin{aligned} (x_1, y_1, z_1) + (x_2, y_2, z_2) &= (x_1 + x_2, y_1 + y_2, z_1 + z_2), \\ a(x, y, z) &= (ax, ay, az). \end{aligned}$$

A vector  $\alpha$  is said to be a *linear combination* of vectors  $\alpha_1, \dots, \alpha_n$  if

$$\alpha = a_1\alpha_1 + \dots + a_n\alpha_n$$

for appropriate choice of  $a_1, \dots, a_n$ . An ordered set  $\{\alpha_1, \dots, \alpha_n\}$  is said to be *independent* if no member of the set is a linear combination of the others or, equivalently, if

$$a_1\alpha_1 + \dots + a_n\alpha_n = 0$$

implies  $a_1 = 0, \dots, a_n = 0$ . If the ordered set  $S = \{\alpha_1, \dots, \alpha_n\}$  is independent and  $\alpha$  is a linear combination of its elements (is *linearly dependent* on  $S$ ), then the scalars  $a_1, \dots, a_n$  can be chosen in only one way so that  $\alpha = \sum_i a_i \alpha_i$ .

If there is a finite set  $S = \{\alpha_1, \dots, \alpha_n\}$  such that every  $\alpha$  in  $V$  is linearly dependent on  $S$ , then  $V$  is said to be of *finite dimension*. For the remainder of this chapter, only vector spaces of finite dimension will be considered; this is, however, not the only case of importance. If  $S = \{\alpha_1, \dots, \alpha_n\}$  is independent and every  $\alpha$  of  $V$  is a linear combination of these vectors, then  $S$  is said to constitute a *basis* for  $V$ . Every finite dimensional vector space has at least one basis, all bases have the same number,  $n$ , of elements;  $n$  is the *dimension* of  $V$ .

A subset  $W$  of  $V$  is said to be a *subspace* of  $V$  if, with the operations as defined in  $V$ ,  $W$  is itself a vector space. A subset  $W$  will be a subspace of  $V$  if, whenever  $\alpha, \beta$  are in  $W$ ,  $\alpha + \beta$  is in  $W$ , and  $a\alpha$  is in  $W$  for every  $a$  in

$F$ . In particular,  $\{0\}$  is a subspace, as is  $V$  itself. The intersection (Chap. 1, Sect. 1)  $W \cap U$  of two subspaces of  $V$  is a subspace of  $V$ ; it is the largest subspace contained in both  $W$  and  $U$ . The union  $W \cup U$  is not usually a subspace; the smallest subspace containing  $W$  and  $U$  is rather their (linear) *sum*  $W + U$ , consisting of all vectors  $\alpha + \beta$ ,  $\alpha$  in  $W$ , and  $\beta$  in  $U$ . If  $W \cap U = 0$ , then  $W + U$  is called a *direct sum*, and is often denoted by  $W \oplus U$  or  $W \dot{+} U$ ; in this case every vector in  $W + U$  is expressible uniquely as  $\alpha + \beta$ ,  $\alpha$  in  $W$ ,  $\beta$  in  $U$ . For any set of vectors  $\{\alpha_1, \dots, \alpha_n\}$ , the set of all their linear combinations constitutes a subspace, and the subspace is *spanned* by them. Every independent set is a subset of a basis. From this it follows that, for each subspace  $W$ , there exists  $U$  (in general, many) such that  $V$  is the direct sum of  $W$  and  $U$ .

## 2. LINEAR TRANSFORMATIONS

Let  $f$  be a transformation (function, mapping) (Chap. 1, Sect. 3) of vector space  $V$  into a second space  $V'$ ;  $f$  is said to be *linear* if for all  $\alpha, \beta, a, b$ ,

$$f(a\alpha + b\beta) = af(\alpha) + bf(\beta).$$

The *image* of  $V$  under  $f$ , denoted by  $f(V)$ , is the set of all vectors  $f(\alpha)$  for  $\alpha$  in  $V$ ;  $f(V)$  is a subspace of  $V'$ . If  $f(V) = V'$ ,  $f$  is said to map  $V$  *onto*  $V'$ . The *null space* of  $f$ , denoted by  $N(f)$ , is the set of all vectors  $\alpha$  in  $V$  such that  $f(\alpha) = 0$ ;  $N(f)$  is a subspace of  $V$ . If  $N(f)$  contains only the element  $0$ ,  $f$  is said to be *nonsingular*; this is equivalent to the condition that  $f$  be one-to-one (Chap. 1, Sect. 3); a nonsingular transformation is termed an *isomorphism* of  $V$  onto  $f(V)$ . The *rank* of  $f$  is defined as the dimension of  $f(V)$ ; this equals the dimension of  $V$  minus that of  $N(f)$ . The mapping  $f$  is nonsingular if and only if its rank is maximal, that is, equals the dimension of  $V$ . If  $W$  is chosen so that  $V$  is the direct sum of  $N(f)$  and  $W$ , and  $W$  has dimension greater than  $0$ , then the restriction of  $f$  to  $W$  is a nonsingular mapping of  $W$  onto  $f(V)$ , that is, an isomorphism of  $W$  onto  $f(V)$ . If  $f$  is an isomorphism of  $V$  onto  $V'$ , then the inverse transformation  $f^T = f^{-1}$  is a linear transformation of  $V'$  onto  $V$ .

The set of all linear transformations of  $V$  into  $V'$  becomes itself a vector space over  $F$ , if addition and multiplication by scalars are defined by the rules:

$f + g$  is the transformation such that  $(f + g)\alpha = f(\alpha) + g(\alpha)$  for all  $\alpha$  in  $V$ ;

$af$  is the transformation such that  $(af)\alpha = a[f(\alpha)]$  for all  $\alpha$  in  $V$ .

If  $f$  maps  $V$  into  $V'$  and  $g$  maps  $V'$  into  $V''$ , following  $f$  by  $g$  defines the *composite* transformation  $fg$  of  $V$  into  $V''$ ; explicitly,  $fg(a) = g[f(a)]$ . If  $f, g$  are linear, so also is  $fg$  (Refs. 2, 8, 9).

## 3. COORDINATES

Let  $\{\alpha_1, \dots, \alpha_n\}$  be a basis for the vector space  $V$ , so that every vector  $\alpha$  in  $V$  can be written uniquely in the form  $\Sigma a_i \alpha_i$ . The  $a_i$  are the *coordinates* of  $\alpha$  relative to the chosen basis; the  $a_i$  are also termed *components*, but this word is sometimes used for the terms  $a_i \alpha_i$ . The choice of a definite basis is often necessary for computation. With a fixed basis understood, one can replace each vector  $\alpha$  by the corresponding  $n$ -tuple  $(a_1, \dots, a_n)$ ; then

$$(a_1, \dots, a_n) + (b_1, \dots, b_n) = (a_1 + b_1, \dots, a_n + b_n),$$

$$c(a_1, \dots, a_n) = (ca_1, \dots, ca_n).$$

A basis that is natural at one stage of a problem may not be the most advantageous at a later stage, so that one must be prepared to change bases.

If a basis  $\alpha_1, \dots, \alpha_n$  is chosen for  $V$  and a basis  $\alpha'_1, \dots, \alpha'_m$  for  $V'$ , then each linear transformation of  $V$  into  $V'$  can be assigned coordinates as follows. The transformation  $f$  is fully determined by the images  $f(\alpha_i)$  of the basis elements for  $V$ . If  $f(\alpha_i) = \Sigma_j a_{ij} \alpha'_j$ , then  $f$  may be characterized by the  $n \cdot m$  scalars  $a_{ij}$ , where  $i = 1, \dots, n, j = 1, \dots, m$ . These numbers are usually thought of as arranged in a rectangular array, or *matrix*

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} = (a_{ij}).$$

One terms  $A$  the matrix *representing* the transformation  $f$  relative to the given bases in  $V$  and  $V'$ .

If  $g$  is a second transformation from  $V$  into  $V'$ , with matrix  $B = (b_{ij})$ , it is clear that the transformation  $f + g$  will have the matrix  $(a_{ij} + b_{ij})$ . Accordingly, one defines the *sum* of two  $n$  by  $m$  matrices as follows:

$$(a_{ij}) + (b_{ij}) = (a_{ij} + b_{ij}).$$

Similarly, the product  $cA$ , which represents  $cf$ , is defined as the matrix  $(ca_{ij})$ .

Now let  $f$  be a linear transformation of  $V$  into  $V'$ ,  $g$  a linear transformation of  $V'$  into  $V''$ , where  $V, V'$  have bases as before, and  $V''$  has a basis  $\alpha''_1, \dots, \alpha''_p$ . Relative to these bases,  $f$  is represented by an  $n$  by  $m$

matrix  $A = (a_{ij})$ ,  $g$  by an  $m$  by  $p$  matrix  $B = (b_{ij})$ ,  $fg$  by an  $n$  by  $p$  matrix  $C = (c_{ij})$ . Since

$$(fg)(\alpha_i) = \sum_j \sum_k a_{ij} b_{jk} \alpha''_k$$

one finds

$$c_{ik} = \sum_{j=1}^m a_{ij} b_{jk};$$

correspondingly, one defines the *product of two matrices*  $A$  and  $B$  (where the number of columns of  $A$  equals the number of rows of  $B$ ) to be the matrix  $C = AB$ , where the elements  $c_{ik}$  of  $C$  are given by the above "row-by-column" rule. Multiplication of matrices is not commutative, but is associative and distributive:  $A(BC) = (AB)C$ ,  $A(B + C) = AB + AC$ ,  $(A + B)C = AC + BC$ .

If  $\alpha = \sum a_i \alpha_i$  is a vector with coordinate representation  $(a_1, \dots, a_n)$ , one can regard the  $n$ -tuple as a 1 by  $n$  matrix. The product  $\alpha A$  can then be evaluated as that of a 1 by  $n$  matrix and an  $n$  by  $m$  matrix. The result is the 1 by  $m$  matrix

$$\alpha A = \left( \sum_{i=1}^n a_i a_{i1}, \dots, \sum_{i=1}^n a_i a_{im} \right)$$

which represents  $f(\alpha)$ :

$$\begin{aligned} f(\alpha) &= f\left(\sum_i a_i \alpha_i\right) = \sum_i a_i f(\alpha_i) = \sum_i \sum_j a_i a_{ij} \alpha'_j \\ &= \sum_{i=1}^m \left( \sum_{i=1}^n a_{ij} a_i \right) \alpha'_j \end{aligned}$$

This shows that, when bases are chosen in  $V$  and  $V'$ , each matrix  $A$  is the matrix of a linear transformation (Ref. 2, 9).

#### 4. ECHELON FORM

The matrix  $A$  associated with a linear transformation  $f$  from  $V$  to  $V'$  can be given an especially simple form by suitable choice of basis for  $V$ , for  $V'$ , or for both. We consider the effect of a change of basis for  $V$ . Every change of basis for  $V$  can be effected by a sequence of *elementary transformations* of the following types: (1) replacement of  $\alpha_i$  by a scalar multiple  $c\alpha_i$ ,  $c \neq 0$ ; (2) renumbering, interchanging  $\alpha_i$  and  $\alpha_j$ ; (3) adding to  $\alpha_i$  some multiple of  $\alpha_j$ ,  $j \neq i$ , so that  $\alpha_i$  is replaced by  $\alpha_i + c\alpha_j$  (and  $\alpha_j$  is left unchanged). The effect of each transformation is to carry out the analogous operation on the rows of the matrix  $A = (a_{ij})$ . Thus (1) multiplies each element of the  $i$ th row by  $c$ , (2) interchanges  $i$ th and  $j$ th rows, (3) replaces the  $i$ th row by  $(a_{i1} + ca_{j1}, \dots, a_{im} + ca_{jm})$ .

A matrix is said to be in (strict) *echelon form* (Ref. 9) if:

- (i) The leading element (first nonzero element) in each nonzero row appears farther to the right than that of any preceding row;
- (ii) The leading elements are all 1;
- (iii) Only zeros appear in the same column with a leading element;
- (iv) All zero rows (if any) appear at the bottom.

By a *zero row* (or column) is meant one consisting wholly of zeros.

**EXAMPLE.** The following matrix is in echelon form.

$$A = \begin{bmatrix} 0 & 1 & 3 & 0 & 0 & 5 \\ 0 & 0 & 0 & 1 & 0 & 7 \\ 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Each matrix can be reduced to echelon form by elementary transformations on its rows, as follows:

**Step 1.** If the first column is a zero column, leave it untouched and proceed to the matrix formed by the remaining columns. If the first column is not a zero column, permute rows so that  $a_{11} \neq 0$ . Dividing this row by  $a_{11}$  gives a new matrix with  $a_{11} = 1$ . Subtracting suitable multiples of this row from the other rows makes all  $a_{i1} = 0$  for  $i \neq 1$ . The matrix now has a first column which is a zero column, or else it has all zeros except for a 1 in the top position. Leave the first row and column untouched and proceed to the matrix formed by the elements not in the first row or column.

**Step 2.** Repeat this process as long as possible. The resulting matrix will satisfy (i), (ii), and (iv).

**Step 3.** To obtain (iii), subtract suitable multiples of each row from earlier rows to convert the elements in these rows above the leading element of the given row into zeros. The result may be stated as follows:

*Every matrix is row-equivalent to an echelon matrix and (it can be shown) to a unique echelon matrix.*

**Application to Systems of Equations** (Ref. 9). A system of  $m$  linear equations in  $n$  unknowns

$$\sum_{j=1}^n a_{ij}x_j = c_i \quad (i = 1, \dots, m)$$

can be replaced by a single matrix equation

$$AX = C,$$

where  $A = (a_{ij})$  and  $X, C$  are column vectors:

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad C = \begin{bmatrix} c_1 \\ \vdots \\ c_m \end{bmatrix}.$$

Let  $B$  be the augmented matrix of the system, obtained by adjoining  $-C$  as  $(n + 1)$ st column to  $A$ . The usual manipulations of equations employed to successively eliminate (so far as possible) the unknowns  $x_1, x_2, \dots, x_n$  correspond to elementary transformations on the matrix  $B$ . If the result were the echelon matrix of the above example, one would have obtained the equivalent system:

$$\begin{aligned} x_2 + 3x_3 &+ 5 = 0 \\ x_4 &+ 7 = 0 \\ x_5 &+ 2 = 0 \end{aligned}$$

Since  $x_1, x_3$  do not appear in leading terms, they can be assigned arbitrary values; the general solution can be obtained immediately from the given equations:

$$x_1 \text{ arbit.}, \quad x_2 = -3x_3 - 5, \quad x_3 \text{ arbit.}, \quad x_4 = -7, \quad x_5 = -2.$$

If a row  $(00 \dots 01)$  had appeared, there would be an equation  $1 = 0$ , as a consequence of the original system, which would therefore be *inconsistent* and have no solution.

## 5. RANK, INVERSES

The rank of a linear transformation  $f$  of  $V$  into  $V'$  was defined (Sect. 2) as the dimension of the image space  $f(V)$ . If  $f$  has matrix  $A$ , then  $f(V)$  is the *row-space* of  $A$ ; that is, the subspace of  $V'$  spanned by the vectors consisting of the rows of  $A$ . The *rank* of  $A$  is defined as the dimension of the row-space of  $A$ ; hence the rank of  $A$  equals the rank of  $f$ . It can be shown that the rank of  $A$  also equals the dimension of the *column-space* of  $A$ . The rank is unaltered by elementary transformations and can be determined by inspection for an echelon matrix, where it is simply the number of nonzero rows.

Let  $f$  be a one-to-one linear transformation of  $V$  onto  $V'$ , so that  $f$  has a linear inverse  $f^{-1}$  (Sect. 2). The spaces  $V, V'$  must have the same dimen-

sion  $m$  and  $f, f^{-1}$  are represented by nonsingular square matrices  $A, B$  such that  $AB = BA = I$ , where

$$I = I_m = \begin{bmatrix} 1 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & 1 \end{bmatrix}$$

is the  $m$  by  $m$  identity matrix. If  $A$  is an arbitrary square nonsingular matrix, there exists a unique inverse  $A^{-1}$  such that  $AA^{-1} = I$  (which implies  $A^{-1}A = I$ ). Hence  $B$  must be  $A^{-1}$ . The echelon matrix for a square nonsingular matrix  $A$  is  $I$ ; the inverse  $A^{-1}$  may be obtained by applying to  $I$  the same sequence of elementary transformations that carry  $A$  into its echelon form  $I$ . The inverse has the properties

$$(cA)^{-1} = c^{-1}A^{-1}, \quad (AB)^{-1} = B^{-1}A^{-1}$$

## 6. DETERMINANTS, ADJOINT

By a permutation  $p$  of the set of integers  $1, 2, \dots, m$  is meant a function  $p:k \rightarrow k' = p(k)$  which is a one-to-one transformation of this set onto itself (Ref. 2). Each such permutation is classified as *even* or *odd* according as the polynomials in  $m$  variables

$$P = \prod_{k < l} (x_k - x_l), \quad P' = \prod_{k < l} (x_{p(k)} - x_{p(l)})$$

are the same or negatives of each other.

EXAMPLE. If  $m = 3$ , and  $p(1) = 3, p(2) = 1, p(3) = 2$ , then  $p$  is even, since

$$(x_1 - x_2)(x_1 - x_3)(x_2 - x_3) = (x_3 - x_1)(x_3 - x_2)(x_1 - x_2).$$

One denotes by  $\text{sgn } p$  the value 1 if  $p$  is even, the value  $-1$  if  $p$  is odd.

The determinant (Refs. 1, 9)  $\det A$  of a square  $m$  by  $m$  matrix  $A = (a_{ij})$  is defined to be the scalar

$$\det A = \sum_p \text{sgn } p \cdot a_{1p(1)} \cdot a_{2p(2)} \cdot \dots \cdot a_{mp(m)}$$

where the sum is over all permutations  $p$  of  $1, 2, \dots, m$ . If  $A$  is singular,  $\det A = 0$ . For nonsingular  $A$ ,  $\det A \neq 0$  and  $\det A$  is  $(-1)^h$  times the product of the scalars  $c$  appearing in the elementary transformation of type (1) (Sect. 4) used in reducing  $A$  to the echelon form  $I$ , where  $h$  is the number of transformations of type (2).

Let  $A_{ij}$  denote the submatrix of  $A$  obtained by deleting the  $i$ th row and  $j$ th column. Then for any fixed  $i$ ,

$$\det A = \sum_{j=1}^m (-1)^{i+j} a_{ij} \cdot \det A_{ij};$$

there is an analogous result for expansion according to a fixed column  $j$ . One calls  $\det A_{ij}$  the *minor* of  $a_{ij}$ , and the expansions of  $\det A$  are called *expansions by minors*.

EXAMPLE.

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \\ = a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) \\ + a_{13}(a_{21}a_{32} - a_{22}a_{31}).$$

The *adjoint* ( $\text{adj}$ )  $A$  of a square matrix  $A$  is the matrix  $B = (b_{ij})$ , where

$$b_{ij} = (-1)^{i+j} \det A_{ji}$$

(note the reversal of indices). One has the rule

$$\text{adj } A \cdot A = (\det A) \cdot I$$

and, if  $\det A \neq 0$ ,

$$A^{-1} = (\det A)^{-1} \cdot \text{adj } A,$$

$$\text{adj } A = (\det A) \cdot A^{-1}.$$

CRAMER'S RULE. If  $\det A \neq 0$ , the system

$$\sum_{j=1}^m a_{ij} x_j = c_j \quad (i = 1 \cdots m),$$

has a unique solution

$$x_i = \frac{\det A(i)}{\det A},$$

where  $A(i)$  is the matrix obtained from  $A$  by replacing the  $i$ -th column by  $c_1, \dots, c_m$  (Ref. 2, 9).

## 7. EQUIVALENCE

Let  $f$  be a linear transformation of  $V$  into  $V'$ . It has been seen (Sect. 4) that the matrix  $A$  for  $f$  can be put in echelon form by a suitable change of basis in  $V$ . If  $V'$  is not the same space as  $V$ , one can further simplify  $A$  by independently changing the basis for  $V'$ . This effects elementary trans-

formations on the *columns* of  $A$ ; by successive subtractions of multiples of earlier columns from later ones, followed possibly by a renumbering of the basis,  $A$  can be reduced to the form

$$J_r = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$$

where  $I_r$  is the  $r$  by  $r$  identity matrix (Sect. 5) and the 0's stand for rows and columns consisting wholly of zeros;  $J_r$  is a rectangular  $n$  by  $m$  matrix, just as was the given matrix  $A$ . For the matrix in echelon form in the example of Sect. 4 the matrix  $J_r$  would be

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The effect of a change of basis in  $V$  is to replace  $A$  by  $PA$ , where  $P$  is a nonsingular  $n$  by  $n$  matrix; the effect of a change of basis in  $V'$  is to replace  $A$  by  $AQ$ , where  $Q$  is a nonsingular  $m$  by  $m$  matrix. The matrix  $B$  is said to be *equivalent* to matrix  $A$  if  $B = PAQ$  for some nonsingular  $P$  and  $Q$ . This is a proper equivalence relation (Chap. 1, Sect. 5). The reasoning given above then gives the conclusion: Every  $A$  is equivalent to a unique matrix of the form  $J_r$ . In other words, the matrices  $J_r$  (for various  $r$ ,  $m$ , and  $n$ ) are a set of *canonical forms* under equivalence (Ref. 9).

## 8. SIMILARITY

One now considers the possible matrices  $A$  representing a linear transformation  $f$  of the vector space  $V$  into itself. The field  $F$  of scalars will be assumed to be the complex number system. Since  $V' = V$ , one can no longer change bases in  $V$  and  $V'$  independently. Indeed, let  $\alpha'_i = \sum_j p_{ij}\alpha_j$  be equations defining a new basis  $\alpha'_1, \dots, \alpha'_n$  in  $V$ . Then  $P = (p_{ij})$  is a nonsingular matrix with inverse  $P^{-1} = (q_{ij})$ , and  $\alpha_k = \sum_h q_{kh}\alpha'_h$ . Let  $f$  have the matrix  $A = (a_{jk})$  relative to the basis  $\alpha_i$ , so that  $f(\alpha_i) = \sum_k a_{jk}\alpha'_k$ . Then

$$f(\alpha'_i) = \sum_h \sum_j \sum_k p_{ij} a_{jk} q_{kh} \alpha'_h,$$

and  $f$  has the matrix  $PAP^{-1}$  relative to the basis  $\alpha'_1, \dots, \alpha'_n$ . The square matrix  $B$  is said to be *similar* to square matrix  $A$  if  $B = PAP^{-1}$  for some nonsingular matrix  $P$ . Hence change of basis in  $V$  replaces the matrix of  $f$  by a similar matrix. Similarity is an equivalence relation (Chap. 1, Sect. 5) in the class of square matrices (Ref. 9).

If  $A$  can be reduced to a similar matrix of sufficiently simple form, most

of the important properties of  $A$  can be read off. The ideal situation is that in which  $A$  is similar to a *diagonal* matrix; that is, a matrix  $(a_{ij})$  in which  $a_{ij} = 0$  for  $i \neq j$ . Unfortunately, not every  $A$  is similar to a diagonal matrix, and the various canonical forms are approximations to the diagonal form.

If  $A$  is similar to

$$B = \text{diag} (\lambda_1, \dots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \cdot & \cdot & 0 \\ 0 & \lambda_2 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \lambda_n \end{bmatrix},$$

then in terms of the new basis  $\alpha_1, \dots, \alpha_n$  associated with  $B$  one has

$$f(\alpha_1) = \alpha_1 B = \lambda_1 \alpha_1, \dots, f(\alpha_n) = \alpha_n B = \lambda_n \alpha_n.$$

In general, if a vector  $\alpha \neq 0$  is such that  $\alpha A = \lambda \alpha$  for some scalar  $\lambda$ , then  $\lambda$  is called an *eigenvalue* (*characteristic value, latent root*) of  $A$ , and  $\alpha$  is called an *eigenvector* belonging to  $\lambda$ . The *characteristic polynomial* for  $A$  is the polynomial  $\phi(x) = \det (xI - A)$ ; this is a polynomial

$$\phi(x) = c_0 + c_1 x + \dots + c_n x^n$$

of degree  $n$ , and its  $n$  roots (real or complex) are the eigenvalues of  $A$ . In particular,

$$(-1)^n c_0 = \det A = \lambda_1 \cdot \lambda_2 \cdot \dots \cdot \lambda_n,$$

$$-c_{n-1} = a_{11} + \dots + a_{nn} = \lambda_1 + \dots + \lambda_n = \text{trace of } A, \text{ and } c_n = 1.$$

The HAMILTON-CAYLEY THEOREM (Ref. 9) states that  $A$  satisfies its characteristic equation:

$$\phi(A) = c_0 I + c_1 A + \dots + c_n A^n = 0.$$

If the roots of  $\phi(x)$  are distinct, then  $A$  is similar to  $B = \text{diag} (\lambda_1, \dots, \lambda_n)$ . In fact, let  $\phi(x)$  be a factor of the  $k$ th power of a polynomial  $\psi(x)$ , whose roots are the distinct numbers  $\lambda_1, \dots, \lambda_p$ ; if  $\psi(A) = 0$ , then  $A$  is similar to a diagonal matrix; if  $\psi(A) \neq 0$ , then  $A$  is not similar to a diagonal matrix.

In the general case of repeated roots, the matrix  $A$  is similar to a matrix  $B$  in *Jordan normal form*; that is, a matrix (in partitioned form, see Ref. 9, Sect. 2.8)

$$B = \text{diag} (B_1, \dots, B_s) = \begin{bmatrix} B_1 & 0 & \cdot & \cdot & 0 \\ 0 & B_2 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & B_s \end{bmatrix}$$

where the  $B_i$  are square matrices of form

$$B_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdot & 0 \\ 0 & \lambda_i & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \lambda_i & 1 \\ 0 & 0 & \cdot & 0 & \lambda_i \end{bmatrix}$$

and  $\lambda_1, \dots, \lambda_s$  are not necessarily distinct. In the matrix  $B$  each characteristic root  $\lambda$  appears on the diagonal a number of times equal to its multiplicity.

An alternative *rational canonical form* for matrix  $A$  has the form  $B = \text{diag}(B_1, \dots, B_p)$ , where  $B_i$  has form

$$\begin{bmatrix} 0 & 1 & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & 0 & 1 \\ -c_{i1} & -c_{i2} & \cdot & \cdot & -c_{ik_i} \end{bmatrix}$$

If  $A$  has rational (real) entries, the  $B_i$  can be chosen so that the  $c_{ij}$  are rational (real).

If  $A$  is a real matrix, the eigenvalues  $\lambda$  need not be real but, since  $\phi(x)$  has real coefficients, they will occur in conjugate complex pairs. In this connection it is useful to note that the matrices

$$\begin{pmatrix} re^{i\theta} & 0 \\ 0 & re^{-i\theta} \end{pmatrix}, \quad \begin{pmatrix} r \cos \theta & -r \sin \theta \\ r \sin \theta & r \cos \theta \end{pmatrix}$$

are similar.

When  $A$  is of small degree or is otherwise especially simple, its eigenvalues and eigenvectors can be found by explicit calculation from the definitions given above; often they can be found from the physical interpretation of the problem. Determination of eigenvalues is a problem in solving an algebraic equation (Chap. 2), but other methods are available (Ref. 4). If  $\lambda$  is an eigenvalue having absolute value greater than that of all other eigenvalues and  $\alpha$  is any reasonable approximation to an eigenvector belonging to  $\lambda$  ( $\alpha$  must not lie in the subspace spanned by the eigenvectors of the remaining eigenvalues), then the sequence  $\alpha = \alpha_1, \alpha_2, \dots, \alpha_n, \dots$  where  $\alpha_{i+1} = \alpha_i A / c_i$  and  $c_i$  is the first nonzero coefficient of  $\alpha_i$  will converge to an eigenvector for  $\lambda$ .

9. ORTHOGONAL AND SYMMETRIC MATRICES

Let  $V$  be a real vector space, with basis  $\alpha_1, \dots, \alpha_n$ , so that each vector has coordinates  $(a_1, \dots, a_n)$ . The *inner product* (Ref. 9)  $(\alpha, \beta)$  of the vectors  $\alpha = (a_1, \dots, a_n)$ ,  $\beta = (b_1, \dots, b_n)$  is defined as the scalar

$$(\alpha, \beta) = a_1 b_1 + \dots + a_n b_n.$$

The *norm* of  $\alpha$  is the scalar  $|\alpha| = (\alpha, \alpha)^{1/2}$ . The *angle*  $\theta$  between  $\alpha, \beta$  is defined by the equation

$$(\alpha, \beta) = |\alpha| |\beta| \cos \theta.$$

These definitions are relative to the given basis but are unaffected if a new *orthonormal* basis is introduced; that is, a basis  $\alpha'_1, \dots, \alpha'_n$  such that  $(\alpha'_i, \alpha'_j) = \delta_{ij} = 1$  or  $0$  according as  $i = j$  or  $i \neq j$ . If

$$\alpha'_i = \sum_j a_{ij} \alpha_j,$$

then the matrix  $A = (a_{ij})$  has as its inverse the *transposed* matrix  $A^T = (b_{ij})$ , where  $b_{ij} = a_{ji}$ ; that is,  $AA^T = I$ . A matrix with this property is called *orthogonal*. Since  $\det A = \det A^T$ , and  $\det A \cdot \det A^T = \det I = 1$ , one concludes that  $\det A = \pm 1$ . When  $\det A = 1$ ,  $A$  is called *proper orthogonal* and is a product of rotations; if  $\det A = -1$ ,  $A$  is a product of rotations and one *reflection*, so that orientation is reversed. The eigenvalues of an orthogonal matrix all have absolute value equal to 1.

A real matrix  $A = (a_{ij})$  is termed *symmetric* if  $A = A^T$ ; if, further, the quadratic form  $\sum_{i,j} a_{ij} x_i x_j$  is  $> 0$  except when  $x_1 = \dots = x_n = 0$ , then  $A$  and the quadratic form are said to be *positive definite*. The  $x$ 's can be interpreted as coordinates of a vector  $\alpha$  with respect to a given basis; then  $\sum_{i,j} a_{ij} x_i x_j = (\alpha A, \alpha)$ . If a new basis is chosen (not necessarily orthonormal), the form is replaced by a new quadratic form. When  $A$  is positive definite, the new basis can be chosen so that  $(\alpha A, \alpha)$  has the form  $\sum_i x_i^2$ ; this is equivalent to the statement that  $A$  can be written as  $PP^T$ , where  $P$  is nonsingular. If  $A$  is symmetric, but not necessarily positive definite, the new basis can be chosen so that  $(\alpha A, \alpha)$  has the form

$$y_1^2 + \dots + y_r^2 - y_{r+1}^2 - \dots - y_s^2,$$

where the numbers  $r, s$  are uniquely determined by  $A$ . This is equivalent to the statement that there exists a nonsingular matrix  $P$  such that  $PAP^T = B$ , where  $B = (b_{ij})$ ,  $b_{ij} = 0$  for  $i \neq j$ ,  $b_{ij} = 0$  or  $\pm 1$  for  $i = j$ . (One terms  $B$  *congruent* to  $A$ .)

The eigenvalues of a symmetric matrix  $A$  are all real, and  $A$  is similar to a real diagonal matrix  $C$ ; indeed  $C = PAP^{-1}$ , where  $P$  may be chosen to be orthogonal (Ref. 9).

An analogous theory holds for complex vector spaces. The *inner product* is defined as

$$(\alpha, \beta) = a_1 \bar{b}_1 + \cdots + a_n \bar{b}_n \quad (\bar{b}_i = \text{conj. of } b_i),$$

so that  $(\alpha, \alpha) = \sum_i |a_i|^2 > 0$ ; the *norm* of  $\alpha$  is defined to be  $(\alpha, \alpha)^{1/2}$ . Orthogonal matrices are replaced by *unitary matrices*, defined by the condition  $A\bar{A}^T = I$ , where the bar denotes replacement of each entry by its conjugate. Symmetric matrices are replaced by *Hermitean* matrices, defined by the condition  $A = \bar{A}^T$ .

## 10. SYSTEMS OF LINEAR INEQUALITIES

Let  $V$  be a real vector space with fixed basis  $\{\alpha_1, \cdots, \alpha_n\}$  as in Sect. 9, so that each vector  $\alpha$  has coordinates  $(a_1, \cdots, a_n)$ . If the vector  $\beta$  has coordinates  $(b_1, \cdots, b_n)$  then one writes

$$\beta > \alpha \text{ or } \alpha < \beta \quad \text{if } a_i < b_i \quad (i = 1, \cdots, n),$$

$$\beta \geq \alpha \text{ or } \alpha \leq \beta \quad \text{if } a_i \leq b_i \quad (i = 1, \cdots, n),$$

$$\beta \geq \alpha \text{ or } \alpha \leq \beta \quad \text{if } \alpha \leq \beta \quad \text{but } \alpha \neq \beta.$$

The relation  $\leq$  is a partial order; the relations  $<$  and  $\leq$  are antisymmetric and transitive but not reflexive (Chap. 1, Sects. 4 and 7). A vector  $\alpha$  is said to be

non-negative if  $\alpha \geq 0$ ,  
 positive if  $\alpha \geq 0$ ,  
 strictly positive if  $\alpha > 0$ .

The set  $Q$  of all non-negative vectors is called the *positive orthant* in  $V$ . A positive vector  $\alpha$  such that  $a_1 + \cdots + a_n = 1$  is called a *probability vector*.

For fixed  $\alpha$  and real number  $k$ , the set of all vectors  $\xi = (x_1, \cdots, x_n)$  such that

$$(\alpha, \xi) + k = a_1 x_1 + \cdots + a_n x_n + k \geq 0$$

is a closed set (Chap. 1, Sect. 8) called a *half-space*  $\mathcal{H}$ . For *example*, the solutions of  $2x_1 + 3x_2 - 6 \geq 0$  form the half-space (half-plane) in two-dimensional space, as shaded in Fig. 1. Similarly, the solutions of  $(\alpha, \xi) + k > 0$  constitute an *open* half-space  $\mathcal{H}_o$ . The solutions of  $(\alpha, \xi) + k = 0$  constitute a *hyperplane*  $\Pi$  which is the boundary of both  $\mathcal{H}$  and  $\mathcal{H}_o$ . By a *system of linear inequalities* is meant a set of relations

$$(\alpha_k, \xi) + k_k R_k 0,$$

where the index  $\kappa$  ranges over a given set (possibly infinite), and for each  $\kappa$ ,  $R_\kappa$  is one of the relations  $>$ ,  $\geq$ ,  $=$ . For example;

$$2x_1 + 3x_2 - 6 \geq 0, \quad x_1 + 5x_2 > 0, \quad x_1 + x_3 = 0$$

is a system of linear inequalities. By a *solution* of the system of inequalities is meant a vector  $\xi = (x_1, \dots, x_n)$  which satisfies all the inequalities.

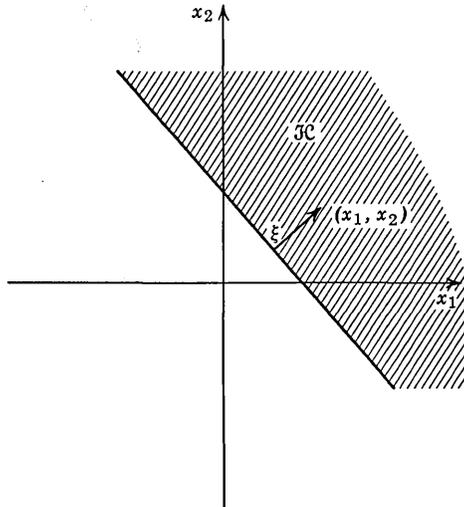


FIG. 1. Half-space in two dimensions.

With each inequality is associated a half-space (or hyperplane)  $\mathcal{H}_\kappa$ . The set of all solutions of the system is the intersection of all  $\mathcal{H}_\kappa$ .

**Convexity.** The vector  $a$  is said to be a *convex combination* of vectors  $\alpha_1, \dots, \alpha_m$  if

$$\alpha = p_1\alpha_1 + \dots + p_m\alpha_m, \quad p_1 + \dots + p_m = 1, \quad p_i \geq 0 \quad (i = 1, \dots, m).$$

A nonempty set  $K$  in  $V$  is said to be *convex* if it contains all convex combinations of its vectors. If  $K$  is interpreted as a point set in  $n$ -dimensional space,  $K$  is convex if and only if, for each pair of points  $\alpha_1, \alpha_2$  in  $K$ , the line segment joining  $\alpha_1$  to  $\alpha_2$  lies in  $K$ . (See Fig. 2.) A half-space  $\mathcal{H}$  is said to be a *support* for a convex set  $K$  in  $V$  if  $K$  is a subset of  $\mathcal{H}$ . If, moreover,  $\Pi$  contains  $n - 1$  independent vectors of  $K$ ,  $\mathcal{H}$  is called an *extreme support* for  $K$ .

Let  $T$  be a subset of  $V$ . The set of all convex combinations of vectors

in  $T$  is a convex set, called the *convex closure* of  $T$ . A convex set  $K$  is said to be *finitely generated* if it is the convex closure of a finite subset of  $K$ .

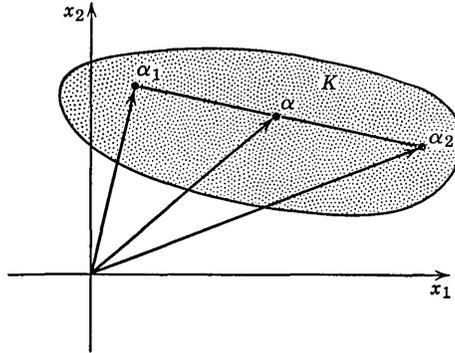


FIG. 2. Convex set.

A set  $T$  is said to be *bounded* if, for some constant  $M$ ,

$$|a_1| + \dots + |a_n| \leq M \text{ for all } \alpha \text{ in } T.$$

**DOUBLE DESCRIPTION THEOREM.** *If  $K$  is a finitely generated convex set in  $V$ , then  $K$  is the intersection of a finite number of supports; moreover, if  $K$  spans  $V$ , then  $K$  is the intersection of its extreme supports. Conversely, if the intersection of a finite collection of half-spaces is nonempty and bounded, then it is finitely generated.*

This theorem, when formulated in algebraic terms, is known as *Farkas' Lemma*:

**FARKAS' LEMMA (strong nonhomogeneous form).** Let  $V$  be  $n$ -dimensional real vector space, let  $V'$  be  $m$ -dimensional real vector space; fixed bases are assumed chosen in each. Let  $A$  be an  $n$  by  $m$  matrix, let  $\delta$  be a vector in  $V'$ , let  $k$  be a scalar, and suppose that there is at least one vector  $\phi$  in  $V$  such that  $\phi A \geq \delta$ . Then a vector  $\alpha$  in  $V$  will satisfy the condition  $(\alpha, \phi) \geq k$  for all  $\phi$  for which  $\phi A \geq \delta$  if and only if there exists a vector  $\gamma \geq 0$  in  $V'$  such that  $\alpha = \gamma A^T$  and  $(\gamma, \delta) \geq k$ .

**FARKAS' LEMMA (weaker homogeneous form,  $k = 0, \delta = 0$ ).** Let  $A$  be an  $n$  by  $m$  matrix. Then a vector  $\alpha$  in  $V$  will satisfy the condition  $(\alpha, \phi) \geq 0$  for all  $\phi$  for which  $\phi A \geq 0$  if and only if there exists a vector  $\gamma \geq 0$  in  $V'$  such that  $\alpha = \gamma A^T$ .

Farkas' Lemma can be used as a foundation for the *Minimax Theorem* in game theory and for the *Duality Theorem* in linear programming. These theorems can also be deduced from the following one:

**THEOREM.** Let  $A = (a_{ij})$  and  $B$  be  $n$  by  $m$  matrices with  $a_{ij} > 0$  for all  $i, j$ . Then there exist probability vectors  $\xi$  in  $V$  and  $\eta$  in  $V'$  and a unique scalar  $k$  such that

$$(kA - B)\eta \geq 0 \quad \text{and} \quad \xi(kA - B) \leq 0;$$

in the first inequality  $\eta$  is regarded as an  $m$  by 1 matrix.

For all of Sect. 10, see Refs. 5, 6, 9.

## REFERENCES

1. A. C. Aitken, *Determinants and Matrices*, Interscience, New York, 1954.
2. Garrett Birkhoff and Saunders MacLane, *A Survey of Modern Algebra* (Revised edition), Macmillan, New York, 1953.
3. M. Bocher, *Introduction to Higher Algebra*, Macmillan, New York, 1930.
4. R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary Matrices*, Cambridge University Press, Cambridge, England, 1938.
5. T. C. Koopmans (Editor), *Activity Analysis of Production and Allocation* (Cowles Commission Monograph No. 13), Wiley, New York, 1951.
6. H. W. Kuhn and A. W. Tucker (Editors), *Contributions to the Theory of Games*, Vols. I, II, III (Annals of Mathematics Studies Nos. 24, 28, 38), Princeton University Press, Princeton, N. J., 1950, 1953, 1956.
7. C. C. MacDuffee, *The Theory of Matrices*, Chelsea, New York, 1946.
8. C. C. MacDuffee, *Vectors and Matrices*, Mathematical Association of America, Buffalo, N. Y., 1943.
9. R. M. Thrall and L. Tornheim, *Vector Spaces and Matrices*, Wiley, New York, 1957.
10. J. H. M. Wedderburn, *Lectures on Matrices*, American Mathematical Society, New York, 1934.



## Finite Difference Equations

G. E. Hay

1. Definitions	4-01
2. Linear Difference Equations	4-03
3. Homogeneous Linear Equations with Constant Coefficients	4-04
4. Nonhomogeneous Linear Equations with Constant Coefficients	4-05
5. Linear Equations with Variable Coefficients	4-07
References	4-08

### 1. DEFINITIONS

By a *difference equation* is meant an equation relating the values of an unspecified function  $f$  at  $x, x + h, x + 2h, \dots, x + nh$ , where  $h$  is fixed. For *example*,

$$(1) \quad f(x + 3) - f(x + 2) - xf(x + 1) - 2f(x) = x^2$$

is a difference equation, in which  $h = 1, n = 3$ . The variable  $x$  will generally be assumed to vary over the discrete set of real values  $x_0 + ph$  ( $p = 0, \pm 1, \pm 2, \dots$ ), where  $x_0$  is a constant. By proper choice of origin  $x_0$  and scale one can make  $x_0 = 0, h = 1$ , so that  $x$  varies over the integers  $0, \pm 1, \pm 2, \dots$ . In the subsequent discussion, this simplification will be assumed made, so that  $x$  varies over the integers and the difference equation thus relates the values of  $f$  at  $x, x + 1, \dots, x + n$ . For the case when  $x$  varies continuously, see Remark at the end of Sect. 4. The values of  $f$  are assumed to be real, although much of the theory extends to the case in which  $f$  has complex values.

A general difference equation is constructed from a function  $\psi(x, y_0, y_1, \dots, y_n)$  of the integer variable  $x$  and the  $n + 1$  real variables  $y_0, \dots, y_n$ . The difference equation is the equation

$$(2) \quad \psi(x, f(x), f(x + 1), \dots, f(x + n)) = 0.$$

By a *solution* of the difference equation is meant a function  $f$  which satisfies it identically. When the equation takes the form

$$(3) \quad f(x + n) = \phi(x, f(x), \dots, f(x + n - 1))$$

and  $\phi(x, y_0, \dots, y_{n-1})$  is defined for all values of  $x, y_0, \dots, y_{n-1}$ , eq. (3) is simply a *recursion formula*. If  $f(0), f(1), \dots, f(n - 1)$  are given arbitrary values, then eq. (3) determines successively  $f(n), f(n + 1), \dots$ ; thus there is a unique solution for  $x \geq 0$  with the given initial values  $f(0), f(1), \dots, f(n - 1)$ .

The *first difference* of  $f(x)$  is  $\Delta f = f(x + 1) - f(x)$ ; the *second difference* is  $\Delta^2 f = \Delta(\Delta f) = f(x + 2) - 2f(x + 1) + f(x)$ ; the *kth difference* is  $\Delta^k f$ . A difference eq. (2) can be written in terms of  $f$  and its differences. For *example*, eq. (1) is equivalent to the equation

$$(1') \quad \Delta^3 f + 2\Delta^2 f + (1 - x)\Delta f - (3 + x)f = x^2.$$

Conversely, an equation relating  $f, \Delta f, \dots, \Delta^n f$  can be written in form (2). Thus, eq. (2) is the general form for difference equations, and this form will be used throughout this section, in preference to an equation relating the differences of  $f$ .

The *order* of the difference eq. (2) is defined as the distance between the most widely separated  $x$ -values at which the values of  $f$  are related. If  $\psi$  definitely depends on  $f(x)$  and  $f(x + n)$ , then the order is  $n$ . However, the order may be less than  $n$ . For *example*, the equation

$$(4) \quad f(x + 4) - 2f(x + 3) - f(x + 1) = 0$$

has order 3, since the most widely separated values are  $x + 1$  and  $x + 4$ . The substitution  $g(x) = f(x + 1)$  reduces this to an equation relating  $g(x), g(x + 2), g(x + 3)$ .

OPERATOR NOTATION. If  $y$  is a function of  $x$ , one writes

$$(5) \quad E^k y = y(x + k) \quad (k = 0, 1, 2, \dots).$$

Thus  $E^0 y = y(x)$ ,  $E^1 y = Ey = y(x + 1)$ . The difference eq. (2) can thus be written

$$(2') \quad \psi(x, y, Ey, E^2 y, \dots, E^n y) = 0.$$

## 2. LINEAR DIFFERENCE EQUATIONS

By a *linear difference equation* is meant an equation of form

$$(6) \quad a_n f(x+n) + a_{n-1} f(x+n-1) + \cdots + a_1 f(x+1) + a_0 f(x) = v(x),$$

where  $a_0, \dots, a_n, v(x)$  are given functions of the integer variable  $x$ . In terms of the operator  $E$  of Sect. 1, the equation can be written:

$$(6') \quad a_n E^n y + \cdots + a_1 E y + a_0 y = v(x),$$

where  $y = f(x)$ . It can be written more concisely as follows:

$$(7) \quad \psi(E)y = v(x),$$

where  $\psi(E)$  is a *linear difference operator*:

$$(8) \quad \psi(E) = a_n E^n + \cdots + a_1 E + a_0.$$

If  $v(x) \equiv 0$ , eq. (6) is termed *homogeneous*; otherwise it is *nonhomogeneous*. In case  $a_0 \neq 0, a_n \neq 0$ , the equation is of order  $n$ . In case  $a_0 \equiv a_1 \equiv \cdots \equiv a_{n-1} \equiv 0$ , but  $a_n a_n \neq 0$ , then it is of order  $n - m$ ; the substitution  $g(x) = f(x - m)$  then reduces eq. (6) to a linear equation for  $g$  of form (6), with nonvanishing first and last coefficients.

**Linear Independence.** Let  $y_1(x), \dots, y_p(x)$  be functions of  $x$  defined for  $a < x < b$ . The functions are said to be *linearly independent* if a relation

$$b_1 y_1(x) + \cdots + b_p y_p(x) \equiv 0$$

with constant  $b_1, \dots, b_p$ , can hold only if  $b_1 = b_2 = \cdots = b_p = 0$ . Otherwise, the functions are said to be *linearly dependent*.

**General Solution.** Let the difference eq. (6) be given, with  $v(x) \equiv 0$  and  $a_0(x)a_n(x) \neq 0$  for  $a < x < b$ ; all coefficients are assumed defined for  $a < x < b$ . Then the equation has order  $n$ , there are  $n$  linearly independent solutions  $y_1(x), \dots, y_n(x)$  for  $a < x < b$  and

$$(9) \quad y = c_1 y_1(x) + \cdots + c_n y_n(x), \quad a < x < b,$$

where  $c_1, \dots, c_n$  are arbitrary constants, is the general solution; that is, all solutions are given by eq. (9). If  $v(x) \neq 0$ , but the other hypotheses hold, then the general solution has form

$$(10) \quad y = c_1 y_1(x) + \cdots + c_n y_n(x) + V(x),$$

where  $V(x)$  is a solution of the nonhomogeneous equation and  $c_1 y_1(x) + \cdots + c_n y_n(x)$  is the general solution of the *related homogeneous equation*, that is, the homogeneous equation obtained by replacing  $v(x)$  by 0.

**EXAMPLE.** The functions  $y_1 \equiv 1, y_2 \equiv x$  are linearly independent solutions of the equation  $(E^2 - 2E + 1)y = 0$ , so that  $y = c_1 + c_2 x$  is the general solution;  $y = 2^{x-1}$  is a solution of the equation  $(E^2 - 2E + 1)y = 2^{x-1}$ , so that  $y = c_1 + c_2 x + 2^{x-1}$  is the general solution.

### 3. HOMOGENEOUS LINEAR EQUATIONS WITH CONSTANT COEFFICIENTS

The equations considered have form

$$(11) \quad \psi(E)y = 0,$$

where

$$(12) \quad \psi(E) = a_n E^n + \cdots + a_1 E + a_0,$$

the coefficients  $a_n, \cdots, a_0$  are constants, and  $a_0 a_n \neq 0$ . Associated with eq. (12) is the characteristic polynomial

$$\psi(\lambda) = a_n \lambda^n + \cdots + a_1 \lambda + a_0$$

in the complex variable  $\lambda$ . The equation

$$(13) \quad \psi(\lambda) = 0$$

is an algebraic equation of degree  $n$ , called the *characteristic equation* or *auxiliary equation* associated with eq. (11). The characteristic equation has  $n$  roots  $\lambda_1, \cdots, \lambda_n$  called *characteristic roots*. (See Chap. 2.) These may be real or complex; since the coefficients are assumed real, the complex roots come in conjugate pairs.

From the set of characteristic roots one obtains a set of  $n$  solutions of the difference eq. (11) by the following rules:

- I. To each simple real root  $\lambda$  one assigns the function  $\lambda^x$ ;
- II. To each real root  $\lambda$  of multiplicity  $k$  one assigns the  $k$  functions  $\lambda^x, x\lambda^x, \cdots, x^{k-1}\lambda^x$ ;
- III. To each pair of simple complex roots  $\alpha + \beta i = \rho(\cos \phi \pm i \sin \phi)$  one assigns the functions  $\rho^x \cos \phi x, \rho^x \sin \phi x$ ;
- IV. To each pair of complex roots  $\alpha \pm \beta i = \rho(\cos \phi \pm i \sin \phi)$  of multiplicity  $k$  one assigns the  $2k$  functions

$$\rho^x \cos \phi x, x\rho^x \cos \phi x, \cdots, x^{k-1}\rho^x \cos \phi x,$$

$$\rho^x \sin \phi x, x\rho^x \sin \phi x, \cdots, x^{k-1}\rho^x \sin \phi x.$$

In all one obtains  $n$  functions  $y_1(x), \cdots, y_n(x)$  which are linearly independent solutions of eq. (11) for all  $x$ , so that

$$y = c_1 y_1(x) + \cdots + c_n y_n(x)$$

is the general solution.

EXAMPLE.  $(E^4 - 8E^3 + 25E^2 - 36E + 20)y = 0$ .  
The characteristic equation is

$$\lambda^4 - 8\lambda^3 + 25\lambda^2 - 36\lambda + 20 = 0.$$

The roots are  $2, 2, 2 \pm i$ . Hence the general solution is

$$y = 2^x(c_1 + c_2x) + 5^{x/2}(c_3 \cos \phi x + c_4 \sin \phi x),$$

where  $\phi = \arctan \frac{1}{2}$ .

REMARKS. The variable  $x$  has heretofore assumed only integral values. If  $x$  is allowed to take on all real values, then the difference equation becomes a *functional equation*. The methods of this section are still applicable and provide the general solution of eq. (11) subject only to the following two modifications: (a) the arbitrary constants  $c_1, c_2, \dots$  may be replaced by arbitrary periodic functions of  $x$ , of period 1; (b) if  $\lambda$  is a negative characteristic root of multiplicity  $k$ , the corresponding solutions become  $(-\lambda)^x \cos \pi x, x(-\lambda)^x \cos \pi x, \dots, x^{k-1}(-\lambda)^x \cos \pi x$ .

#### 4. NONHOMOGENEOUS LINEAR EQUATIONS WITH CONSTANT COEFFICIENTS

The equation considered is

$$(14) \quad \psi(E)y = v(x),$$

where  $\psi(E)$  satisfies the same conditions as in Sect. 3. By the rule stated at the end of Sect. 2, the general solution of eq. (14) has the form

$$(15) \quad y = c_1y_1(x) + \dots + c_ny_n(x) + V(x),$$

where  $V(x)$  is a particular solution and the other terms are the general solution of the related homogeneous equation  $\psi(E)y = 0$ .

The procedures for finding the particular solution  $V(x)$  can be described concisely by means of an operational calculus which parallels that used for differential equations (Chap. 8). The operators  $\psi(E)$  with constant coefficients can be added, subtracted, multiplied, and multiplied by constants just as polynomials. The operators can be converted into operators  $\chi(\Delta)$  by the relation

$$(16) \quad \Delta = E - 1.$$

For *example*,

$$E^2 - 1 = (\Delta + 1)^2 - 1 = \Delta^2 + 2\Delta.$$

The powers  $\Delta, \Delta^2, \dots, \Delta^k$  are the first, second,  $\dots, k$ th difference, as defined in Sect. 1.

If  $y = V(x)$  is a solution of eq. (14), one writes

$$(17) \quad V(x) = \frac{1}{\psi(E)}v(x) = [\psi(E)]^{-1}v(x).$$

TABLE 1. RULES FOR PARTICULAR SOLUTIONS

No.	$\psi(E)$	$v(x)$	$[\psi(E)]^{-1}v$
1.	$\psi(E)$	$c_1v_1(x) + c_2v_2(x); c_1, c_2 \text{ const.}$	$c_1[\psi(E)]^{-1}v_1 + c_2[\psi(E)]^{-1}v_2$
2.	$\psi_1(E)\psi_2(E)$	$v(x)$	$\frac{1}{\psi_1(E)} \left( \frac{1}{\psi_2(E)} v \right)$
3.	$\Delta = E - 1$	$v(x), x = 0, \pm 1, \dots$	$\Delta^{-1}v = \sum_{k=0}^{x-1} v(k)$
4.	$\Delta^2 = E^2 - 2E + 1$	$v(x), x = 0, \pm 1, \dots$	$\Delta^{-2}v = \sum_{k=0}^{x-1} \sum_{s=0}^{k-1} v(s)$
5.	$\Delta^r = (E - 1)^r$	$\binom{x}{n} = \frac{\{x(x-1)\dots(x-n+1)\}}{n!}$	$\Delta^{-r} \binom{x}{n} = \binom{x}{n+r}$
6.	$\psi(E)$	$a^x u(x)$	$a^x \frac{1}{\psi(aE)} u(x)$
7.	$\psi(E)$	$a^x, \psi(a) \neq 0$	$\frac{a^x}{\psi(a)}$
8.	$(E - a)^k \phi(E)$	$a^x, \phi(a) \neq 0$	$\frac{x^k a^{x-k}}{\phi(a)k!}$
9.	$\psi(E)$	$a^x u(x), \psi(a) \neq 0, u$ a polynomial of degree $s$	$a^x [p(a) + \frac{a}{1!} p'(a)\Delta + \dots + \frac{a^s}{s!} p^s(a)\Delta^s] u(x),$ $p(\lambda) = 1/\psi(\lambda)$
10.	$(E - a)^k \phi(E)$ $\phi(a) \neq 0$	$a^x u(x), u(x)$ a polynomial of degree $s$	$a^{x-k} [q(a)\Delta^{-k} + \frac{a}{1!} q'(a)\Delta^{1-k} + \dots + \frac{a^s}{s!} q^s(a)\Delta^{s-k}] u(x),$ $q(\lambda) = 1/\phi(\lambda)$
11.	$E - a$	$v(x)$	$a^{x-1} \Delta^{-1}(a^{-x}v)$
12.	$(E - a)^k$	$v(x)$	$a^{x-k} \Delta^{-k}(a^{-x}v)$
13.	$(E - \alpha)^2 + \beta^2$ $\alpha + i\beta = p(\cos \phi + i \sin \phi)$	$v(x)$	$(p^{x-1}/\beta) [\sin(\phi x - \phi) \cdot \Delta^{-1}(p^{-x} \cos \phi x v) - \cos(\phi x - \phi) \cdot \Delta^{-1}(p^{-x} \sin \phi x v)]$

Thus the inverse operator  $[\psi(E)]^{-1}$ , when applied to  $v(x)$ , yields one solution of the eq. (14). The rules for finding particular solutions can now be summarized in a table, which evaluates  $[\psi(E)]^{-1}v$  for various choices of  $\psi$  and  $v$ . This is carried out in Table 1. The last column gives one choice  $V(x)$  of  $[\psi(E)]^{-1}v$ ; the general solution is given by eq. (15).

The binomial coefficient  $\binom{x}{n}$  of Rule 5, Table 1, is defined for  $n = 1, 2, \dots$ . When  $n = 0$ , it is defined to equal 1, and Rule 5 remains valid. Corresponding to this inverse rule is the direct rule:

$$(18) \quad \Delta^r \binom{x}{n} = \binom{x-r}{n}, \quad 0 < r \leq n \\ = 0, \quad r > n.$$

A general power of  $x$  can be expanded in terms of these coefficients:

$$(19) \quad x^n = T_n^{-1}(1!)\binom{x}{1} + \dots + T_n^n(n!)\binom{x}{n},$$

where the  $T_n^j$  are *Stirling numbers of the second kind*. They are tabulated on page 170 of Ref. 2. If the polynomial  $u(x)$  is expanded in terms of the coefficients by eq. (19) and Rule 5 or eq. (18) is applied, then Rules 9, 10 are easier to use.

A general expression  $1/\psi(E)$  can be regarded as a rational function of  $E$  and expanded in partial fractions, just as if  $E$  were a numerical variable. Rules 11, 12, 13 then permit evaluation of the terms. For *example*,

$$\frac{1}{E^2 - 3E + 2} v(x) = \frac{1}{(E-1)(E-2)} v(x) = \left( \frac{1}{E-2} - \frac{1}{E-1} \right) v(x) \\ = 2^{x-1} \Delta^{-1}(2^{-x}v) - \Delta^{-1}v.$$

Rule 12 is needed for multiple roots. Rule 13 is needed for complex roots; it can be generalized to take care of repeated complex roots (Ref. 2).

## 5. LINEAR EQUATIONS WITH VARIABLE COEFFICIENTS

The general solution of the first order linear equation

$$(20) \quad [E - p(x)]y = v(x), \quad x \geq a,$$

where  $p(x) \neq 0$  for  $x \geq a$ , is

$$(21) \quad y = q(x) \left[ c + \sum_{s=a}^{x-1} \frac{v(s)}{q(s+1)} \right],$$

$$(22) \quad q(x) = p(a)p(a+1) \cdots p(x-1), \quad x > a \\ = 1, \quad x = a.$$

**Laplace's Method.** For equations

$$(23) \quad [a_n(x)E^n + \cdots + a_1(x)E + a_0(x)]y = 0$$

with polynomial coefficients, one seeks a solution

$$(24) \quad y(x) = \int_a^b t^{x-1}v(t) dt,$$

where  $a$ ,  $b$ , and  $v(t)$  are to be determined.

Let  $(x)_n$  denote  $n!(\binom{x}{n})$ , so that  $(x)_n = x(x-1)\cdots(x-n+1)$  for  $n = 1, 2, \dots$ , and let  $(x)_0 = 1$ . It follows from eq. (19) that an arbitrary polynomial can be expressed as a linear combination of the polynomials  $(x)_n$ . Hence the coefficients  $a_k(x)$  can be considered as linear combinations of the  $(x)_n$ . Now by integration by parts one obtains from eq. (24) the relation

$$(25) \quad (x+m-1)_m E^x y \\ = \left[ \sum_{s=1}^m (-1)^{s+1} (x+m-1)_{m-s} t^{x-1+s} D^{s-1} \{t^n v(t)\} \right]_a^b \\ + (-1)^m \int_a^b t^{x+m-1} D^m \{t^n v(t)\} dt,$$

where  $D^s = d^s/dt^s$ . Hence the difference eq. (23) takes the form

$$(26) \quad [F(x, v, t)]_a^b + \int_a^b t^{x-1} G(v, t) dt = 0.$$

The function  $v(t)$  is chosen so that  $G(v, t) \equiv 0$ . In fact, the equation  $G(v, t) = 0$  is usually a homogeneous linear differential equation for  $v(t)$ . The constants  $a$  and  $b$  are then chosen so that  $F(x, v, t)$  vanishes when  $t = a$  and  $t = b$ , so that eq. (26) is satisfied. Once  $a$ ,  $b$  and  $v(t)$  have been determined in this way, eq. (24) then yields  $y(x)$ .

For further details see Ref. 2, Sect. 174.

### REFERENCES

1. T. Fort, *Finite Differences*, Oxford University Press, Oxford, England, 1948.
2. C. Jordan, *Calculus of Finite Differences*, 2nd edition, Chelsea, New York, 1950.
3. N. E. Nörlund, *Vorlesungen über Differenzenrechnung*, Springer, Berlin, 1924.

## Differential Equations

G. E. Hay and W. Kaplan

1. Basic Concepts	5-01
2. Equations of First Order and First Degree	5-02
3. Linear Differential Equations	5-04
4. Equations of First Order but not of First Degree	5-07
5. Special Methods for Equations of Higher than First Order	5-09
6. Solutions in Form of Power Series	5-10
7. Simultaneous Linear Differential Equations	5-12
8. Numerical Methods	5-14
9. Graphical Methods—Phase Plane Analysis	5-15
10. Partial Differential Equations	5-20
References	5-22

### 1. BASIC CONCEPTS

An *ordinary differential equation* is an equation of form

$$(1) \quad \psi(x, y, y', \dots, y^{(n)}) = 0,$$

expressing a relationship between an unspecified function  $y$  of  $x$  and its derivatives  $y' = dy/dx, \dots, y^{(n)} = d^n y/dx^n$ . An example is the following:

$$y' - xy = 0.$$

The *order* of the equation is  $n$ , which is the order of the highest derivative appearing. A *solution* is a function  $y$  of  $x$ ,  $a < x < b$ , which satisfies the

equation identically. For many equations one can obtain a function

$$(2) \quad y = f(x, c_1, \dots, c_n),$$

expressing  $y$  in terms of  $x$  and  $n$  independent arbitrary constants  $c_1, \dots, c_n$  such that, for each choice of the constants, eq. (2) is a solution of eq. (1), and every solution of eq. (1) is included in eq. (2). When these conditions are satisfied, eq. (2) is called the *general solution* of eq. (1). A *particular solution* is the general solution with all of the  $n$  arbitrary constants given particular values.

If eq. (1) is an algebraic equation in  $y^{(n)}$  of degree  $k$ , then the differential eq. (1) is said to have degree  $k$ . For example, the equation

$$(3) \quad y''''2 + y''y''' + y^4 = e^x$$

has order 3 and degree 2. When the degree is 1, the equation has the form

$$(4) \quad p(x, y, \dots, y^{(n-1)})y^{(n)} + q(x, y, \dots, y^{(n-1)}) = 0$$

or, where  $p \neq 0$ , the equivalent form

$$(5) \quad y^{(n)} = F(x, y, \dots, y^{(n-1)}), \quad F = -q/p.$$

The EXISTENCE THEOREM asserts that, if in eq. (5)  $F$  is continuous in an open region  $R$  of the space of the variables  $x, y, \dots, y^{(n-1)}$ , and  $(x_0, y_0, \dots, y_0^{(n-1)})$  is a point of  $R$ , then there exists a solution  $y(x)$  of eq. (5),  $|x - x_0| < h$ , such that

$$(6) \quad y = y_0, y' = y'_0, \dots, y^{(n-1)} = y_0^{(n-1)} \quad \text{for } x = x_0.$$

Thus there exists a solution satisfying *initial conditions* (6). If  $F$  has continuous partial derivatives with respect to  $y, y', \dots, y^{(n-1)}$  in  $R$ , then the solution is unique.

## 2. EQUATIONS OF FIRST ORDER AND FIRST DEGREE

An equation of first order and first degree can be written in either of the equivalent forms

$$(7) \quad y' = F(x, y),$$

$$(8) \quad M(x, y) dx + N(x, y) dy = 0.$$

For equations of special form, explicit rules can be given for finding the general solution. Some of the most important types are listed here.

**Equations with Variables Separable.** If in eq. (8)  $M$  depends only on  $x$ ,  $N$  only on  $y$ , then eq. (8) is said to have the *variables separable*. The equation may then be written with the  $x$ 's separated from the  $y$ 's, and the general solution may be obtained by integration.

EXAMPLE.  $y' = 3x^2y$ . An equivalent separated form is  $3x^2 dx - y^{-1} dy = 0$ . Hence

$$\int 3x^2 dx - \int y^{-1} dy = c.$$

Integrating and solving for  $y$  one finds  $y = c_1 e^{x^3}$  as the general solution, where  $c_1 = e^{-c}$ .

**Homogeneous Equations.** A function  $F(x, y)$  is said to be homogeneous of degree  $n$  if  $F(\lambda x, \lambda y) \equiv \lambda^n F(x, y)$ . The differential eq. (7) is said to be homogeneous if  $F(x, y)$  is homogeneous of degree 0. To solve such a differential equation write  $y = vx$  and express the differential equation in terms of  $v$  and  $x$ . The resulting differential equation has variables separable and can be solved as above. In general,  $y' = F(x, y)$  becomes

$$xv' + v = F(x, vx) = x^0 F(1, v) = G(v),$$

$$\frac{dx}{x} + \frac{dv}{v - G(v)} = 0.$$

**Exact Equations.** The differential eq. (8) is exact if for some function  $u(x, y)$

$$(9) \quad \frac{\partial u}{\partial x} = M(x, y), \quad \frac{\partial u}{\partial y} = N(x, y),$$

so that  $du = M dx + N dy$ . The equation is exact if and only if  $\partial M/\partial y \equiv \partial N/\partial x$ . The general solution is given (implicitly) by  $u(x, y) = c$ .

EXAMPLE.  $(3x^2 - 2xy) dx + (2y - x^2) dy = 0$ . Here  $\partial M/\partial y = -2x = \partial N/\partial x$ , so that the equation is exact. Then

$$\frac{\partial u}{\partial x} = 3x^2 - 2xy, \quad \frac{\partial u}{\partial y} = 2y - x^2.$$

From the first equation,  $y = x^3 - x^2y + g(y)$ , where  $g(y)$  is an arbitrary function of  $y$ . Substitution in the second equation yields the relation  $-x^2 + g'(y) = 2y - x^2$ , so that  $g(y) = y^2 + c$ . Hence the general solution is  $x^3 - x^2y + y^2 = c$ .

**Integrating Factors.** If the eq. (8) is not exact, it may be possible to make it exact by multiplying by a function  $\phi(x, y)$ , called an *integrating factor*.

EXAMPLE. The equation  $(3xy + 2y^2) dx + (x^2 + 2xy) dy = 0$  is not exact, but after multiplication by  $x$  becomes the exact equation

$$(3x^2y + 2xy^2) dx + (x^3 + 2x^2y) dy = 0.$$

The general solution is  $x^3y + x^2y^2 = c$ . The integrating factor is  $x$ .

**Linear Equations.** A differential equation is *linear* if it is of the first degree in the dependent variable and its derivatives. If such an equation is also of the first order, it may be written in the form

$$(10) \quad y' + p(x)y = q(x).$$

Here  $u = e^{\int p dx}$  is an integrating factor and the general solution is

$$(11) \quad y = u^{-1} \left( \int Q(x)u dx + c \right), \quad u = e^{\int p dx}.$$

EXAMPLE.  $y' + x^{-1}y = 4x^2$ . Here  $u = x$  and eq. (11) gives

$$y = x^{-1} \left( \int 4x^3 dx + c \right) = x^3 + cx^{-1}.$$

### 3. LINEAR DIFFERENTIAL EQUATIONS

The linear differential equation of order  $n$  can be written in the form

$$(12) \quad a_0 D^n y + a_1 D^{n-1} y + \cdots + a_{n-1} D y + a_n y = Q(x),$$

where the coefficients  $a_0, \dots, a_n$  may depend on  $x$ , and  $D^k y \equiv d^k y / dx^k$ . When the  $a_j$  are constant, eq. (12) is said to have *constant coefficients*. When  $Q(x) \equiv 0$ , the equation is said to be *homogeneous*. The homogeneous equation obtained from eq. (12) by replacing  $Q(x)$  by 0 is called the *related homogeneous equation*. It will generally be assumed that  $a_0 \neq 0$  throughout the interval of  $x$  considered.

The general solution of eq. (12) is given by

$$(13) \quad y = c_1 y_1(x) + \cdots + c_n y_n(x) + y^*(x),$$

where  $y^*(x)$  is one particular solution and  $y_1(x), \dots, y_n(x)$  are particular solutions of the related homogeneous equation which are *linearly independent*; that is, a relation

$$b_1 y_1(x) + b_2 y_2(x) + \cdots + b_n y_n(x) \equiv 0,$$

with constant  $b_1, \dots, b_n$  can hold only if  $b_1 = 0, \dots, b_n = 0$ . When  $Q(x) \equiv 0$ , one can choose  $y^*(x)$  to be 0.

**Homogeneous Linear Equations with Constant Coefficients.** The equation has the form

$$(14) \quad a_0 D^n y + \cdots + a_{n-1} D y + a_n y = 0, \quad a_0 \neq 0,$$

where  $a_0, \dots, a_n$  are constants. Particular solutions are obtained by setting  $y = e^{rx}$ . Substitution in eq. (14) leads to the equation for  $r$ :

$$(15) \quad a_0 r^n + \cdots + a_{n-1} r + a_n = 0.$$

This is called the *auxiliary equation* or *characteristic equation*. In general it has  $n$  roots, real or complex, some of which may be coincident (Chap. 2). From these roots one obtains  $n$  linearly independent solutions of the differential eq. (14) by the following rules:

I. To each real root  $r$  of multiplicity  $k$  one assigns the functions  $e^{kx}$ ,  $xe^{kx}$ ,  $\dots$ ,  $x^{k-1}e^{kx}$ .

II. To each pair of conjugate complex roots  $\alpha \pm \beta i$  of multiplicity  $k$  one assigns the  $2k$  functions

$$e^{\alpha x} \cos \beta x, e^{\alpha x} \sin \beta x, xe^{\alpha x} \cos \beta x, xe^{\alpha x} \sin \beta x, \\ \dots, x^{k-1}e^{\alpha x} \cos \beta x, x^{k-1}e^{\alpha x} \sin \beta x.$$

The  $n$  function  $y_1(x), \dots, y_n(x)$  thus obtained are linearly independent and

$$y = c_1y_1(x) + \dots + c_ny_n(x)$$

is the general solution of eq. (14).

EXAMPLE 1.  $D^2y - 3Dy + 2y = 0$ . The auxiliary equation is  $r^2 - 3r + 2 = 0$ , the roots are 1, 2; the general solution is  $y = c_1e^x + c_2e^{2x}$ .

EXAMPLE 2.  $D^6y - 9D^4y + 24D^2y - 16y = 0$ . The auxiliary equation is  $r^6 - 9r^4 + 24r^2 - 16 = 0$ , the roots are  $\pm 1, \pm 2, \pm 2i$ ; the general solution is  $y = c_1e^x + c_2e^{-x} + e^{2x}(c_3 + c_4x) + e^{-2x}(c_5 + c_6x)$ .

EXAMPLE 3.  $D^4y + 4D^3y + 12D^2y + 16Dy + 16 = 0$ . The auxiliary equation is  $r^4 + 4r^3 + 12r^2 + 16r + 16 = 0$ , the roots are  $-1 \pm i\sqrt{3}, -1 \pm i\sqrt{3}$ . The general solution is  $y = e^{-x}[(c_1 + c_2x) \cos \sqrt{3}x + (c_3 + c_4x) \sin \sqrt{3}x]$ .

### Nonhomogeneous Linear Equations with Constant Coefficients.

The equation considered is

$$(16) \quad a_0D^n + \dots + a_{n-1}Dy + a_ny = Q(x), \quad a_0 \neq 0,$$

where  $a_0, \dots, a_n$  are constants and  $Q(x)$  is, for example, continuous for  $a < x < b$ . The general solution of the related homogeneous equation is found as in the preceding paragraphs; it is called the *complementary function*. Here are presented methods for finding a particular solution  $y^*(x)$  of eq. (16). As indicated in eq. (13), addition of the complementary function and  $y^*(x)$  gives the required general solution of eq. (16).

**Method of Undetermined Coefficients.** If  $Q(x)$  is of form

$$(17) \quad e^{\alpha x}[p(x) \cos \beta x + q(x) \sin \beta x],$$

where  $p(x)$  and  $q(x)$  are polynomials of degree at most  $h$ , then there is a particular solution

$$(18) \quad y^* = x^k e^{\alpha x}[\phi(x) \cos \beta x + \psi(x) \sin \beta x],$$

where  $\phi(x)$  and  $\psi(x)$  are polynomials of degree at most  $h$  and  $\alpha \pm \beta i$  is a root of multiplicity  $k$  (possibly 0) of the auxiliary equation. If  $\beta = 0$ ,  $Q$  is of form  $e^{\alpha x}p(x)$ ; also  $p$  and  $q$  may reduce to constants ( $h = 0$ ). The coefficients of the polynomials  $\phi$ ,  $\psi$  can be considered as undetermined coefficients; substitution of eq. (18) in eq. (16) leads to relations between these coefficients from which all can be determined. As an *example* consider the equation

$$(D^2 + 1)y = 3 \cos x.$$

Here  $\alpha = 0$ ,  $\beta = 1$ ,  $h = 0$ . Since  $\pm i$  are roots of the auxiliary equation,  $k = 1$  and

$$y^* = x(A \cos x + B \sin x).$$

Substitution in the differential equation leads to the relation

$$2(-A \sin x + B \cos x) \equiv 3 \cos x.$$

Hence  $B = \frac{3}{2}$ ,  $A = 0$ ;  $y^* = \frac{3}{2}x \sin x$  and the general solution is

$$y = \frac{3}{2}x \sin x + c_1 \cos x + c_2 \sin x.$$

**Superposition Principle.** If in eq. (16)  $Q(x)$  is a linear combination of functions  $Q_1(x), \dots, Q_N(x)$  and  $y_1^*(x), \dots, y_N^*(x)$  are particular solutions of the respective equations obtained by replacing  $Q(x)$  by  $Q_1(x), \dots, Q_N(x)$ , then the corresponding linear combination of  $y_1^*(x), \dots, y_N^*(x)$  is a solution of eq. (16); that is, if

$$Q(x) = b_1 Q_1(x) + \dots + b_N Q_N(x),$$

then

$$y^*(x) = b_1 y_1^*(x) + \dots + b_N y_N^*(x)$$

is a particular solution of eq. (16). For *example*, particular solutions of

$$(D^2 + 1)y = 3 \cos x, \quad (D^2 + 1)y = 5e^{2x}$$

are found by undetermined coefficients to be  $\frac{3}{2}x \sin x$ ,  $e^{2x}$  respectively. Hence a particular solution of

$$(D^2 + 1)y = 12 \cos x + 10e^{2x}$$

is given by  $6x \sin x + 2e^{2x}$ .

**Variation of Parameters.** Let the complementary function be

$$y = c_1 y_1(x) + \dots + c_n y_n(x).$$

Then a particular solution is

$$(19) \quad y^* = v_1(x)y_1(x) + \dots + v_n(x)y_n(x),$$

where

$$v_1(x) = \int w_1(x) dx, \dots, v_n(x) = \int w_n(x) dx$$

and  $w_1(x), \dots, w_n(x)$  are defined by the linear equations

$$\begin{aligned} & y_1(x)w_1(x) + \dots + y_n(x)w_n(x) = 0, \\ & y'_1(x)w_1(x) + \dots + y'_n(x)w_n(x) = 0, \\ & \dots, \\ & y_1^{(n-1)}(x)w_1(x) + \dots + y_n^{(n-1)}(x)w_n(x) = Q(x)/a_0. \end{aligned} \tag{20}$$

The determinant of coefficients of eqs. (20) is the *Wronskian determinant*

$$W = \begin{vmatrix} y_1(x) & \dots & y_n(x) \\ y'_1(x) & \dots & y'_n(x) \\ \dots & \dots & \dots \\ y_1^{(n-1)}(x) & \dots & y_n^{(n-1)}(x) \end{vmatrix}. \tag{21}$$

Under the assumptions made,  $W$  cannot equal 0 for any  $x$  of the interval considered, so that eqs. (20) have a unique solution (Chap. 2). This method is applicable if  $a_0, \dots, a_n$  are functions of  $x$ , provided  $a_0(x) \neq 0$ .

**Operational Methods.** The operational methods based on the *Heaviside calculus* provide another powerful tool for obtaining solutions of non-homogeneous linear equations with constant coefficients (see Chap. 8). Closely related are the methods based on the Laplace transform (Chap. 9).

#### 4. EQUATIONS OF FIRST ORDER BUT NOT OF FIRST DEGREE

The equations considered have form

$$\psi(x, y, p) = 0, \tag{22}$$

where  $p = dy/dx$ . Equation (22) can be solved for  $p$ , except where  $\psi_p = 0$ . The locus defined by the two equations

$$\psi(x, y, p) = 0, \quad \psi_p(x, y, p) = 0 \tag{23}$$

is called the *singular locus*. It may contain curves  $y = f(x)$  which are solutions of eq. (22); such solutions are called *singular solutions*. The solutions of eq. (22) (with the possible exception of the singular solutions) can often be obtained by one of the following special methods.

**Factorization.** If eq. (22) can be factored in the form

$$[p - F_1(x, y)][p - F_2(x, y)] \dots [p - F_k(x, y)] = 0, \tag{24}$$

then its solutions are obtained by combining all solutions of the first degree equations

$$(25) \quad p = F_1(x, y), \dots, p = F_k(x, y) \quad (p = dy/dx).$$

For *example*, the equation

$$p^2 - (2x + y)p + 2xy = 0$$

can be factored into the equations

$$p = 2x, \quad p = y;$$

the solutions are  $y = x^2 + c_1$ ,  $y = c_2e^x$ . If  $\psi(x, y, p)$  is of second degree in  $p$ , the expressions for  $p$  and the equivalent factorization (24) can be obtained by the quadratic formula.

*Solving for y or x.* If eq. (22) is of first degree in  $y$ , one can solve for  $y$  to obtain an equation

$$(26) \quad y = F(x, p).$$

Differentiation of this equation with respect to  $x$  yields a relation of form

$$(27) \quad \frac{dp}{dx} = G(x, p),$$

that is, a first order equation relating  $p$  and  $x$ . If the general solution of eq. (27) is given by

$$(28) \quad \phi(x, p) = c,$$

then the equations

$$(29) \quad y = F(x, p), \quad \phi(x, p) = c$$

together define solutions of eq. 22;  $p$  may be eliminated between the equations or treated as a parameter. As an *example*, consider the *Clairaut equation*:

$$(30) \quad y = xp + F(p).$$

The method described leads to the "general solution"

$$(31) \quad y = cx + F(c).$$

There is, in general, a *singular* solution defined by the equations

$$(32) \quad x + F'(p) = 0, \quad y = xp + F(p).$$

If the eq. (22) is solvable for  $x$ , one can differentiate with respect to  $y$ , re-

placing  $dx/dy$  by  $1/p$ ; one obtains the solutions in the form

$$(33) \quad \phi(y, p) = c, \quad y = F(x, p).$$

### 5. SPECIAL METHODS FOR EQUATIONS OF HIGHER THAN FIRST ORDER

**Equations with Dependent Variable Missing.** Let the given equation be

$$(34) \quad F(x, y', y'', \dots, y^{(n)}) = 0,$$

so that  $y$  does not appear. Set  $p = dy/dx$ . Then

$$y'' = \frac{dp}{dx}, \quad y''' = \frac{d^2p}{dx^2}, \dots,$$

and so eq. (34) becomes

$$(34') \quad F\left(x, p, \frac{dp}{dx}, \dots, \frac{d^{n-1}p}{dx^{n-1}}\right) = 0,$$

an equation of order  $n - 1$  for  $p$  in terms of  $x$ . If its solutions are known, then the solutions of eq. (34) are obtained from the relation  $y = \int p dx$ .

**EXAMPLE.** Consider the equation

$$x^3 y'' - x^2 y' = 3 + x^2.$$

The substitution  $p = y'$  leads to the first order linear equation

$$x^3 \frac{dp}{dx} - x^2 p = 3 + x^2.$$

Its general solution is found (Sect. 2) to be

$$p = -\frac{1}{x^2} + 1 + c_1 x.$$

Hence integration yields  $y$ :

$$y = \frac{1}{x} + x + \frac{1}{2} c_1 x^2 + c_2.$$

**Equations with Independent Variable Missing.** Let the given equation be the  $n$ th order equation

$$(35) \quad F(y, y', y'', \dots, y^{(n)}) = 0,$$

so that  $x$  does not appear. Set  $p = y'$ . Then

$$\frac{d^2y}{dx^2} = \frac{dp}{dx} = \frac{dp}{dy} \frac{dy}{dx} = p \frac{dp}{dy},$$

$$\frac{d^3y}{dx^3} = p^2 \frac{d^2p}{dy^2} + \left(\frac{dp}{dy}\right)^2, \dots$$

Thus eq. (35) becomes an equation of order  $n - 1$ . If its solutions are known, in the form

$$p = \phi(y, c_1, \dots, c_{n-1}),$$

then

$$\frac{dy}{dx} = \phi, \quad \frac{dy}{\phi(y, c_1, \dots)} = dx.$$

Thus integration yields an implicit form of the solutions of the given equation.

**Linear Equations with One Known Solution.** Let a linear equation be given:

$$(36) \quad a_0(x)y^{(n)}(x) + \dots + a_n(x)y = Q(x).$$

Let  $y_1(x)$  be a solution of the related homogeneous equation. Then the substitutions

$$(37) \quad y = y_1(x)v, \quad w = v'$$

leads to an equation of order  $n - 1$  for  $w$ . If  $w$  has been found, integration and multiplication by  $y_1(x)$  yields  $y$ .

## 6. SOLUTIONS IN FORM OF POWER SERIES

**Formation of Taylor Series.** Let an equation of order  $n$  be given:

$$(38) \quad y^{(n)} = F(x, y, y', \dots, y^{(n-1)})$$

and let  $F$  be expressible in an absolutely convergent power series in powers of  $x, y, y', \dots$  for  $|x| < a, |y| < b_1, \dots, |y^{(n-1)}| < b_n$ , so that  $F$  is an analytic function of the  $n + 1$  variables. Then the solution  $f(x)$  of eq. (38) with initial conditions:  $y = 0, \dots, y^{(n-1)} = 0$  at  $x = 0$  is expressible as a power series in  $x$  for  $|x| < \rho$ , provided  $\rho$  is sufficiently small. The series is the *Taylor series* of  $f(x)$ :

$$(39) \quad f(x) = f(0) + xf'(0) + \dots + x^k \frac{f^{(k)}(0)}{k!} + \dots$$

The values of  $f(0), \dots, f^{(n-1)}(0)$  are given to be 0. The values of  $f^{(n)}(0),$

$f^{(n+1)}(0), \dots$  are obtained from the differential equation:

$$\begin{aligned} f^{(n)}(0) &= F(0, 0, \dots, 0), \\ f^{(n+1)}(x) &= F_x + F_y y' + \dots + F_{y^{(n-1)}} y^{(n)}, \dots, \\ f^{(n+1)}(0) &= F_x(0, \dots, 0) + F_{y^{(n-1)}} F(0, \dots, 0), \dots \end{aligned}$$

If the initial value  $x_0$  of  $x$  is not 0, one can introduce a new independent variable  $x_1 = x - x_0$ . If the initial values  $y_0, \dots$ , of  $y, \dots, y^{(n-1)}$  are not 0, one can introduce a new dependent variable

$$u = y - \left( y_0 + y'_0 x + \dots + y_0^{(n-1)} \frac{x^{n-1}}{(n-1)!} \right);$$

then the initial values of  $u, u', \dots, u^{(n-1)}$  are 0.

**Special Methods for Linear Equations.** For simplicity attention will be restricted to the second order equation

$$(40) \quad y'' + p(x)y' + q(x)y = 0.$$

If  $p(x)$  and  $q(x)$  are analytic at  $x = 0$ , then  $x = 0$  is called an *ordinary point* of eq. (40). The Taylor series solution eq. (39) can then be obtained as above. One can also substitute a series

$$(41) \quad y = \sum_{n=0}^{\infty} c_n x^n$$

in eq. (40) and obtain conditions on the coefficient  $c_n$ . In general  $c_0, c_1$  turn out to be arbitrary constants and the other coefficients are expressible in terms of these two with the aid of a *recursion formula*. The series eq. (41) converges and represents a solution of eq. (40) for  $|x| < a$ , provided  $p(x)$  and  $q(x)$  are analytic functions of the *complex* variable  $x$  for  $|x| < a$ ; see Chap. 7.

If  $p(x)$  and  $q(x)$  are not both analytic at  $x = 0$ , but  $xp(x)$  and  $x^2q(x)$  are analytic at  $x = 0$ , then  $x = 0$  is called a *regular singular point* of Eq. (40). In this case a solution is sought in the form

$$(42) \quad y = \sum_{n=0}^{\infty} c_n x^{n+k},$$

where the  $c$ 's and  $k$  are constants. If one substitutes in eq. (40) and equates to zero the coefficients of the various powers of  $x$ , one obtains a set of equations relating the  $c$ 's and  $k$ . The coefficient of the lowest power of  $x$  gives rise to a quadratic equation in  $k$ , called the *indicial equation*. Four

cases arise, depending on whether the two roots of this equation do not differ by an integer, are equal, or differ by an integer. In two of these cases, the solution consists partly of power series multiplied by  $\log x$ . For further details and *examples* of the important applications, one is referred to Chap. 7, Sect. 9. (See also Refs. 3, 9, 11.)

## 7. SIMULTANEOUS LINEAR DIFFERENTIAL EQUATIONS

Attention will be restricted to systems of first order equations:

$$(43) \quad \frac{dy_i}{dx} = \sum_{j=1}^n a_{ij}(x)y_j + Q_i(x), \quad i = 1, \dots, n.$$

A variety of other systems can be reduced to this form by appropriate substitutions. It will be assumed that the  $a_{ij}(x)$  and  $Q_1(x), \dots, Q_n(x)$  are continuous for  $a \leq x \leq b$ .

By a *solution* of eqs. (43) is meant an  $n$ -tuple of functions

$$(44) \quad y_1 = f_1(x), \dots, y_n = f_n(x), \quad a \leq x \leq b$$

which together satisfy eqs. (43) identically. The general solution of eqs. (43) can be shown to have form

$$(45) \quad y_i = c_1 f_{1i}(x) + c_2 f_{2i}(x) + \dots + c_n f_{ni}(x) + g_i(x);$$

here  $i = 1, \dots, n$ , the  $c$ 's are arbitrary constants;  $y_i = g_i(x)$  ( $i = 1, \dots, n$ ) defines one solution of eqs. (53); for each  $j$  the functions  $f_{ji}(x)$  ( $i = 1, \dots, n$ ) define a solution of the *related homogeneous system*

$$(46) \quad \frac{dy_i}{dx} = \sum_{j=1}^n a_{ij}y_j, \quad i = 1, \dots, n,$$

and these solutions of eqs. (46) are *linearly independent*; that is,

$$b_1 f_{1i}(x) + b_2 f_{2i}(x) + \dots + b_n f_{ni}(x) \equiv 0, \quad i = 1, \dots, n$$

implies  $b_1 = 0, \dots, b_n = 0$ .

**Homogeneous Systems with Constant Coefficients.** Consider the homogeneous system eq. (46) in which the  $a_{ij}$  are constants. Particular solutions are obtained by setting

$$(47) \quad y_1 = \alpha_1 e^{\lambda x}, \dots, y_n = \alpha_n e^{\lambda x}.$$

Equations (47) define a solution provided  $\lambda$  satisfies the *characteristic equation*

$$(48) \quad \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0.$$

If  $\lambda$  satisfies eq. (48), then values of  $\alpha_1, \dots, \alpha_n$  not all 0 can be found such that

$$(49) \quad \sum_{j=1}^n (a_{ij} - \lambda \delta_{ij}) \alpha_j = 0 \quad (i = 1, \dots, n),$$

where  $\delta_{ij} = 0$  if  $i \neq j$ ,  $= 1$  if  $i = j$ . With these values of the  $\alpha_i$ , eq. (47) defines a solution of the system (46). If the characteristic equation has distinct real roots  $\lambda_1, \dots, \lambda_n$ , then one obtains  $n$  linearly independent solutions in this manner, so that the general solution is obtained in form (45), with the  $g_i(x)$  replaced by 0. If  $\lambda$  is a repeated root of multiplicity  $k$ , one replaces eq. (47) by

$$(50) \quad y_1 = p_1(x)e^{\lambda x}, \dots, y_n = p_n(x)e^{\lambda x},$$

where  $p_1(x), \dots, p_n(x)$  are polynomials of degree at most  $k - 1$ ; one can obtain  $k$  linearly independent solutions in this form, which can be used to build the general solution. The procedure is easily modified to take care of complex roots (Refs. 1, 3).

The methods described are more easily formulated in terms of matrices (Chap. 3, and Ref. 3).

**Nonhomogeneous Systems.** If the general solution of the related homogeneous system is known:

$$(51) \quad y_i = c_1 f_{1i}(x) + \cdots + c_n f_{ni}(x) \quad (i = 1, \dots, n),$$

then one can obtain a particular solution by the method of *variation of parameters*. One replaces  $c_1, \dots, c_n$  in eq. (51) by unknown functions  $v_1(x), \dots, v_n(x)$ . Substitution in eqs. (43) leads to the equations

$$(52) \quad \sum_{j=1}^n f_{ji} \frac{dv_j}{dx} = Q_i(x), \quad i = 1, \dots, n,$$

which can be solved for the functions  $dv_j/dx$ ; integration then yields  $v_1(x)$ ,

$\dots, v_n(x)$  and the solution

$$(53) \quad y_i = v_1(x)f_{1i}(x) + \dots + v_n(x)f_{ni}(x), \quad i = 1, \dots, n.$$

For systems with constant coefficients one can also employ *operational methods* (Chaps. 8, 9).

## 8. NUMERICAL METHODS

**Method of Picard.** Let a system

$$(54) \quad \frac{dy_i}{dx} = F_i(x, y_1, \dots, y_n), \quad i = 1, \dots, n,$$

be given and let a solution

$$(55) \quad y_1 = f_1(x), \dots, y_n = f_n(x), \quad a \leq x \leq b,$$

be sought with initial values

$$(56) \quad f_1(x_0) = k_1, \dots, f_n(x_0) = k_n, \quad a < x_0 < b.$$

One forms successive approximations to the desired solution by the formulas:

$$f_1^1(x) \equiv k_1, \dots, f_n^1(x) \equiv k_n,$$

$$f_i^2(x) = \int_{x_0}^x F_i(t, k_1, \dots, k_n) dt + k_i, \quad i = 1, \dots, n,$$

and in general

$$f_i^{m+1}(x) = \int_{x_0}^x F_i(t, f_1^m(t), \dots, f_n^m(t)) dt + k_i,$$

for  $i = 1, \dots, n$ . Under appropriate hypotheses (see Ref. 3) the sequences  $f_i^m(x)$  ( $i = 1, \dots, n$ ) converge to the desired solution (55) as  $m \rightarrow \infty$ .

An equation of order  $n$

$$(57) \quad y^{(n)} = F(x, y, \dots, y^{(n-1)})$$

can be replaced by the system

$$(58) \quad \frac{dy_1}{dx} = y_2, \quad \frac{dy_2}{dx} = y_3, \dots, \frac{dy_{n-1}}{dx} = y_n, \quad \frac{dy_n}{dx} = F(x, y_1, \dots, y_n)$$

and then treated in the same way.

**Step-by-Step Integration.** An approximation to the solution (55) can be obtained by replacing (54) by the corresponding incremental equations

$$(59) \quad \Delta y_i = F_i(x, y_1, \dots, y_n) \Delta x.$$

The values of  $y_1, \dots, y_n$  at  $x = x_0$  are the given constants  $k_1, \dots, k_n$ . The values at  $x_0 + \Delta x$  are  $k_1 + \Delta y_1, \dots, k_n + \Delta y_n$ , where the  $\Delta y_i$  are computed from eq. (59) with  $F_i$  evaluated at  $x_0, k_1, \dots, k_n$ . Proceeding in this manner, step by step, one obtains values of  $y_1, \dots, y_n$  at discrete values  $x_0, x_0 + \Delta x, \dots$ . The increments  $\Delta x$  can be varied or kept constant. Under appropriate hypotheses, the "solution" thus computed converges to the desired solution as the increments  $\Delta x$  approach 0. (See Ref. 3.)

**Other Methods.** A variety of more refined numerical methods have been developed, in many cases well suited to analysis on digital computers. For details see Chap. 14.

## 9. GRAPHICAL METHODS—PHASE PLANE ANALYSIS

The discussion will be limited to the first order equation

$$(60) \quad y' = f(x, y)$$

and to the system

$$(61) \quad \frac{dx}{dt} = F(x, y), \quad \frac{dy}{dt} = G(x, y).$$

Elimination of  $t$  between the two eqs. (61) leads to an equation of form (60) and, indeed, eqs. (61) should be thought of as a parametric form of eq. (60);  $t$  can be interpreted as *time*. It should be remarked that certain second order equations can be reduced to these forms. For *example* an equation

$$(62) \quad \frac{d^2x}{dt^2} = f\left(x, \frac{dx}{dt}\right)$$

is equivalent to the pair

$$(63) \quad \frac{dx}{dt} = y, \quad \frac{dy}{dt} = f(x, y).$$

**Method of Isoclines.** For the eq. (60) the isoclines are the loci of constant slope; that is, the loci

$$(64) \quad f(x, y) = m.$$

In general, eq. (64) defines a family of curves in the  $xy$ -plane. If one draws a series of short line segments, all with slope  $m$ , and each with its center on the curve, then one has obtained a series of tangent lines to the unknown solutions. If the process is repeated for many different values of

$m$ , one obtains a dense set of such tangent lines. From these tangent lines, the solutions are usually readily sketched. An *example* is suggested in Fig. 1.

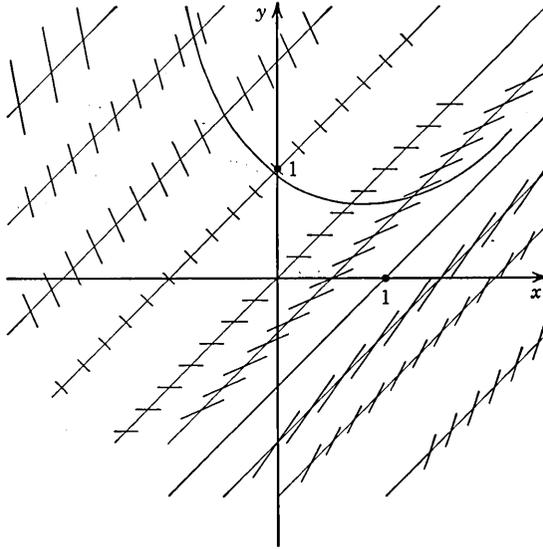


FIG. 1. Solution of  $y' = x - y$  by isoclines.

**Singular Points.** The form (60) has certain disadvantages which can be avoided by using an appropriate equivalent form (61). In particular,  $f(x, y)$  may have discontinuities, while the equivalent form (61) need have none. For example, the equation  $y' = (x^2 - y^2)/xy$  can be replaced by

$$(65) \quad \frac{dx}{dt} = xy, \quad \frac{dy}{dt} = x^2 - y^2;$$

the new system has no discontinuities. Equations (65) can be thought of as defining a vector field in the  $xy$ -plane. Instead of drawing tangent lines as above, one can draw many vectors at scattered points in the  $xy$ -plane. These suggest the solution curves in the same way as do the tangent lines.

The vector field fails to define a direction only where both  $dx/dt$  and  $dy/dt$  are 0. For eqs. (65) this holds only at the origin. For the general system (61) this holds where both equations

$$(66) \quad F(x, y) = 0, \quad G(x, y) = 0$$

are satisfied. The solutions of eqs. (66) are called *singular points*.

A graphical analysis near the singular points is generally difficult to complete with the aid of isoclines alone. It is possible to obtain a qualitative

picture of the solutions near the singular point by expanding  $F(x, y)$ ,  $G(x, y)$  in power series and neglecting all but the linear terms. If the singular

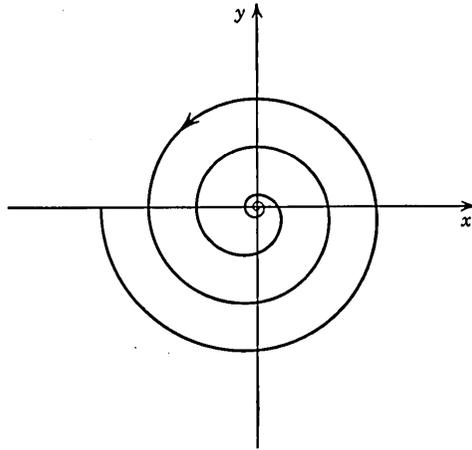


FIG. 2. Solution near focus type singular point.

point is at  $(0, 0)$ , one thus replaces eqs. (61) by the approximating linear system

$$(67) \quad \frac{dx}{dt} = a_1x + b_1y, \quad \frac{dy}{dt} = a_2x + b_2y.$$

The appearance of the solutions of eqs. (61) near the singular point in typical cases is suggested in Figs. 2, 3, 4, 5. The arrows on the curves indi-

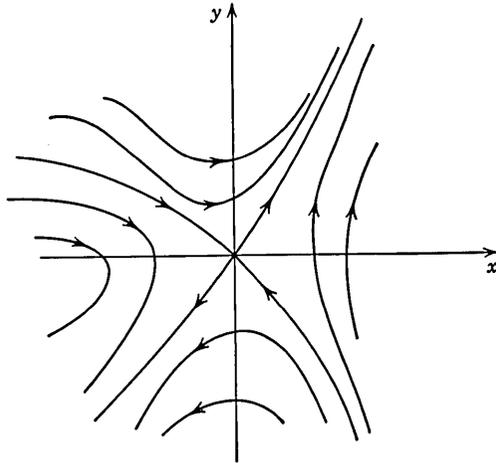


FIG. 3. Solution near saddle-point type singular point.

cate the direction of increasing time  $t$ . The four cases illustrated correspond to four cases for the roots of the characteristic equation

$$(68) \quad \begin{vmatrix} a_1 - \lambda & b_1 \\ a_2 & b_2 - \lambda \end{vmatrix} = 0.$$

In Fig. 2 the roots are  $-\alpha \pm \beta i$ , with  $\alpha > 0$ ,  $\beta > 0$ ; in Fig. 3 the roots are  $\lambda_1, \lambda_2$  with  $\lambda_1 < 0 < \lambda_2$ ; in Fig. 4 the roots are  $\pm \beta i$ ,  $\beta > 0$ ; in Fig. 5

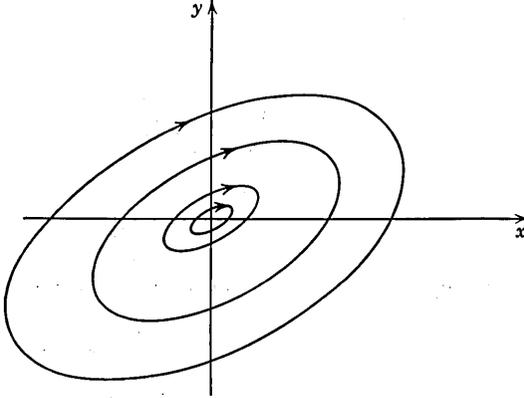


FIG. 4. Solution near center type singular point.

the roots are  $\lambda_1, \lambda_2$  with  $\lambda_1 < \lambda_2 < 0$ . The solutions of the system (61) will have the same appearance near  $(0, 0)$  as the solution of eqs. (67), except in borderline cases; of the four cases illustrated, only that of Fig. 4 is of borderline type. For a full discussion, see Ref. 2.

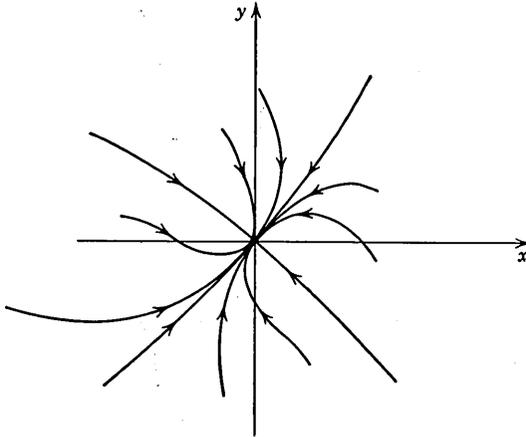


FIG. 5. Solution near node type singular point.

**Limit Cycles.** Of much importance for applications are the solutions represented by closed curves in the  $xy$ -plane. These are termed *limit cycles*. For the parametric eqs. (61) such a solution is represented by equations  $x = p(t)$ ,  $y = q(t)$ , where  $p$  and  $q$  have a common period  $T$ . A typical solution family containing a limit cycle  $C$  is illustrated in Fig. 6. The

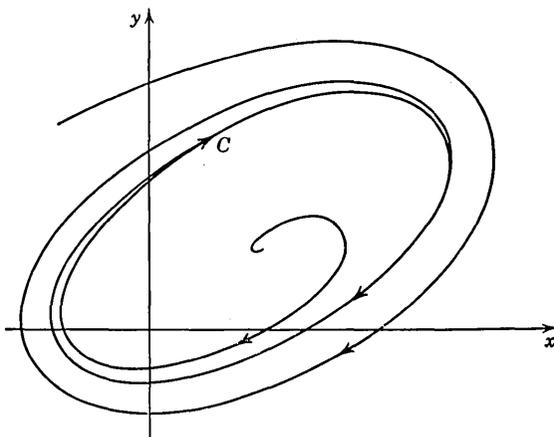


FIG. 6. Limit cycle.

cycle  $C$  is *stable* in this case; that is, all solutions starting near  $C$  approach  $C$  as time  $t$  increases. In many cases simple properties of the isoclines allow one to conclude existence of limit cycles in particular regions. A theorem of Bendixson states that a region in which  $F_x + G_y > 0$  can contain no limit cycle of eqs. (61) (Refs. 2, 3).

**Phase Plane.** For the motion of a particle of mass  $m$  on a line, classical mechanics gives an equation of the form

$$(69) \quad m \frac{d^2x}{dt^2} = F\left(t, x, \frac{dx}{dt}\right).$$

When  $F$  is independent of  $t$ , the substitution  $v = dx/dt$  leads to an equation

$$(70) \quad mv \frac{dv}{dx} = F(x, v)$$

which can be analyzed as above. The pair  $(x, v)$  represents a *phase* of the mechanical system and the  $xv$ -plane is termed the *phase plane*. Second order equations arising in other contexts can be treated similarly and the term phase is used for the pair  $(x, v)$  or  $(x, y)$  regardless of the physical

significance of the variables. An especially simple graphical discussion can be given for the *conservative* case of eq. (69):

$$(71) \quad m \frac{d^2x}{dt^2} = F(x).$$

See Ref. 2.

## 10. PARTIAL DIFFERENTIAL EQUATIONS

This section presents a brief discussion of partial differential equations of second order. Some further information is given in Chap. 6. (See Refs. 4, 10.)

**Classification.** Consider an equation

$$(72) \quad A \frac{\partial^2 u}{\partial x^2} + 2B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D \frac{\partial u}{\partial x} + E \frac{\partial u}{\partial y} + Fu + G = 0$$

where  $u$  is an unknown function of  $x$  and  $y$  and the coefficients  $A, \dots, G$  are given functions of  $x$  and  $y$  (perhaps constants). The eq. (72) is termed

*elliptic* if  $B^2 - AC < 0$ ,

*parabolic* if  $B^2 - AC = 0$ ,

*hyperbolic* if  $B^2 - AC > 0$ .

The three types are illustrated by the

$$\text{Laplace equation: } \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0,$$

$$\text{heat equation: } \frac{\partial u}{\partial t} - k^2 \frac{\partial^2 u}{\partial x^2} = 0,$$

$$\text{wave equation: } \frac{\partial^2 u}{\partial t^2} - k^2 \frac{\partial^2 u}{\partial x^2} = 0.$$

Attention will be restricted to the three special types.

**Dirichlet Problem.** One seeks a solution  $u(x, y)$  of the Laplace equation in an open region  $D$ , with given boundary values on the boundary of  $D$ . This problem can be treated by conformal mapping (Chap. 10, Sect. 5).

**Heat Equation.** A typical problem is the following. One seeks a solution  $u(x, t)$  of the heat equation  $u_t - k^2 u_{xx} = 0$  for  $t > 0$ ,  $0 < x < 1$ , with given initial values  $\phi(x) = u(x, 0)$  and boundary values  $u(0, t) = 0$ ,  $u(1, t) = 0$ . To obtain a solution one can employ the method of *separa-*

tion of variables. One seeks solutions of the differential equation and boundary conditions of form

$$(73) \quad u = f(x)g(t).$$

From the differential equation one finds that one must have

$$\frac{g'(t)}{g(t)} - k^2 \frac{f''(x)}{f(x)} = 0.$$

Hence  $g'/g$  must be a constant  $\lambda$ , and  $f''/f$  must equal  $\lambda/k^2$ :

$$(74) \quad g'(t) - \lambda g(t) = 0, \quad k^2 f''(x) - \lambda f(x) = 0.$$

From the boundary conditions at  $x = 0$  and  $x = 1$  one finds that

$$(75) \quad f(0) = f(1) = 0.$$

From eqs. (74) and (75) one concludes that  $f(x)$  and  $\lambda$  must have the form

$$(76) \quad f(x) = b \sin n\pi x, \quad \lambda = -k^2 n^2 \pi^2, \quad n = 1, 2, \dots$$

From eqs. (74)  $g(t)$  has form  $\text{const.} \cdot e^{\lambda t}$ . Hence particular solutions of form (73) have been found:

$$(77) \quad u = e^{-k^2 n^2 \pi^2 t} \sin n\pi x, \quad n = 1, 2, \dots$$

Each linear combination of the functions (77) is also a solution of both the heat equation and the boundary conditions at  $x = 0$  and  $x = 1$ . Accordingly, each convergent series

$$(78) \quad u = \sum_{n=1}^{\infty} b_n e^{-k^2 n^2 \pi^2 t} \sin n\pi x$$

also represents a solution. By proper choice of the constants  $b_n$  the initial values can be satisfied. One requires that

$$(79) \quad \phi(x) = \sum_{n=1}^{\infty} b_n \sin n\pi x.$$

Thus the  $b_n$  are determined from the expansion of  $\phi(x)$  in its Fourier sine series (Chap. 8, Sect. 8). With the  $b_n$  so chosen, eq. (78) represents the desired solution of the given problem.

**Wave Equation.** One seeks a solution  $u(x, t)$  of the wave equation  $u_{tt} - k^2 u_{xx} = 0$  for  $0 < x < \pi$ ,  $t > 0$  with given initial values  $u(x, 0) = \phi(x)$  and initial velocities  $u_t(x, 0) = \psi(x)$  and given boundary values  $u(0, t) = u(\pi, t) = 0$ . This is the problem of the *vibrating string*. The

method of separation of variables can be used as above and one obtains the solution in the form of a series

$$(80) \quad u = \sum_{n=1}^{\infty} \sin nx [\alpha_n \sin knt + \beta_n \cos knt],$$

where  $\alpha_n$  and  $\beta_n$  are determined from the expansions:

$$(81) \quad \phi(x) = \sum_{n=1}^{\infty} \beta_n \sin nx, \quad \psi(x) = \sum_{n=1}^{\infty} nk\alpha_n \sin nx.$$

**Relaxation Methods.** One can obtain an approximation to the solution of a partial differential equation by replacing it by a corresponding difference equation. The method has been especially successful for the Dirichlet problem, which is discussed here. The differential equation  $u_{xx} + u_{yy} = 0$  is replaced by the equation

$$(82) \quad u(x+h, y) + u(x, y+h) + u(x-h, y) \\ + u(x, y-h) - 4u(x, y) = 0.$$

If the given region is the square  $0 \leq x \leq 1$ ,  $0 \leq y \leq 1$ , one chooses  $h = 1/n$  for some positive integer  $n$  and requires eq. (82) to hold at the *lattice points*  $(k_1h, k_2h)$ ,  $0 < k_1 < n$ ,  $0 < k_2 < n$ . The values of  $u$  on the boundary ( $x = 0$  or  $1$ ,  $y = 0$  or  $1$ ) are given, and eq. (82) becomes a system of simultaneous linear equations for the unknowns  $u(k_1h, k_2h)$ . These can be solved by the *relaxation method*. One chooses an initial set of values for the unknowns, then obtains a next approximation by replacing  $u(x, y)$  by

$$(83) \quad \frac{1}{4}[u(x+h, y) + u(x, y+h) + u(x-h, y) + u(x, y-h)]$$

at each lattice point. Repetition of the process generates a sequence  $u_n(x, y)$  which can be shown to converge to the solution of eq. (82). As  $h \rightarrow 0$ , the solution of eq. (82) can be shown to converge to the desired solution of the Dirichlet problem (Ref. 10).

## REFERENCES

1. R. P. Agnew, *Differential Equations*. McGraw-Hill, New York, 1942.
2. A. Andronow and C. E. Chaikin, *Theory of Oscillations*, Princeton University Press, Princeton, N. J., 1949.
3. E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
4. R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. I, Interscience, New York, 1953.

5. E. L. Ince, *Ordinary Differential Equations*, Longmans, Green, London, 1927.
6. E. Kamke, *Differentialgleichungen, Lösungsmethoden und Lösungen*, Vol.1, 2nd edition, Akademische Verlagsgesellschaft, Leipzig, 1943.
7. E. Kamke, *Differentialgleichungen reeller Funktionen*, Akademische Verlagsgesellschaft, Leipzig, 1933.
8. E. D. Rainville, *Elementary Differential Equations*, Macmillan, New York, 1952.
9. E. D. Rainville, *Intermediate Differential Equations*, Wiley, New York, 1943.
10. R. V. Southwell, *Relaxation Methods in Engineering Sciences*, Oxford University Press, Oxford, England, 1946.
11. E. T. Whittaker and G. M. Watson, *A Course of Modern Analysis*, 4th edition, Cambridge University Press, Cambridge, England, 1940.



# Integral Equations

E. H. Rothe

1. Definitions and Main Problems	6-01
2. Relation to Boundary Value Problems	6-03
3. General Theorems	6-05
4. Theorems on Eigenvalues	6-06
5. The Expansion Theorem and Some of Its Consequences	6-07
6. Variational Interpretation of the Eigenvalue Problem	6-08
7. Approximation Methods	6-10
References	6-17

## 1. DEFINITIONS AND MAIN PROBLEMS

A *linear integral equation of first kind* is an equation of form

$$(1) \quad \int_a^b K(s, t)x(t) dt = f(s);$$

$f(s)$  and  $K(s, t)$  are considered to be given, and a function  $x(t)$  satisfying eq. (1) is called a *solution* of the integral equation.

**Fredholm Integral Equation.** This is the *linear integral equation of second kind* and has the form

$$(2) \quad x(s) - \lambda \int_a^b K(s, t)x(t) dt = f(s).$$

Here  $K(s, t)$  and  $f(s)$  are given real functions, and  $\lambda$  is a given real constant; a solution of the integral equation is a function  $x(s)$  satisfying eq. (2) for  $a \leq s \leq b$ .

**Volterra Integral Equation.** If in eq. (2) the upper limit  $b$  is replaced by the variable  $s$ , the resulting equation

$$(3) \quad x(s) - \lambda \int_a^s K(s, t)x(t) dt = f(s)$$

is called a *Volterra integral equation*. Equation (3) can be considered as a special case of eq. (2); namely, the case for which  $K(s, t) = 0$  for  $t \geq s$ .

The preceding definitions relate to integral equations for functions of one real variable. There are analogous definitions for functions of two or more real variables. It is also of importance to allow  $x, K, f$  to take on complex values and to allow  $\lambda$  to be complex. For simplicity the results will be formulated for functions of one variable; essentially no change is required to extend the results to functions of several variables. Only functions with real values will be considered here. The discussion will furthermore be restricted to the integral equation of second kind; for the equation of first kind, see Ref. 10, Chap. 2.

**REMARK.** The equations defining Laplace and Fourier transforms can be regarded as integral equations of first kind. Solving the equations is equivalent to finding the inverse transforms. See Chaps. 8, 9.

The function  $K(s, t)$  in eq. (2) is called the *kernel* of the integral equation. The eq. (2) is said to be *homogeneous* if  $f(s) = 0$ ; otherwise it is *nonhomogeneous*. The homogeneous equation

$$(4) \quad x(s) - \lambda \int_a^b K(s, t)x(t) dt = 0$$

obtained from eq. (2) by replacing  $f(s)$  by 0 is called the *homogeneous equation associated* with eq. (2).

A number  $\lambda$  such that eq. (4) has a solution  $x = \phi(s)$  not identically 0 is called a *characteristic value* or *eigenvalue* of eq. (4) or of the kernel  $K(s, t)$ ; the solution  $\phi(s)$  is called an *eigenfunction* associated with  $\lambda$ . For each eigenvalue  $\lambda$  there may be several associated eigenfunctions. From the definition it follows that 0 cannot be an eigenvalue.

The *eigenvalue problem* associated with eq. (4) is the determination of whether, for a given kernel, eigenvalues exist, what they are, and what the corresponding eigenfunctions are.

The *expansion problem* associated with eq. (3) is the determination of the possibility of expanding every function  $g(s)$  of a given class in an in-

finite series:

$$g(s) = \sum_{\alpha=1}^{\infty} c_{\alpha} \phi_{\alpha}(s),$$

where the  $\phi_{\alpha}(s)$  are eigenfunctions.

The *solvability problem* associated with eq. (2) is the determination of whether eq. (2) has a solution  $x(s)$  and whether the solution is unique.

2. RELATION TO BOUNDARY VALUE PROBLEMS

The problems described in Sect. 1 arise naturally in the analysis of boundary value problems associated with partial differential equations.

A TYPICAL EXAMPLE is presented. The equation is the *wave equation*

$$(5) \quad u_{tt} - \nabla^2 u = 0, \quad u = u(x, y, z, t),$$

where  $\nabla^2$  is the Laplacian operator:

$$(6) \quad \nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}.$$

A finite domain  $D$ , with smooth boundary  $B$ , is given in  $x, y, z$  space; a function  $g(x, y, z)$  is given in  $D$ . One seeks a function  $u$  satisfying eq. (5) for  $t \geq 0$  and for  $(x, y, z)$  in  $D$  and satisfying the boundary conditions

$$(7) \quad u(x, y, z, t) = 0, \quad (x, y, z) \text{ on } B, t \geq 0;$$

$$(8) \quad u(x, y, z, 0) = g(x, y, z), \quad (x, y, z) \text{ in } D;$$

$$(9) \quad u_t(x, y, z, 0) = 0, \quad (x, y, z) \text{ in } D.$$

The *classical attack* on this problem is to “separate” the time and space variables; that is, to set

$$(10) \quad u(x, y, z, t) = S(x, y, z)T(t).$$

Then one is led to the boundary value problems:

$$(11) \quad \nabla^2 S + \lambda S = 0, \quad S = 0 \text{ for } (x, y, z) \text{ on } B.$$

$$(12) \quad T''(t) + \lambda T(t) = 0, \quad T'(0) = 0.$$

If  $S$  and  $T$  satisfy eqs. (11), (12) for some constant  $\lambda$ , then  $u = ST$  satisfies eqs. (5) and both (7) and (9), but not necessarily eq. (8).

**Integral Equation.** The problem (11) can be replaced by an integral equation by the following reasoning. It is shown in the theory of partial differential equations (Ref. 5) that there exists a uniquely determined function  $K(s, \sigma)$  of the two points

$$s: (x, y, z), \quad \sigma: (\xi, \eta, \zeta)$$

in  $D$ , the so-called *Green's function*, with the following properties:  $K$  has continuous second partial derivatives as long as  $s \neq \sigma$ ;  $K(s, \sigma) = 0$  for  $s$  on  $B$ ,  $\sigma$  in  $D$ ; if  $\phi(\sigma)$  has continuous first partial derivatives in  $D$  and  $D_1$  is an arbitrary subdomain of  $D$  (with smooth boundary), then

$$(13) \quad \nabla^2 \iiint_{D_1} K(x, y, z, \xi, \eta, \zeta) \phi(\xi, \eta, \zeta) d\xi d\eta d\zeta = -\phi(x, y, z)$$

for  $(x, y, z)$  in  $D_1$ . Identifying  $\phi$  with  $-\lambda S$  and  $D_1$  with  $D$ , one sees that the boundary value problem (11) is equivalent to the homogeneous integral equation in these variables:

$$(14) \quad S(x, y, z) - \lambda \iiint_D K(x, y, z, \xi, \eta, \zeta) S(\xi, \eta, \zeta) d\xi d\eta d\zeta = 0.$$

*Solution.* Let eq. (14) have a sequence of eigenfunctions  $S_\alpha(x, y, z)$  associated with the positive eigenvalues  $\lambda_\alpha$  ( $\alpha = 1, 2, \dots$ ). Then  $S_\alpha$  satisfies eq. (11) with  $\lambda = \lambda_\alpha$ ; for  $\lambda = \lambda_\alpha$  eq. (12) has the solution  $T_\alpha \equiv \cos \sqrt{\lambda_\alpha} t$ , so that  $u = S_\alpha T_\alpha$  satisfies eqs. (5), (7), (9). To satisfy eq. (8), one notes that each series

$$(15) \quad u = \sum_{\alpha=1}^{\infty} c_\alpha S_\alpha(x, y, z) T_\alpha(t) = \sum c_\alpha S_\alpha \cos \sqrt{\lambda_\alpha} t$$

also satisfies eqs. (5), (7), (9), if the  $c$ 's are constants and the series satisfies appropriate convergence conditions. The condition (8) now becomes

$$(16) \quad g(x, y, z) = \sum_{\alpha=1}^{\infty} c_\alpha S_\alpha(x, y, z);$$

thus one is led to the expansion problem. If  $g$  can be expanded as in eq. (8), then (15) defines a solution of the given problem.

Suppose that the 0 on the right-hand side of eq. (5) is replaced by  $F_0(x, y, z, t)$ ; this corresponds to an external force. If  $F_0(x, y, z, t) = F(x, y, z)T(t)$ , where  $T(t)$  satisfies eq. (12) for some  $\lambda = \lambda_0$ , then the substitution of eq. (10) leads to the nonhomogeneous integral equation

$$(17) \quad S(x, y, z) - \lambda_0 \iiint_D K(x, y, z, \xi, \eta, \zeta) S(\xi, \eta, \zeta) d\xi d\eta d\zeta = f(x, y, z),$$

where

$$(18) \quad f(x, y, z) = \iiint_D K(x, y, z, \xi, \eta, \zeta) F(\xi, \eta, \zeta) d\xi d\eta d\zeta.$$

## 3. GENERAL THEOREMS

In what follows  $K(s, t)$  will be assumed continuous for  $a \leq s \leq b$ ,  $a \leq t \leq b$ . Such a continuity condition is not always satisfied, e.g., for the Green's function of Sect. 2. See Ref. 5 (pp. 543 ff.) and Ref. 6 (pp. 355 ff.) for reduction of the discontinuous case to the continuous case.

**Definitions.** The following definitions relate to functions of  $t$  defined and continuous for  $a \leq t \leq b$ . If  $x, y$  are two such functions, their *scalar product* is

$$(19) \quad (x, y) = \int_a^b x(t)y(t) dt.$$

The *norm*  $\|x\|$  of  $x = x(t)$  is defined as  $(x, x)^{1/2}$ . Functions  $x_1, \dots, x_n$  are *linearly independent* if

$$(20) \quad c_1x_1(t) + \dots + c_nx_n(t) \equiv 0,$$

with constant  $c_1, \dots, c_n$ , implies  $c_1 = 0, \dots, c_n = 0$ ; if the functions are not linearly independent, they are termed *linearly dependent*. An infinite system of functions

$$(21) \quad \phi_1, \phi_2, \dots$$

is called linearly independent if  $\phi_1, \dots, \phi_k$  are linearly independent for every  $k$ .

Two functions  $x, y$  are said to be *orthogonal* if  $(x, y) = 0$ . The system (21) is orthogonal if  $(\phi_\alpha, \phi_\beta) = 0$  for  $\alpha \neq \beta$ . A system of orthogonal functions none of which is identically zero is necessarily linearly independent. The system (21) is called *orthonormal* if it is orthogonal and *normalized*, that is,

$$\|\phi_\alpha\| = 1 \text{ for all } \alpha.$$

If the system (21) is *linearly independent*, it can be orthogonalized and normalized; that is, an orthonormal system  $\{\psi_\alpha\}$  can be found such that, for every  $n$ ,  $\psi_n$  is a linear combination of  $\phi_1, \dots, \phi_n$  and  $\phi_n$  is a linear combination of  $\psi_1, \dots, \psi_n$ . For details, see Ref. 4, p. 50.

THE SCHWARZ INEQUALITY states that for every  $x, y$ ,  $|(x, y)| \leq \|x\| \cdot \|y\|$ , with equality if and only if  $x, y$  are linearly dependent.

THE BESSEL INEQUALITY states that, if  $\{\phi_\alpha\}$  is an orthonormal system, then for every  $x$

$$(22) \quad \sum_{\alpha=1}^{\infty} |(x, \phi_\alpha)|^2 \leq \|x\|^2.$$

Now consider three related integral equations:

$$(23) \quad \phi(s) - \lambda \int_a^b K(s, t)\phi(t) dt = f(s),$$

$$(24) \quad \phi(s) - \lambda \int_a^b K(s, t)\phi(t) dt = 0,$$

$$(25) \quad \psi(s) - \lambda \int_a^b K(t, s)\psi(t) dt = 0.$$

Equation (24) is the homogeneous equation related to (23); eq. (25) is called the *adjoint* or *transposed* equation of (24).

**THEOREM 1.** *If  $\lambda$  is an eigenvalue of eq. (24), then  $\lambda$  is an eigenvalue of eq. (25). There are at most a finite number  $k$  of linearly independent eigenfunctions of eq. (24) associated with eigenvalue  $\lambda$ ; this maximal number  $k$  is the same for eq. (25).*

The number  $k$  is called the *multiplicity* of the eigenvalue  $\lambda$ .

**THEOREM 2.** *Equation (23) has a solution if and only if  $f$  is orthogonal to all solutions of the adjoint eq. (25).*

**Conclusions Based on Theorems 1 and 2.** *Let  $\lambda$  not be an eigenvalue of  $K(s, t)$ . Then  $\lambda$  is also not an eigenvalue of  $K(t, s)$ ; that is,  $\psi \equiv 0$  is the only solution of eq. (25). Hence  $(f, \psi) = 0$  for all solutions  $\psi$  of eq. (25), and eq. (23) has a solution for arbitrary  $f$ . For each  $f$ , the solution is unique; for the difference  $\phi$  of two solutions is a solution of eq. (24), hence  $\phi \equiv 0$ .*

*Let  $\lambda$  be an eigenvalue of  $K(s, t)$ . Then eq. (25) is satisfied for at least one  $\psi$  not identically zero and eq. (23) is not satisfied for some  $f$ , in particular for  $f = \psi$ . (In the problem of Sect. 2 this case arises if the frequency  $\lambda_0$  of the time factor  $T$  of  $F_0(x, y, z, t)$  is an eigenvalue of the homogeneous eq. (14); this is the case of *resonance*.)*

One is thus led to the following *alternative of Fredholm*: either (i) the nonhomogeneous eq. (23) has a solution for arbitrary  $f$  or (ii) the homogeneous eq. (24) has at least one (not identically vanishing) solution. Case (i) can also be characterized by the statement: eq. (23) has at most one solution for each  $f$ ; for the uniqueness implies existence of a solution.

#### 4. THEOREMS ON EIGENVALUES

The kernel  $K(s, t)$  is said to be *symmetric* if  $K(s, t) \equiv K(t, s)$ . This case occurs in many applications; for example in the problem of Sect. 2.

**THEOREM 3.** *A symmetric kernel has at least one and at most a countable infinity of eigenvalues. Eigenfunctions corresponding to distinct eigenvalues are orthogonal. The eigenvalues can be numbered to form a sequence  $\{\lambda_\alpha\}$ , in which each eigenvalue is repeated as many times as its multiplicity, and such*

that  $|\lambda_1| \leq |\lambda_2| \leq \dots$ ; if there are infinitely many eigenvalues, then  $|\lambda_\alpha| \rightarrow \infty$  as  $\alpha \rightarrow \infty$ . An eigenfunction  $\phi_\alpha$  can be assigned to each  $\lambda_\alpha$  in such a fashion that the sequence  $\{\phi_\alpha\}$  is orthonormal and every eigenfunction  $\phi$  is a linear combination of a finite number of the  $\phi_\alpha$ 's.

The sequence  $\{\phi_\alpha\}$  is called a full system of eigenfunctions of the kernel.

REMARK. While restricting  $K(s, t)$  to be real, one can consider complex eigenvalues  $\lambda$  and eigenfunctions  $x(t) = x_1(t) + ix_2(t)$ . Some kernels have only complex eigenvalues; some kernels have no eigenvalues at all. A symmetric kernel has only real eigenvalues.

5. THE EXPANSION THEOREM AND SOME OF ITS CONSEQUENCES

THEOREM 4. Let  $\{\phi_\alpha\}$  be a full system of eigenfunctions for the symmetric kernel  $K(s, t)$ . Then in order that a function  $g(s)$  can be expanded in a uniformly convergent series:

$$(26) \quad g(s) = \sum_{\alpha} c_{\alpha} \phi_{\alpha}(s),$$

where

$$(27) \quad c_{\alpha} = (g, \phi_{\alpha}),$$

it is sufficient that  $g(s)$  can be written in the form

$$(28) \quad g(s) = \int_a^b K(s, t)G(t) dt,$$

where  $G(t)$  is continuous.

In many applications the form (28) for the function  $g(s)$  to be expanded arises in a natural way. For example, the function (18) is of this form.

The coefficients (27) can be written in a different form which is often useful. From eq. (28) and from the facts that  $\phi_\alpha$  satisfies eq. (24) with  $\lambda = \lambda_\alpha$  and that  $K$  is symmetric, one deduces the expression

$$(29) \quad c_{\alpha} = \frac{(G, \phi_{\alpha})}{\lambda_{\alpha}}.$$

As a first application of the expansion theorem, let  $\lambda$  be a number which is not an eigenvalue, and seek to expand the solution  $x = \phi(s)$  of eq. (23) in terms of the eigenfunctions. To do this, note that by eq. (23)  $x - f$  is of form (28) with  $G = \lambda x$ . By Theorem 4 and eq. (29) one deduces the expansion

$$(30) \quad x(s) - f(s) = \sum_{\alpha} \frac{\lambda}{\lambda_{\alpha}} (x, \phi_{\alpha}) \phi_{\alpha}(s).$$

If this relation is multiplied by  $\phi_\beta(s)$  and integrated from  $a$  to  $b$ , one ob-

tains a linear equation for  $(x, \phi_\beta)$ . Solving this equation and substituting the result in eq. (30) gives the desired formula

$$(31) \quad x(s) = f(s) + \lambda \sum_{\alpha} \frac{(f, \phi_{\alpha})}{\lambda_{\alpha} - \lambda} \phi_{\alpha}(s).$$

(The series is meaningless if  $\lambda$  is one of the  $\lambda_{\alpha}$ , unless  $(f, \phi_{\alpha}) = 0$ ; this is in agreement with Theorem 2 of Sect. 3.)

A *second application* concerns the "quadratic form"

$$(32) \quad I\{x, x\} = \int_a^b \int_a^b K(s, t)x(s)x(t) dt ds,$$

whose importance will become clear in the next section. If one applies the expansion theorem to the integral of  $K(s, t)x(t)$ , one obtains the formula:

$$(33) \quad I\{x, x\} = \sum_{\alpha} \frac{k_{\alpha}^2}{\lambda_{\alpha}}, \quad k_{\alpha} = (x, \phi_{\alpha}).$$

The transition from eq. (32) to (33) is the analogue of choosing coordinates which represent a conic section in its "principal axis" form.

## 6. VARIATIONAL INTERPRETATION OF THE EIGENVALUE PROBLEM

In this section the hypotheses and notations are the same as those of Sect. 5. It is convenient to denote the *positive*  $\lambda_{\alpha}$ 's by

$$(34) \quad 0 < p_1 \leq p_2 \leq p_3 \leq \dots$$

and the *negative* ones by

$$(35) \quad 0 > -n_1 \geq -n_2 \geq -n_3 \geq \dots$$

There may be no  $p$ 's or no  $n$ 's; as remarked in Sect. 1, 0 is not an eigenvalue. Equation (33) now becomes

$$(36) \quad I\{x, x\} = \sum_j \frac{k_j^2}{p_j} - \sum_j \frac{l_j^2}{n_j},$$

where  $k_j = (x, \psi_j)$ ,  $l_j = (x, \chi_j)$  and  $\psi_j$  is the eigenfunction associated with  $p_j$ ,  $\chi_j$  the eigenfunction associated with  $n_j$ . From eq. (36) and Bessel's inequality (22) one now concludes:

**THEOREM 5.** *If there are positive eigenvalues of the symmetric kernel  $K(s, t)$ , then*

$$(37) \quad I\{x, x\} \leq \frac{\|x\|^2}{p_1},$$

where  $p_1$  is the smallest positive eigenvalue. The maximum of  $I\{x, x\}$  for  $x$  within the class of  $x$  having norm 1 is attained when  $x = \psi_1$  and equals  $1/p_1$ . If there is a positive eigenvalue  $p_n$ , then the maximum of  $I\{x, x\}$  within the class of  $x$  for which

$$(38) \quad \|x\| = 1, \quad (x, \psi_j) = 0 \quad (j = 1, \dots, n - 1)$$

is attained when  $x = \psi_n$  and equals  $1/p_n$ .

If  $K(s, t)$  is replaced by  $-K(s, t)$ , one obtains a characterization of the negative eigenvalues and corresponding eigenfunctions.

The characterization of eigenvalues in Theorem 5 is recursive; that is, in order to characterize  $p_n$  and  $\psi_n$  one has to know  $\psi_1, \dots, \psi_{n-1}$ . A direct characterization is obtainable as follows. Let  $M\{y_1, \dots, y_{n-1}\}$  denote the least upper bound of  $I\{x, x\}$  among all  $x$  such that

$$(39) \quad (x, y_j) = 0 \quad (j = 1, \dots, n - 1).$$

It can be shown that, among all choices of  $y_1 = y_1(t), \dots, y_{n-1} = y_{n-1}(t)$ ,  $M$  has its smallest value, namely  $1/p_n$ , when  $y_1 = \psi_1, \dots, y_{n-1} = \psi_{n-1}$ . See Ref. 4, p. 132.

**Rayleigh-Ritz Quotient.** This is the quotient

$$(40) \quad Q\{x\} = I\{x, x\} \div \int_a^b \left[ \int_a^b K(s, t) x(t) dt \right]^2 ds.$$

Assume that there are at most a finite number of negative eigenvalues and assume all the eigenvalues are numbered so that  $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \dots$ . From eq. (33) one finds

$$(41) \quad I\{x, x\} = \sum_{\alpha} \left( \frac{k_{\alpha}}{\lambda_{\alpha}} \right)^2 \lambda_{\alpha} \geq \lambda_1 \sum_{\alpha} \left( \frac{k_{\alpha}}{\lambda_{\alpha}} \right)^2.$$

From the expansion theorem of Sect. 5, with  $G(t) = x(t)$ , one deduces that

$$(42) \quad \int_a^b \left[ \int_a^b K(s, t) x(t) dt \right]^2 ds = \sum_{\alpha} \left( \frac{k_{\alpha}}{\lambda_{\alpha}} \right)^2.$$

From eqs. (40), (41), (42) one thus obtains the inequality

$$(43) \quad Q\{x\} \geq \lambda_1.$$

Furthermore one can show that  $Q\{x\}$  takes on its minimum  $\lambda_1$  when  $x = \phi_1$ . Thus the smallest eigenvalue and associated eigenfunction are obtainable by minimizing  $Q\{x\}$ . This is the basis of a very effective computational procedure.

The quotient  $Q\{x\}$  can be written in another way, more familiar in the theory of differential equations. One sets

$$(44) \quad u(s) = \int_a^b K(s, t)x(t) dt,$$

so that

$$(45) \quad Q\{x\} = \int_a^b u(t)x(t) dt \div \int_a^b u^2 dt.$$

The analogous definition, and integration by parts, for the problem of Sect. 2 leads to the expression

$$(46) \quad Q\{x\} = \frac{\iiint_D (u_x^2 + u_y^2 + u_z^2) dx dy dz}{\iiint_D u^2 dx dy dz},$$

where  $u$  is the solution of the problem

$$(47) \quad \nabla^2 u = -x \text{ in } D, \quad u = 0 \text{ on } B.$$

## 7. APPROXIMATION METHODS

The first four methods to be described are devices for replacing the integral equation by a system of linear algebraic equations.

**Approximation of Integrals.** Let a subdivision of the interval  $a \leq t \leq b$  be given:

$$a = t_1 < t_2 < \cdots < t_n < t_{n+1} = b$$

and let  $\delta = \max(t_{j+1} - t_j)$ . Then for continuous  $h(t)$ , the difference

$$\int_a^b h(t) dt - \sum_{j=1}^n h(t_j)(t_{j+1} - t_j)$$

can be made as small as desired, in absolute value, by making  $\delta$  sufficiently small. Hence one can take the sum as approximation to the integral. If this is done for the Fredholm eq. (2), one obtains the approximating equation

$$(48) \quad x(s) - \lambda \sum_{j=1}^n K(s, t_j)(t_{j+1} - t_j) = f(s).$$

If one now writes

$$(49) \quad x(t_i) = x_i, \quad K(t_i, t_j)(t_{j+1} - t_j) = a_{ij}, \quad f(t_i) = b_i$$

for  $i = 1, \dots, n$ , then at  $s = t_i$  eq. (48) becomes

$$(50) \quad x_i - \lambda \sum_{j=1}^n a_{ij}x_j = b_i \quad (i = 1, \dots, n).$$

This is a system of linear equations for  $x_1, \dots, x_n$ . A solution can be regarded as giving the values of the desired  $x(s)$  at  $t_1, \dots, t_n$ ; one can interpolate linearly between these points to obtain an approximation to  $x(s)$ . The first proof by Fredholm of the main theorems of Sect. 3 was based on eq. (50) and subsequent passage to the limit ( $n \rightarrow \infty, \delta \rightarrow 0$ ).

For numerical purposes the procedure may be improved by using better approximations for the integral such as those given by the trapezoidal rule, Simpson's rule or Gauss's quadrature (Ref. 9, Chap. 7). Each of these methods replaces the integral by a sum  $\sum h(t_j)A_j$  with properly chosen abscissas  $t_j$  and "weights"  $A_j$ . For more details and also the question of convergence, see Ref. 1 (pp. 105 ff.), Ref. 3 (pp. 437 ff.), Ref. 9 (p. 455).

**Method of Degenerate Kernels.** A kernel  $A(s, t)$  is called *degenerate* if it can be written as a finite sum of products of a function of  $s$  by a function of  $t$ ; that is, if it is of the form

$$(51) \quad A(s, t) = \sum_{j=1}^n A_j(s)B_j(t).$$

Every continuous kernel  $K(s, t)$  can be approximated by a continuous degenerate kernel  $A(s, t)$ ; that is, for every  $\epsilon > 0$  there exists a continuous  $A(s, t)$  such that  $|K(s, t) - A(s, t)| < \epsilon$  for  $a \leq s \leq b, a \leq t \leq b$ . One therefore obtains an approximate solution of eq. (2) by replacing  $K$  by  $A$ . For the question of convergence one is referred to Ref. 4 (pp. 118 ff.), Ref. 1 (Abschnitt IV), and Ref. 3 (p. 464).

If  $K$  is replaced by  $A$ , the Fredholm eq. (2) is replaced by the equation

$$(52) \quad x(s) - \lambda \sum_{j=1}^{\infty} A_j(s) \int_a^b B_j(t)x(t) dt = f(s),$$

whose solution is found by solving a system of linear equations. To see this, one multiplies eq. (52) by  $B_i(s)$  and integrates with respect to  $s$  from  $a$  to  $b$ . With the notations

$$\int_a^b x(t)B_j(t) dt = x_j, \quad \int_a^b A_j(t)B_i(t) dt = a_{ij}, \quad \int_a^b f(t)B_j(t) dt = f_j,$$

one obtains the system

$$(53) \quad x_i - \lambda \sum_{j=1}^n a_{ij} x_j = f_i \quad (i = 1, \dots, n).$$

It can be verified that if  $x_1, \dots, x_n$  is a solution of this linear system, then

$$(54) \quad x(s) = f(s) + \lambda \sum_{j=1}^n x_j A_j(s)$$

is a solution of eq. (52) and, conversely, every solution of eq. (52) is obtained in this way.

**The Ritz-Galerkin Method.** This is a method for finding approximations to the eigenvalues and eigenfunctions of the homogeneous eq. (3) with symmetric kernel. Let  $\{v_\alpha\}$  be an orthonormal system. Such a system is called *complete* (in the class of continuous functions on the interval  $a \leq s \leq b$ ) if for every continuous function  $x(s)$  the sums

$$(55) \quad \sum_{i=1}^n x_i v_i(s), \quad x_i = \int_a^b x_i(t) v_i(t) dt$$

converge "in the mean" to  $x(s)$ ; that is, if

$$\lim_{n \rightarrow \infty} \int_a^b [x(s) - \sum_{i=1}^n x_i v_i(s)]^2 ds = 0.$$

**EXAMPLE.** The functions

$$v_1 = (2\pi)^{-1/2}, \quad v_{2m+1} = \pi^{-1/2} \cos ms, \quad v_{2m} = \pi^{-1/2} \sin ms, \quad (m = 1, 2, \dots)$$

form a complete orthonormal system for  $-\pi \leq s \leq \pi$ ; see Chap. 8, Sect. 8.

Now let  $\phi_1(s)$  be a normalized eigenfunction of eq. (3) corresponding to the smallest positive eigenvalue  $\lambda_1$ . (If there are no positive eigenvalues, one follows a similar procedure starting with the negative eigenvalue of smallest absolute value.) One now seeks an approximation  $\phi$  to  $\phi_1$  of form

$$(56) \quad \phi = \sum_{i=1}^n c_i v_i(s),$$

where  $\{v_\alpha\}$  is a complete orthonormal system. In order to determine the  $c_i$ , note (Theorem 5, Sect. 6) that  $1/\lambda_1$  is the maximum of  $I\{x, x\}$  when  $\|x\| = 1$ , and that this maximum is reached for  $x = \phi_1$ . Restricting at-

tention now to functions of form (56), one finds

$$(57) \quad I\{\phi, \phi\} = \sum_{\alpha=1}^n \sum_{\beta=1}^n b_{\alpha\beta} c_{\alpha} c_{\beta},$$

with

$$(58) \quad b_{\alpha\beta} = \int_a^b \int_a^b K(s, t) v_{\alpha}(s) v_{\beta}(t) ds dt.$$

The condition  $\|\phi\| = 1$  becomes

$$(59) \quad \sum_{i=1}^n c_i^2 = 1.$$

Maximizing the quadratic form (57) with side condition (59) can be analyzed by the method of Lagrange multipliers (Ref. 4). One obtains the equations

$$(60) \quad c_i - \lambda \sum_{j=1}^n b_{ij} c_j = 0 \quad (i = 1, \dots, n)$$

which, together with eq. (59), determine the  $c_i$  and  $\lambda$ . In particular,  $\lambda$  is a root of the algebraic equation obtained by setting the determinant of eq. (60) equal to zero. If  $\lambda_1^*$  is the smallest positive root of this equation, then  $\lambda_1^*$  is an approximation to  $\lambda_1$  and  $\lambda_1^* \geq \lambda_1$ ; for  $\lambda = \lambda_1^*$  eqs. (59) and (60) determine  $c_1, \dots, c_n$  and, by eq. (56), a desired approximation  $\phi$  of the eigenfunction  $\phi_1$ .

**Method of Enskog.** The method will be discussed for the Fredholm eq. (2) with symmetric kernel, with  $\lambda$  not an eigenvalue. (For less restrictive assumptions, see Ref. 7, p. 109.) It is based on a complete linearly independent system  $v_1, v_2, \dots$  with the additional property that the functions

$$(61) \quad y_n(s) = v_n(s) - \lambda \int_a^b K(s, t) v_n(t) dt$$

are orthonormal and complete. Such a system can be constructed as follows: Let  $w_1, w_2, \dots$  be a complete linearly independent system (e.g., the system of sines and cosines given above). One then defines

$$(62) \quad z_n(s) = w_n(s) - \lambda \int_a^b K(s, t) w_n(t) dt.$$

It can be proved that the  $z_n$  are likewise linearly independent and com-

plete. From the  $z_n$  one constructs an equivalent orthonormal system  $y_1, y_2, \dots$  (Sect. 3), so that relations

$$(63) \quad y_n(s) = \sum_{m=1}^n c_{nm} z_m(s)$$

hold, with constant  $c_{nm}$ . One now defines:

$$(64) \quad v_n(s) = \sum_{m=1}^n c_{nm} w_m(s).$$

It follows from eq. (62) that eq. (61) holds. Moreover, the system  $\{v_n\}$  is a complete linearly independent system.

Having a system  $\{v_n\}$  of the properties indicated, one can find an approximate solution  $x(s)$  of the Fredholm eq. (2) of form

$$x(s) = \sum_{i=1}^n c_i y_i(s), \quad c_i = (x, y_i).$$

Multiply eq. (2) by  $v_i(s)$  and integrate with respect to  $s$  from  $a$  to  $b$ , to obtain the relations:

$$\begin{aligned} (f, v_i) &= \int_a^b x(s) v_i(s) ds - \lambda \int_a^b \int_a^b K(s, t) x(t) v_i(s) dt ds \\ &= \int_a^b x(s) [v_i(s) - \lambda \int_a^b K(t, s) v_i(t) dt] ds. \end{aligned}$$

Because of the symmetry of the kernel, the expression in brackets is  $y_i(s)$ . Hence

$$(f, v_i) = \int_a^b x(s) y_i(s) ds = c_i.$$

*Iteration* is the basis of the following methods:

**Successive Approximations.** The Fredholm equation can be written in the form

$$(65) \quad x(s) = f(s) + \lambda \int_a^b K(s, t) x(t) dt.$$

This form suggests defining successive approximations  $x^{(i)}(s)$  to the solution  $x(s)$  as follows:

$$(66) \quad x^{(0)}(s) = f(s), \quad x^{(i+1)}(s) = f(s) + \lambda \int_a^b K(s, t) x^{(i)}(t) dt,$$

where  $i = 0, 1, 2, \dots$ . One can prove by induction that

$$(67) \quad x^{(n)}(s) = f(s) + \sum_{i=1}^n \lambda^i \int_a^b K^{(i)}(s, t) f(t) dt,$$

where the so-called *iterated kernels* are defined by the relations

$$(68) \quad K^{(1)}(s, t) = K(s, t), \quad K^{(i+1)}(s, t) = \int_a^b K(s, u) K^{(i)}(u, t) du,$$

for  $i = 1, 2, \dots$ . It can be proved that

$$(69) \quad x(s) = \lim_{n \rightarrow \infty} x^{(n)}(s) = f(s) + \sum_{i=1}^{\infty} \lambda^i \int_a^b K^{(i)}(s, t) dt$$

exists if  $|\lambda|$  is less than  $[(b - a) \text{Max } K(s, t)]^{-1}$ ; the series, known as *Neumann's series*, converges uniformly for  $a \leq s \leq b$ . The function  $x(s)$  defined by eqs. (69) is the solution of (65) for  $\lambda$  restricted as stated. For the Volterra eq. (3) the Neumann series converges for all  $\lambda$  and the solution is valid for all  $\lambda$ .

**The Schwarz Constants.** Write  $I\{x, x, K\}$  for the quadratic form  $I\{x, x\}$  defined by eq. (32) to express more clearly the dependence on  $K$ . The Schwarz constants are then defined as follows:

$$(70) \quad a_0 = (x, x), \quad a_i = I\{x, x, K^{(i)}\}, \quad (i = 1, 2, \dots),$$

where the  $K^{(i)}$  are defined by eq. (68). These constants (which obviously depend on the choice of the function  $x$ ) are important for the theory as well as for estimating eigenvalues. Note the following facts, supposing always that  $K(s, t)$  is symmetric:

If  $P$  is an arbitrary real number, subject only to the restriction that

$$(71) \quad a_{i+1} - Pa_{i+2} \neq 0,$$

and

$$(72) \quad Q = \frac{a_i - a_{i+1}P}{a_{i+1} - a_{i+2}P},$$

then the interval with end points  $P, Q$  contains at least one eigenvalue provided that at least one of the following assumptions is satisfied: (a)  $i$  is even, (b)  $K$  is a *positive definite kernel*, that is,  $I\{x, x\} > 0$  unless  $x \equiv 0$ . (For a proof, see Ref. 1, p. 30.) The quotients  $Q$  are termed *Temple quotients*. Setting  $P = 0$  in eq. (72) leads to consideration of the quotients  $Q_i = a_{i-1}/a_i$ . It can be shown that the sequence  $|Q_{2i-1}|$  is monotone nonincreasing and converges to  $|\lambda_1|$ , where  $\lambda_1$  is the eigenvalue of smallest

absolute value. (For further applications of the Schwarz constants, see Refs. 1, 3, 9.)

**Method of Steepest Descent.** The basis of this method is the fact that  $x$  is a solution of the Fredholm eq. (2) with symmetric kernel if and only if  $x$  minimizes the expression

$$(73) \quad F\{x\} = \frac{1}{2}[(x, x) - \lambda \int_a^b \int_a^b K(s, t)x(s)x(t) ds dt] - (f, x).$$

Let  $x_0$  be a first approximation to  $x$ . One seeks a better approximation  $x_1 = x_0 + h$ , and tries to choose  $h$  so that in going from  $x_0$  to  $x_1$  the value of  $F$  descends as rapidly as possible. With the notation

$$(74) \quad L[x] = x - \lambda \int_a^b K(s, t)x(t) dt,$$

one finds that

$$(75) \quad F\{x_0 + h\} = F\{x_0\} + (L[x_0] - f, h) + \frac{1}{2}(L[h], h).$$

Now if  $F$  were a function of a finite number of real variables, the analogue of the second term on the right side of eq. (75) would be the scalar product of  $\text{grad } F$  with  $h$ . One therefore defines here

$$(76) \quad g[x] = L[x] - f$$

as the gradient of  $F$ . This suggests that, as in the case of a function of a finite number of real variables, the *direction of steepest descent is given by the negative gradient*; this can be proved to be true. One therefore sets  $h = -\alpha g[x_0]$ , where  $\alpha$  is a real constant to be determined. Replace  $h$  by  $-\alpha g[x_0]$  in eq. (75); then  $F[x_0 + h]$  becomes a function of the real variable  $\alpha$ . Now determine  $\alpha$  by minimizing this function by the ordinary methods of calculus. The result for the desired next approximation  $x_1 = x_0 + h$  is

$$(77) \quad x_1 = x_0 - \frac{\|g[x_0]\|^2}{(L[g[x_0]], g[x_0])} g[x_0].$$

If one repeats the procedure starting with  $x_1$  instead of  $x_0$ , one obtains a new approximation  $x_2$ ; continuing thus, one obtains a sequence  $x_1, x_2, \dots, x_n, \dots$ . If the kernel  $K$  is symmetric and positive definite and  $|\lambda|$  is less than  $|\lambda_\alpha|$  for every eigenvalue  $\lambda_\alpha$ , then  $x_n$  converges in the mean to the solution  $x$  of the Fredholm eq. (2). For proofs and details, see Ref. 8 (pp. 103 and 136). The method can also be applied to finding eigenvalues (Ref. 8, p. 142).

REFERENCES

1. H. Bückner, *Die Praktische Behandlung von Integralgleichungen, Ergebnisse der Angewandten Mathematik*, Vol. 1, Springer Verlag, Berlin, Göttingen, Heidelberg, 1952.
2. L. Collatz, *Eigenwertprobleme and ihre numerische Behandlung*, Chelsea Publishing Company, New York, 1948.
3. L. Collatz, *Numerische Behandlung von Differentialgleichungen, Die Grundlehren der Mathematischen Wissenschaften in Einzeldarstellungen*, Vol. LX, 2nd edition, Springer Verlag, Berlin, Göttingen, Heidelberg, 1955.
4. R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 1, Interscience Publishers, New York-London, 1953.
5. Ph. Frank and R. v. Mises, *Die Differential- und Integralgleichungen der Mechanik und Physik*, 2nd edition, Vieweg, Braunschweig, 1930 (republished Rosenberg, New York, 1943).
6. E. Goursat, *Cours d'Analyse Mathématique*, Vol. 3, 3rd edition, Gauthier-Villars, Paris, 1923.
7. G. Hamel, *Integralgleichungen, Einfuehrung in Lehre und Gebrauch*, Springer, Berlin, 1937 (Edwards Brothers, Ann Arbor, Mich., 1946).
8. L. V. Kantorovich, *Functional Analysis and Applied Mathematics*, National Bureau of Standards Report 1509, 1952 [translated from *Uspekhi Matemat. Nauk*, 3 (6), 89-185 (1948)].
9. Z. Kopal, *Numerical Analysis*, Wiley, New York, 1955.
10. W. Schmeidler, *Integralgleichungen mit Anwendungen in Physik und Technik*, I. *Lineare Integralgleichungen*, Akademische Verlagsgesellschaft Geest u. Portig, Leipzig, 1950.



# Complex Variables

W. Kaplan

1. Functions of a Complex Variable	7-01
2. Analytic Functions. Harmonic Functions	7-04
3. Integral Theorems	7-05
4. Power Series. Laurent Series	7-08
5. Zeros. Singularities. Residues. Argument Principle	7-11
6. Analytic Continuation	7-16
7. Riemann Surfaces	7-17
8. Elliptic Functions	7-18
9. Functions Defined by Linear Differential Equations	7-21
10. Other Transcendental Functions	7-25
References	7-28

## 1. FUNCTIONS OF A COMPLEX VARIABLE

**Complex Numbers.** Throughout Chap. 7,  $z = x + iy$  and  $w = u + iv$  denote complex numbers;  $i$  is the *imaginary unit*,  $i^2 = -1$ ;  $x, y, u, v$  are arbitrary real numbers;  $x$  is the *real part* of  $z$ ,  $y$  the *imaginary part* of  $z$ :

$$(1) \quad x = \operatorname{Re}(x + iy), \quad y = \operatorname{Im}(x + iy).$$

The complex numbers  $z$  can be represented geometrically by the points  $(x, y)$  of an  $xy$ -plane (or  $z$ -plane), as in Fig. 1. The polar coordinates  $(r, \theta)$  of  $z$  are termed respectively the *modulus* (or *absolute value*) of  $z$  and *argument* (or *amplitude*) of  $z$ :

$$(2) \quad r = |z| = \operatorname{mod} z; \quad \theta = \arg z = \operatorname{amp} z; \quad z = r(\cos \theta + i \sin \theta).$$

The conjugate of  $z = x + iy$  is:

$$(3) \quad \bar{z} = x - iy.$$

Algebraic properties of complex numbers are discussed in Chap. 2. In general, complex numbers are combined as are real numbers, with the relation  $i^2 = -1$  used to simplify the results. *Addition* is the same as vector

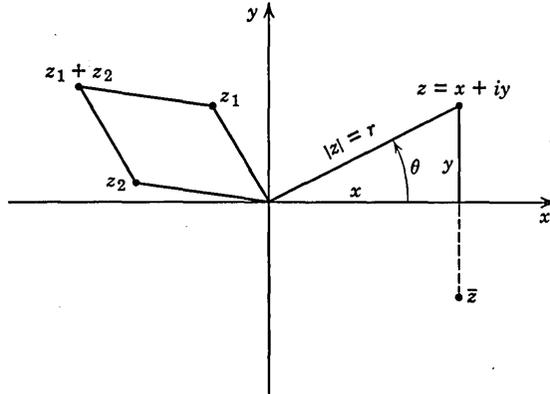


FIG. 1. The complex  $z$ -plane.

addition (Fig. 1). *Multiplication* of  $z_1$  by  $z_2$  yields a number  $z_1 \cdot z_2$  whose modulus is  $|z_1| \cdot |z_2|$  and whose argument is  $\arg z_1 + \arg z_2$ .

#### Useful Rules.

$$(4) \quad \begin{aligned} \overline{z_1 + z_2} &= \bar{z}_1 + \bar{z}_2, & \overline{z_1 \cdot z_2} &= \bar{z}_1 \cdot \bar{z}_2, \\ z + \bar{z} &= 2 \operatorname{Re}(z), & z - \bar{z} &= 2i \operatorname{Im}(z), & z \cdot \bar{z} &= |z|^2, \\ |z_1 + z_2| &\leq |z_1| + |z_2|, & |z_2 - z_1| &\geq ||z_2| - |z_1||, \\ z^n &= [r(\cos \theta + i \sin \theta)]^n = r^n(\cos n\theta + i \sin n\theta), & n &= \pm 1, \pm 2, \dots \end{aligned}$$

**Complex Functions.** By a *function* of the complex variable  $z$  will be meant an assignment of a value  $w$  to each  $z$  of a certain set  $D$  in the  $z$ -plane (see Chap. 1, Sects. 1 and 3); one then writes:

$$(5) \quad w = f(z).$$

(Some formulas will assign several values of  $w$  to each  $z$  in  $D$ . One then speaks of a "multiple-valued function.") The set  $D$  is generally an *open region* (e.g., interior of a circle); see Chap. 1, Sect. 8. From the equation  $u + iv = f(x + iy)$  one deduces two equations of the form

$$(6) \quad u = u(x, y), \quad v = v(x, y) \quad ((x, y) \text{ in } D),$$

and conversely a pair (6) of functions of two variables determines a complex function (5) of  $z$ .

*Limits and continuity* for complex functions are defined as for real functions. The phrase “ $z$  approaches  $z_0$ ” is interpreted to mean:  $|z - z_0| \rightarrow 0$ , or that the distance from  $z$  to  $z_0$  becomes arbitrarily small. The basic theorems on sums, products, quotients hold without change from the real case. Continuity of  $w = f(z)$  is equivalent to continuity of both  $u(x, y)$  and  $v(x, y)$  in (6).

Each complex function  $w = f(z)$  can be interpreted as a *mapping* (Chap. 10) of the set  $D$  into a set  $E$  in the  $w$ -plane. If  $f(z)$  is continuous, then as  $z$  traces a curve in the  $z$ -plane,  $w$  traces a curve in the  $w$ -plane.

*Derivatives* of complex functions are defined as for real functions:

$$(7) \quad \frac{d}{dz} f(z) = f'(z) = \lim_{\Delta z \rightarrow 0} \frac{f(z + \Delta z) - f(z)}{\Delta z},$$

and the formal rules of differentiation carry over. Higher derivatives  $f''(z), \dots$  are defined similarly.

*Definite integrals* of complex functions are defined as line integrals:

$$(8) \quad \int_C^{z_2} f(z) dz = \int_C (u + iv)(dx + i dy) \\ = \int_C u dx - v dy + i \int_C v dx + u dy.$$

Here  $C$  is a continuous path of finite length from  $z_1$  to  $z_2$ . Again the formal rules carry over.

**Examples of Complex Functions** are the following:

*polynomials:*  $w = a_0 z^n + \dots + a_{n-1} z + a_n,$

*rational functions:*  $w = \frac{a_0 z^n + \dots + a_n}{b_0 z^m + \dots + b_m},$

*exponential function:*  $w = e^z = e^x (\cos y + i \sin y) = \exp z,$

(9) *logarithm:*  $w = \log z = \log |z| + i \arg z \quad (z \neq 0),$

*power function:*  $w = z^a = \exp (a \log z),$

*trigonometric functions:*  $\sin z = \frac{e^{iz} - e^{-iz}}{2i}, \quad \cos z = \frac{e^{iz} + e^{-iz}}{2},$

*hyperbolic functions:*  $\sinh z = \frac{e^z - e^{-z}}{2}, \quad \cosh z = \frac{e^z + e^{-z}}{2},$

inverse trigonometric functions:

$$\sin^{-1} z = \frac{1}{i} \log (iz \pm \sqrt{1 - z^2}), \quad \cos^{-1} z = \frac{1}{i} \log (z \pm i\sqrt{1 - z^2})$$

The logarithm is a *multiple-valued function* and can be made single-valued (so that continuity can be discussed) by properly restricting  $z$  and the choice of  $\arg z$ . The principal value is:

$$(10) \quad \log z = \log |z| + i\theta, \quad (r > 0, -\pi < \theta \leq \pi),$$

a function continuous except for  $\theta = \pi$ .

If  $a$  is a rational number (e.g.,  $\frac{2}{3}$ ),  $z^a$  has a finite number of values. For example,  $z^{1/2} = \sqrt{z}$  has two values:

$$(11) \quad \begin{aligned} z^{1/2} &= e^{1/2 \log z} = e^{1/2(\log r + i \arg z)} = \sqrt{r} e^{(1/2)i \arg z} \\ &= \sqrt{r} \left( \cos \frac{\theta}{2} + i \sin \frac{\theta}{2} \right) \text{ or } \sqrt{r} \left( \cos \left( \frac{\theta}{2} + \pi \right) + i \sin \left( \frac{\theta}{2} + \pi \right) \right), \end{aligned}$$

if  $\theta$  is one choice of  $\arg z$ .

**Identities** satisfied by the exponential function, logarithm, and trigonometric functions:

$$(12) \quad \begin{aligned} e^{z_1+z_2} &= e^{z_1} \cdot e^{z_2}, \quad e^{z_1-z_2} = e^{z_1} \div e^{z_2}, \quad (e^z)^n = e^{nz} \quad (n = \pm 1, \pm 2, \dots), \\ \log(z_1 \cdot z_2) &= \log z_1 + \log z_2, \quad \log z^n = n \log z, \\ \sin(z_1 + z_2) &= \sin z_1 \cos z_2 + \cos z_1 \sin z_2, \quad \sin^2 z + \cos^2 z = 1, \dots \end{aligned}$$

In the case of the logarithm, the identities are true only for proper choice of value of each logarithm concerned. The rules for differentiation also carry over:

$$(13) \quad \frac{d}{dz} (z^n) = nz^{n-1}, \quad \frac{d}{dz} e^z = e^z, \quad \frac{d}{dz} \sin z = \cos z, \dots$$

## 2. ANALYTIC FUNCTIONS. HARMONIC FUNCTIONS

The function  $w = f(z)$  is said to be *analytic* (regular, holomorphic) in an open region  $D$  if it has a derivative  $f'(z)$  in  $D$ . The function  $f(z)$  is analytic in  $D$  if and only if  $u = \operatorname{Re}(f(z))$  and  $v = \operatorname{Im}(f(z))$  have continuous

first partial derivatives in  $D$  and the *Cauchy-Riemann equations* hold in  $D$ :

$$(14) \quad \frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}.$$

Furthermore, if  $f(z)$  is analytic,

$$(15) \quad f'(z) = \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x} = \frac{\partial v}{\partial y} + i \frac{\partial v}{\partial x} = \dots.$$

If  $f(z)$  is analytic in  $D$ , the derivatives of all orders of  $f$ ,  $u$ ,  $v$  exist and are continuous in  $D$ . From eq. (14) one deduces that

$$(16) \quad \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0;$$

that is,  $u$  and  $v$  are *harmonic functions*. Relations (14) are described by the statement: " $u$  and  $v$  form a pair of conjugate harmonic functions." One says " $v$  is conjugate to  $u$ ," but should note that  $u$  is conjugate to  $-v$ .

In polar coordinates (14) and (15) become

$$(17) \quad \frac{\partial u}{\partial r} = \frac{1}{r} \frac{\partial v}{\partial \theta}, \quad \frac{1}{r} \frac{\partial u}{\partial \theta} = -\frac{\partial v}{\partial r},$$

$$(18) \quad f'(z) = e^{-i\theta} \left( \frac{\partial u}{\partial r} + i \frac{\partial v}{\partial r} \right).$$

All the functions (9) are analytic, provided the logarithms are restricted so as to be continuous and division by zero is excluded. A function analytic for all  $z$  is called an *entire* function or an *integral* function; examples are polynomials and  $e^z$ .

A function cannot be analytic merely at a single point or merely along a curve. The definition requires always that analyticity holds in an open region. The phrases "analytic at  $z_0$ " or "analytic along curve  $C$ " are understood to mean "analytic in an open region containing  $z_0$ " or "analytic in an open region containing curve  $C$ ." If  $f$  is analytic in an open region  $D$ , then the values  $w = f(z)$  form an open region in the  $w$ -plane.

### 3. INTEGRAL THEOREMS

The open region  $D$  is termed *simply connected* if every simple closed path  $C$  in  $D$  (Fig. 2) has its interior in  $D$ . If  $D$  is not simply connected it is

*multiply connected*; for example, the region between two concentric circles is multiply connected; it is *doubly connected*, because its boundary is formed of two pieces or "components."

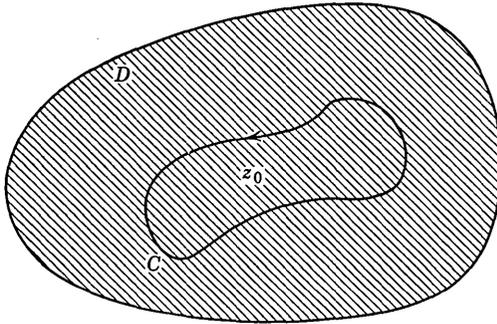


FIG. 2. Simply connected region.

All paths in the following line integrals are assumed to be "rectifiable," i.e., to have finite length.

**CAUCHY INTEGRAL THEOREM.** *If  $f(z)$  is analytic in a simply connected open region  $D$ , then*

$$\oint_C f(z) dz = 0$$

*on every simple closed path  $C$  in  $D$  or, equivalently,  $\int f(z) dz$  is independent of path in  $D$ .*

**MORERA'S THEOREM (converse of Cauchy theorem).** *If  $f(z)$  is continuous in the open region  $D$  and*

$$\oint_C f(z) dz = 0$$

*on every simple closed path  $C$  in  $D$ , then  $f(z)$  is analytic in  $D$ .*

An indefinite integral of  $f(z)$  is a function  $F(z)$  whose derivative is  $f(z)$ . If  $f(z)$  is continuous in  $D$  and has an indefinite integral  $F(z)$ , then

$$(19) \quad \int_{z_1}^{z_2} f(z) dz = F(z_2) - F(z_1);$$

in particular, the integral is independent of path, so that  $f(z)$  must be analytic; since  $F'(z) = f(z)$ ,  $F(z)$  must also be analytic. If  $f(z)$  is a given analytic function in  $D$ , then existence of an indefinite integral of  $f(z)$  can be

proved, provided  $D$  is simply connected. In particular,

$$(20) \quad F(z) = \int_{z_0}^z f(z) dz \quad (z_0 \text{ in } D)$$

has meaning, since the integral is independent of path, and  $F'(z) = f(z)$ , so that  $F$  is an indefinite integral.

CAUCHY INTEGRAL FORMULAS. Let  $f(z)$  be analytic in  $D$ . Let  $C$  be a simple closed path in  $D$  and having its interior in  $D$ . Let  $z_0$  be interior to  $C$  (Fig. 2). Then

$$(21) \quad \begin{aligned} f(z_0) &= \frac{1}{2\pi i} \oint_C \frac{f(z)}{z - z_0} dz, & f'(z_0) &= \frac{1}{2\pi i} \oint_C \frac{f(z)}{(z - z_0)^2} dz, \dots, \\ f^{(n)}(z_0) &= \frac{n!}{2\pi i} \oint_C \frac{f(z)}{(z - z_0)^{n+1}} dz, \dots \end{aligned}$$

At the heart of this theorem is the special case  $f(z) \equiv 1$ :

$$2\pi i = \oint_C \frac{dz}{z - z_0}; \quad 0 = \oint_C \frac{dz}{(z - z_0)^n}, \quad n = 2, 3, \dots$$

Cauchy's theorem and integral formulas can be extended to multiply connected domains. Let  $D$  be a domain bounded by curves  $C_1, C_2, \dots, C_k$

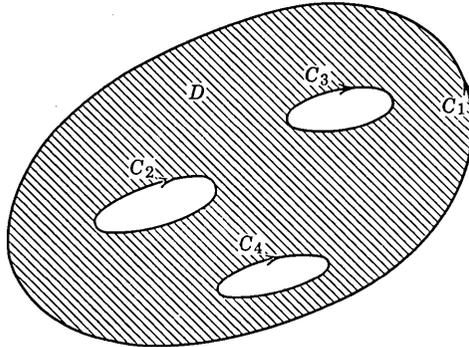


FIG. 3. Multiply connected region.

as in Fig. 3. Let  $f(z)$  be analytic in a somewhat larger region, including all of  $D$  and its boundary. Then

$$(22) \quad \oint_{C_1} f(z) dz + \oint_{C_2} f(z) dz + \dots + \oint_{C_n} f(z) dz = 0;$$

that is, the integral of  $f(z)$  around the complete boundary  $B$  of  $D$  is zero,

provided one integrates on the boundary in the direction which keeps the region  $D$  "on the left":

$$(23) \quad \int_B f(z) dz = 0.$$

Under the same conditions, if  $z_0$  is in  $D$ ,

$$(24) \quad f(z_0) = \frac{1}{2\pi i} \int_B \frac{f(z)}{z - z_0} dz, \quad f^{(n)}(z_0) = \frac{n!}{2\pi i} \int_B \frac{f(z)}{(z - z_0)^{n+1}} dz.$$

**CAUCHY INEQUALITIES.** Under the hypotheses stated above for eqs. (21), let  $|f(z)| = M$  on  $C$  and let  $C$  be a circle with center  $z_0$  and radius  $R$ . Then

$$(25) \quad |f^{(n)}(z_0)| \leq \frac{Mn!}{R^n} \quad (n = 0, 1, 2, \dots).$$

**LILOVILLE THEOREM.** If  $f(z)$  is analytic for all finite  $z$  and  $|f(z)| \leq M$ , where  $M$  is a constant, for all  $z$ , then  $f(z)$  is identically constant.

**MAXIMUM PRINCIPLE.** Let  $f(z)$  be analytic in the open region  $D$ . If  $|f(z)|$  has a weak relative maximum at a point  $z_0$  of  $D$  (that is, if  $|f(z)| \leq |f(z_0)|$  for  $z$  sufficiently close to  $z_0$ ), then  $f(z)$  is identically constant.

For proofs of these theorems see Refs. 2, 3, 8.

#### 4. POWER SERIES. LAURENT SERIES

Infinite series whose terms are complex numbers are defined as for real numbers and, in general, the theory of convergence is the same. In particular, a series  $\sum_{n=1}^{\infty} b_n$  of complex numbers is termed *absolutely convergent* if the series of real numbers  $\sum |b_n|$  converges. *Absolute convergence implies convergence.*

**Power Series.** A power series in  $z$  has the form

$$(26) \quad \sum_{n=0}^{\infty} c_n(z - z_0)^n,$$

where  $z_0$  is fixed. Each such series has a radius of convergence  $\rho$ ,  $0 \leq \rho \leq +\infty$ . If  $\rho = 0$ , the series converges only for  $z = z_0$ . Otherwise, the series converges (in fact, absolutely) for  $|z - z_0| < \rho$ , i.e., inside the *circle of convergence* (whose radius  $\rho$  may be infinite). Outside this circle, for  $|z - z_0| > \rho$ , the series diverges. On the circle:  $|z - z_0| = \rho$ , the series may converge at some points and diverge at others. The radius can be evaluated by the formulas:

$$(27) \quad \rho = \lim_{n \rightarrow \infty} \left| \frac{c_n}{c_{n+1}} \right|, \quad \rho = \lim_{n \rightarrow \infty} \frac{1}{\sqrt[n]{|c_n|}},$$

provided the limit exists, and in any case by the formula

$$(28) \quad \rho = \liminf_{n \rightarrow \infty} \frac{1}{\sqrt[n]{|c_n|}},$$

where  $\liminf$  denotes the lower limit.

Let the power series (26) have radius of convergence  $\rho > 0$ , so that its sum is a well-defined function  $f(z)$  inside the circle of convergence. One can then prove that the series converges uniformly to  $f(z)$  in each circle  $|z - z_0| \leq \rho' < \rho$ , so that  $f(z)$  is continuous. Furthermore, the differentiated series  $\sum n c_n (z - z_0)^{n-1}$  converges uniformly in each circle  $|z - z_0| \leq \rho' < \rho$ . From this it follows that the differentiated series converges to  $f'(z)$  and that  $f'(z)$  is continuous. Hence  $f(z)$  is itself analytic for  $|z - z_0| < \rho$ . *Every power series defines an analytic function inside its circle of convergence.* In general, all derivatives of  $f(z)$  can be evaluated by repeated differentiation of the series. One hence concludes that

$$(29) \quad c_n = \frac{f^{(n)}(z_0)}{n!};$$

that is, the power series is the Taylor series of  $f(z)$ . From this it follows that equality of the sum of two power series:

$$\sum_{n=0}^{\infty} c_n (z - z_0)^n = \sum_{n=0}^{\infty} C_n (z - z_0)^n, \quad |z - z_0| < \rho,$$

implies equality of corresponding coefficients:

$$c_n = C_n \quad (n = 0, 1, 2, \dots).$$

Now let  $f(z)$  be given as an analytic function in an open region  $D$  of arbitrary shape and let  $z_0$  be a point of  $D$ . With  $z_0$  as center one can then construct a circle of maximum radius  $r_0$  having its interior in  $D$  (Fig. 4).

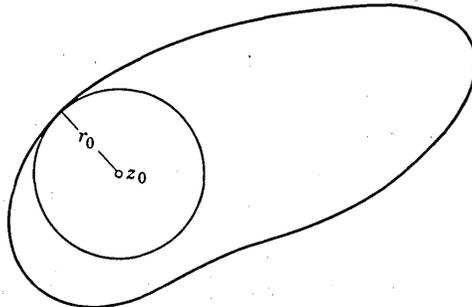


FIG. 4. Taylor series expansion.

Within this circle  $f(z)$  can be represented by a power series, its Taylor series about  $z_0$ :

$$(30) \quad f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!} (z - z_0)^n, \quad |z - z_0| < r_0;$$

the series may have a radius of convergence  $\rho$  larger than  $r_0$ . From this theorem one deduces the following expansions:

$$\begin{aligned} e^z &= \sum_{n=0}^{\infty} \frac{z^n}{n!}, \text{ all } z; \quad \sin z = z - \frac{z^3}{3!} + \cdots, \text{ all } z; \\ \cos z &= 1 - \frac{z^2}{2!} + \cdots, \text{ all } z; \\ (31) \quad \log(1+z) &= z - \frac{z^2}{2} + \frac{z^3}{3} - \cdots, \quad |z| < 1; \\ \frac{1}{1-z} &= 1 + z + z^2 + \cdots, \quad |z| < 1; \\ (1+z)^k &= 1 + kz + \frac{k(k-1)}{2!} z^2 + \cdots, \quad |z| < 1. \end{aligned}$$

**Laurent Series.** A series of form

$$\sum_{n=1}^{\infty} \frac{b_n}{(z - z_0)^n}$$

is reducible by the substitution  $z' = 1/(z - z_0)$  to the form of an ordinary power series and accordingly converges for  $|z'| < \rho$ , i.e., for  $|z - z_0| > \rho_1 = \frac{1}{\rho}$ . If now a series  $\sum_{n=0}^{\infty} a_n(z - z_0)^n$  converges for  $|z - z_0| < \rho_2$  and  $\rho_1 < \rho_2$ , then the sum

$$\sum_{n=1}^{\infty} \frac{b_n}{(z - z_0)^n} + \sum_{n=0}^{\infty} a_n(z - z_0)^n$$

has meaning for  $\rho_1 < |z - z_0| < \rho_2$ , that is, in a certain annular region  $D$  (Fig. 5). Here  $\rho_1$  may be 0 and  $\rho_2$  may be  $+\infty$ . Let the sum be  $f(z)$ , so that  $f(z)$  is analytic in  $D$ . If one writes  $b_n = a_{-n}$  ( $n = 1, 2, \dots$ ); then one has

$$(32) \quad f(z) = \sum_{n=-\infty}^{-1} a_n(z - z_0)^n + \sum_{n=0}^{\infty} a_n(z - z_0)^n = \sum_{n=-\infty}^{\infty} a_n(z - z_0)^n.$$

This series is termed the *Laurent expansion* of  $f(z)$ . The coefficients can be shown to be uniquely determined as follows:

$$(33) \quad a_n = \frac{1}{2\pi i} \oint_C \frac{f(z) dz}{(z - z_0)^{n+1}} \quad n = 0, \pm 1, \pm 2, \dots,$$

where  $C$  is any path about the ring, as in Fig. 5.

If  $f(z)$  is an arbitrary function analytic in a ring domain  $D$ , then one can compute the coefficients  $a_n$  by eq. (33) and form the Laurent series,

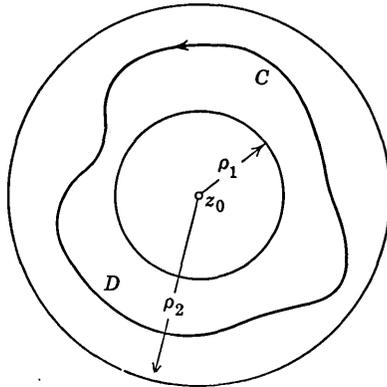


FIG. 5. Laurent expansion.

which will then converge to  $f(z)$  in  $D$ . In practice there are easier ways of obtaining the coefficients. One way is to write  $f(z)$  as the sum of two functions  $f_2(z), f_1(z)$ , the first analytic for  $|z - z_0| < \rho_2$ , the second analytic for  $|z - z_0| > \rho_1$  and approaching 0 as  $|z| \rightarrow \infty$ . Under the substitution  $\zeta = 1/(z - z_0)$ ,  $f_1(z)$  becomes a function of  $\zeta$  analytic for  $|\zeta| < 1/\rho_1$ , so that  $f_1(z) = \sum_{n=1}^{\infty} b_n \zeta^n$  or

$$(34) \quad f_1(z) = \sum_{n=1}^{\infty} \frac{b_n}{(z - z_0)^n}, \quad |z - z_0| > \rho_1.$$

For  $f_2(z)$  one has a Taylor series about  $z_0$ . Addition of the two series provides the desired Laurent series. For example, if  $f(z) = 1/(z - 1)(z - 2)$  and  $D$  is the ring  $1 < |z| < 2$ , then one can choose  $f_1(z) = -1/(z - 1)$ ,  $f_2(z) = 1/(z - 2)$ .

### 5. ZEROS. SINGULARITIES. RESIDUES. ARGUMENT PRINCIPLE

**Zeros.** Let  $f(z)$  be analytic in domain  $D$  and let  $f(z_0) = 0$ . Then  $z_0$  is called a *zero* of  $f(z)$ . If  $f(z)$  is not identically zero, then each zero has a

definite order (or multiplicity)  $n$ , a positive integer, and  $f(z) = (z - z_0)^n g(z)$  where  $g$  is analytic in  $D$  and  $g(z_0) \neq 0$ . The order  $n$  is the smallest value of  $k$  such that  $f^{(k)}(z_0) \neq 0$ . If  $f(z)$  is not identically zero, then each zero of  $f(z)$  is *isolated*; that is, for each zero  $z_0$  one can choose a circular region  $|z - z_0| < a$  containing no other zero.

**Singularities.** Let  $f(z)$  be not identically zero and have a zero of order  $n$  at  $z_0$ . Then  $h(z) = 1/f(z)$  is analytic in some circular region  $|z - z_0| < a$  except at the center  $z_0$ . By definition,  $h(z)$  has a *pole of order  $n$*  at  $z_0$ . One can write:  $h(z) = (z - z_0)^{-n} p(z)$ , where  $p(z)$  is analytic for  $|z - z_0| < a$  and  $p(z_0) \neq 0$ . Since  $f(z_0) = 0$ ,  $|h(z)| \rightarrow \infty$  as  $z \rightarrow z_0$ . One conventionally assigns the value  $\infty$  to  $h(z)$  at  $z_0$ .

In general, let  $f(z)$  be analytic in a punctured disk:  $0 < |z - z_0| < a$ , but not at  $z_0$ . Then  $f(z)$  is said to have an *isolated singularity* at  $z_0$ . One can form a Laurent expansion of  $f(z)$  in the ring domain  $\rho_1 = 0 < |z - z_0| < a = \rho_2$ . Three cases can then arise.

I. *No negative powers in the Laurent series.* Then

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n,$$

so that  $f(z)$  can be treated as a function analytic for  $|z - z_0| < a$  without exception. The singularity is termed *removable*. The new value of  $f(z)$  at  $z_0$  is  $a_0 = \lim_{z \rightarrow z_0} f(z)$ .

II. *A finite number of negative powers in the Laurent series.* Here, for proper choice of  $N$ ,

$$\begin{aligned} f(z) &= \frac{a_{-N}}{(z - z_0)^N} + \cdots + \frac{a_{-1}}{z - z_0} + a_0 + a_1(z - z_0) + \cdots \\ (35) \quad &= \frac{g(z)}{(z - z_0)^N}, \quad g(z_0) = a_{-N} \neq 0. \end{aligned}$$

Hence  $f(z)$  has a *pole* of order  $N$  at  $z_0$ .

III. *Infinitely many negative powers in the Laurent series.* In this case  $f(z)$  is said to have an *essential singularity* at  $z_0$ .

By a theorem of Riemann, the three cases can be distinguished as follows: I.  $|f(z)|$  is bounded for  $0 < |z - z_0| < b$  for some  $b$ . II.  $|f(z)| \rightarrow \infty$  as  $z \rightarrow z_0$ . III. Neither  $|f(z)|$  nor  $1/|f(z)|$  is bounded in each punctured disk  $0 < |z - z_0| < b$ . In Case III, by a theorem of Weierstrass and Casorati,  $f(z)$  comes arbitrarily close to every complex number in every neighborhood of  $z_0$ .

If  $f(z)$  is analytic for  $|z| > R$ , then  $f(z)$  is considered to have an *isolated singularity at  $z = \infty$* . A Laurent expansion is available, with  $\rho_1 = R$  and  $\rho_2 = \infty$ . The classification is similar to the above, with "negative" replaced by "positive." Also the type of singularity of  $f(z)$  at  $\infty$  is the same as that of  $f(1/z)$  at  $z = 0$ .

A function analytic for all finite  $z$  except for poles is termed a *meromorphic function*.

**Residues.** The residue of  $f(z)$  at an isolated singularity  $z_0$  is defined as

$$(36) \quad \text{Res } [f(z), z_0] = \frac{1}{2\pi i} \int_C f(z) dz,$$

where  $C$  is a circle  $|z - z_0| = c$ , enclosing no singularity other than  $z_0$ , and the integration is in the counterclockwise direction. The residue of  $f(z)$  at  $z = \infty$ , denoted by  $\text{Res } [f(z), \infty]$ , is defined by the same integral, where  $C$  is a circle  $|z| = c$  outside of which  $f(z)$  has no singularity other than  $\infty$  and where the integration is in the *clockwise* direction. If  $z_0$  is finite,

$$(37) \quad \text{Res } [f(z), z_0] = a_{-1},$$

where  $a_{-1}$  is the coefficient of  $(z - z_0)^{-1}$  in the Laurent expansion about  $z_0$ . If  $z_0$  is  $\infty$ ,

$$(38) \quad \text{Res } [f(z), \infty] = -a_{-1},$$

where  $a_{-1}$  is the coefficient of  $z^{-1}$  in the Laurent expansion of  $f(z)$  for  $|z| > R$ .

The CAUCHY RESIDUE THEOREM asserts that, if  $f(z)$  is analytic in an open region containing the path  $C$ , then

$$(39) \quad \oint_C f(z) dz = 2\pi i \cdot (\text{sum of residues of } f(z) \text{ inside } C),$$

provided  $f(z)$  is analytic inside  $C$  except for a finite number of isolated singularities. Similarly,

$$(40) \quad \oint_C f(z) dz = 2\pi i \cdot (\text{sum of residues of } f(z) \text{ outside } C, \text{ including } \text{Res } [f(z), \infty]),$$

provided  $f(z)$  is analytic outside  $C$  except for a finite number of isolated singularities. Hence, if  $f(z)$  is analytic for all  $z$ , except for a finite number of singularities, the sum of all residues of  $f(z)$ , including that at  $\infty$ , is 0.

**Calculation of residues** may be simplified by the following rules:

1. At a pole  $z_0$  of first order,

$$(41) \quad \text{Res } [f(z), z_0] = \lim_{z \rightarrow z_0} (z - z_0)f(z).$$

2. At a pole  $z_0$  of order  $N$  ( $N = 2, 3, \dots$ ),

$$(42) \quad \text{Res } [f(z), z_0] = \lim_{z \rightarrow z_0} \frac{g^{(N-1)}(z)}{(N-1)!},$$

where  $g(z) = (z - z_0)^N f(z)$ .

3. Let

$$(43) \quad f(z) = \frac{A(z)}{B(z)},$$

where  $A(z)$  and  $B(z)$  are analytic at  $z_0$ . If  $A(z_0) \neq 0$  and  $B(z)$  has a zero of first order at  $z_0$ , then

$$(44) \quad \text{Res } [f(z), z_0] = \frac{A(z_0)}{B'(z_0)}.$$

If  $A(z_0) \neq 0$  and  $B(z)$  has a zero of second order at  $z_0$ , then

$$(45) \quad \text{Res } [f(z), z_0] = \frac{6A'(z_0)B''(z_0) - 2A(z_0)B'''(z_0)}{3[B''(z_0)]^2}.$$

If  $A(z_0) \neq 0$  and  $B(z)$  has a zero of third order at  $z_0$ , then

$$(46) \quad \text{Res } [f(z), z_0] = \frac{120A''B'''^2 - 60A'B'''B^{iv} - 12AB'''B^v + 15AB^{iv}{}^2}{40B'''^3},$$

where all quantities are evaluated at  $z_0$ . If  $A(z)$  has a first order zero at  $z_0$  and  $B(z)$  a second order zero, then

$$(47) \quad \text{Res } [f(z), z_0] = \frac{2A'(z_0)}{B''(z_0)}.$$

**ARGUMENT PRINCIPLE.** Let  $f(z)$  be analytic in an open region  $D$  containing the simple closed path  $C$ ; let  $f(z)$  have at most a finite number of singularities inside  $C$ , all of which are poles, and let  $f(z) \neq 0$  on  $C$ . Then

$$(48) \quad \frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z)} dz$$

= number of zeros of  $f$  inside  $C$  - number of poles of  $f$  inside  $C$ ,

where zeros and poles are counted according to multiplicity.

The left-hand side of eq. (48) is termed the *logarithmic residue* of  $f(z)$  on  $C$ . It can be written as

$$\frac{1}{2\pi i} \oint_C d \log f.$$

As  $z$  traces  $C$ ,  $w = f(z)$  traces a path  $C_w$  in the  $w$ -plane. The integral  $\int d \log f(z)$  equals  $i$  times the total change in the argument of  $w$  as the path  $C_w$  is traced. Hence it equals  $2\pi i$  times the “winding number” of  $C_w$  about  $w = 0$ , i.e., the number of times that  $C_w$  effectively winds about  $w = 0$  in the positive direction.

**THE FUNDAMENTAL THEOREM OF ALGEBRA** (see Chap. 2). From the argument principle one deduces that every polynomial in  $z$  of degree  $N$  has precisely  $N$  zeros in the complex plane.

**ROUCHÉ’S THEOREM** may also be deduced: if both  $f_1(z)$  and  $f_2(z)$  are analytic in a simply connected open region containing the simple closed path  $C$  and  $|f_1(z) - f_2(z)| < |f_2(z)|$  on  $C$ , then  $f_1(z)$  and  $f_2(z)$  have the same number of zeros inside  $C$ .

**Evaluation of Definite Integrals by Residues.** A great variety of definite integrals can be evaluated with the aid of residues. For example, if  $R(u, v)$  is a rational function of  $u$  and  $v$ , then

$$(49) \quad \int_0^{2\pi} R(\sin \theta, \cos \theta) d\theta = \oint_{|z|=1} R\left(\frac{z^2 - 1}{2iz}, \frac{z^2 + 1}{2z}\right) \frac{dz}{iz}$$

and the integral on the right can be computed by residues. Also, in general

$$(50) \quad \int_{-\infty}^{\infty} f(x) dx = 2\pi i \{\text{sum of residues of } f(z) \text{ in the half-plane } y > 0\},$$

provided  $f(z)$  is analytic for  $y \geq 0$  except for a finite number of points in  $y > 0$ ,  $\int_{-\infty}^{\infty} f(x) dx$  exists and

$$\lim_{R \rightarrow \infty} \int_0^\pi f(Re^{i\theta}) Re^{i\theta} d\theta = 0.$$

The last condition is satisfied if  $f(z)$  is rational and has a zero of order greater than 1 at  $\infty$ , or if  $f(z) = e^{miz}g(z)$  where  $m > 0$ ,  $g(z)$  is rational, and  $g(z)$  has a zero at  $\infty$ . For further applications one is referred to Chap. VI of Ref. 12.

## 6. ANALYTIC CONTINUATION

Let  $f_1(z)$  be analytic in the open region  $D_1$ ,  $f_2(z)$  in  $D_2$ . If  $D_2$  and  $D_1$  have a common part and  $f_1(z) = f_2(z)$  in that common part, then  $f_2(z)$  is said to be a *direct analytic continuation* of  $f_1(z)$  from  $D_1$  to  $D_2$ . Given  $f_1(z)$ ,  $D_1$ ,  $D_2$ , the function  $f_2(z)$  may or may not exist; however, if it does exist, there can be only one such function (*uniqueness* of analytic continuation).

Let  $D_1, D_2, \dots, D_n$  be regions such that each has a common part with the next and let  $f_j(z)$  be analytic in  $D_j$  ( $j = 1, \dots, n$ ). If  $f_j(z) = f_{j+1}(z)$  ( $j = 1, \dots, n - 1$ ) in the common part of  $D_j, D_{j+1}$ , then one says that  $f_1(z)$  has been continued analytically from  $D_1$  to  $D_n$  via  $D_2, \dots, D_{n-1}$  and calls  $f_n(z)$  an (*indirect*) *analytic continuation* of  $f_1(z)$ . Given  $f_1(z)$  and the regions  $D_1, \dots, D_n$ , there is at most one analytic continuation of  $f_1(z)$  to  $D_n$  via  $D_2, \dots, D_{n-1}$ . There may exist other continuations of  $f_1(z)$  to  $D_n$  via other chains of regions.

Given a function  $f(z)$  analytic in region  $D$ , one can form all possible continuations of  $f(z)$  to other regions. The totality of such continuations is said to form an *analytic function in the broad sense* (Weierstrassian analytic function). In this sense  $\log z$ ,  $\sqrt{z}$ ,  $\sin^{-1} z$  can each be considered as one analytic function. The importance of the concept is illustrated by the fact that every identity satisfied by  $f(z)$  will be satisfied by all its analytic continuations. The term "identity" includes linear differential equations with polynomial coefficients.

**Example of Analytic Continuation.** The functions

$$f_1(z) = \sum_{n=0}^{\infty} \frac{z^n}{2^{n+1}}, \quad |z| < 2,$$

$$f_2(z) = \sum_{n=0}^{\infty} \frac{(z+1)^n}{3^{n+1}}, \quad |z+1| < 3$$

are analytic continuations of each other. Indeed, both are power series expansions of  $f_3(z) = 1/(2-z)$  and have the same sum for  $|z| < 2$ . Also  $f_2(z)$  can be regarded as the Taylor series of  $f_1(z)$  about  $z = -1$ . This series happens to converge outside of  $|z| < 2$  and hence provides an analytic continuation.

**Analytic Continuation from Reals.** Let  $f_1(z)$  be defined only for  $y = 0$ ,  $a < x < b$ , i.e., only when  $z$  is real and between  $a$  and  $b$ . Let  $f_2(z)$  be analytic in an open region  $D$  which includes the interval of definition of  $f_1(z)$ . If  $f_2(z) = f_1(z)$  on this interval, then  $f_2(z)$  is said to be an analytic continuation of  $f_1(z)$  from reals. Again continuation, if possible, is unique. *Examples.*  $e^z$  as a continuation of  $e^x$ ,  $\sin z$  as a continuation of  $\sin x$ ,  $\log z$  as a continuation of  $\log x$ .

7. RIEMANN SURFACES

The function  $w = z^{1/2}$  can be considered as an analytic function in the broad sense; that is, it is formed of several functions which are analytic continuations of each other. The resulting totality has the defect that it is two-valued: for each  $z$ , there are two possible values (except for  $z = 0$ ). To remedy this defect one regards  $w = z^{1/2}$  as a function defined not in the  $z$ -plane, but on a Riemann surface over the  $z$ -plane. In this case, the Riemann surface can be constructed as follows. One takes two copies of the  $z$ -plane, calling them Sheet I and Sheet II. Each sheet is considered as cut open along a *branch line*, the positive real axis. Sheet II is placed directly over Sheet I, with axes in the same position, and then the two sheets are attached by joining upper edge of the cut line of each sheet to the lower edge of the cut line of the other, as suggested in Fig. 6. Un-

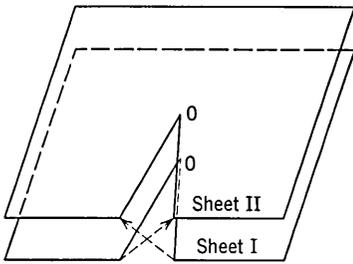


FIG. 6. Riemann surface of  $z^{1/2}$ .

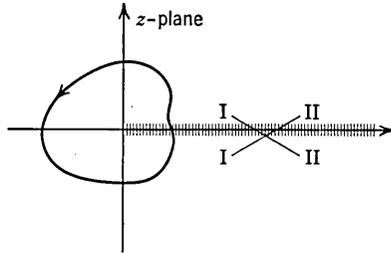


FIG. 7. Branch line for  $w = z^{1/2}$ .

fortunately this cannot be carried out in space. For each point in the  $z$ -plane, one has then two points in the Riemann surface, one in each sheet. As one traces a path about  $z = 0$  in the  $z$ -plane, one can describe a corresponding path in the Riemann surface by assigning a sheet to each position; no change of sheet can be made except when crossing the branch line, and a change of sheet must be made at such a crossing (Fig. 7). A closed path in the  $z$ -plane will not in general lead to a closed path on the Riemann surface. A path which closes up after two encirclements of the origin will be closed on the Riemann surface. The origin itself appears as a point common to the two sheets and is termed a *branch point*.

On the Riemann surface just constructed one can now define  $\sqrt{z}$  as a single-valued function as follows:  $\sqrt{z} = \sqrt{r} e^{i\theta/2}$ ,  $0 < \theta < 2\pi$ , on Sheet I;  $z = \sqrt{r} e^{i\theta/2}$ ,  $2\pi < \theta < 4\pi$ , on Sheet II. Above the branch line continuity determines the proper value to be assigned.

The procedure described can be generalized to

$$w = \sqrt[3]{z}, \quad w = \sqrt{(z - 1)(z - 2)(z - 3)}$$

and to all algebraic functions. In general  $n$  sheets will be required and several branch lines and branch points. The surface for

$$w = \sqrt{(z-1)(z-2)(z-3)}$$

is suggested in Fig. 8.

The procedure can be extended to nonalgebraic functions, but in general infinitely many sheets are required. An important case is  $\log z$ , for

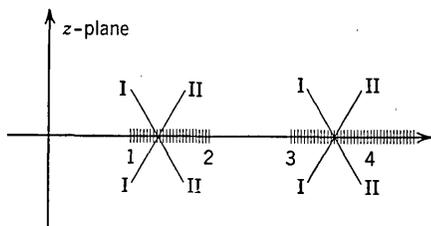


FIG. 8. Riemann surface of  $w = [(z-1)(z-2)(z-3)]^{1/2}$ .

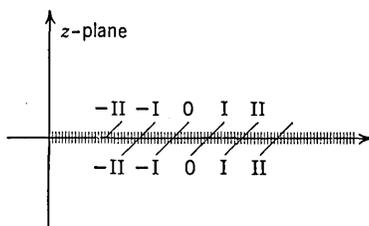


FIG. 9. Riemann surface of  $\log z$ .

which sheets  $0, \pm I, \pm II, \dots$  are needed, as in Fig. 9. In this case  $z = 0$  is a *logarithmic branch point* and is not regarded as a point of the Riemann surface.

## 8. ELLIPTIC FUNCTIONS

Let  $f(z)$  be a meromorphic function (analytic except for poles);  $f(z)$  is said to have period  $\omega$ ,  $\omega \neq 0$ , if  $f(z + \omega) = f(z)$  for all  $z$ ;  $f(z)$  is called an *elliptic* or *doubly periodic* function if  $f$  is not constant and has periods  $\omega_1, \omega_2$  and if  $\omega_1/\omega_2$  is not real. It then follows that  $n_1\omega_1 + n_2\omega_2$  are also periods, for every choice of the integers  $n_1, n_2$ . For proper choice of  $\omega_1, \omega_2$  these are all the periods of  $f$  and it will always be assumed that  $\omega_1, \omega_2$  are so chosen. The numbers  $\Omega = n_1\omega_1 + n_2\omega_2$  form the vertices of a paving of the plane by parallelograms, any one of which can be chosen as a *period parallelogram* of  $f(z)$ ; it is convenient to exclude the points on a pair of adjacent sides from each period parallelogram. It can be proved that  $f(z)$  has a finite number  $N$  of poles (counted according to multiplicity) in a period parallelogram;  $N$  is the *order* of  $f(z)$  as an elliptic function;  $N$  is always at least 2. In general,  $f(z) - a$  has  $N$  zeros in the parallelogram.

**Jacobian Elliptic Functions.** Examples of elliptic functions are provided by the functions

$$\operatorname{sn} z, \operatorname{cn} z, \operatorname{dn} z$$

of Jacobi. These can be defined as follows. For fixed  $k$ ,  $0 < k < 1$ , let

$$(51) \quad F(w) = \int_0^w \frac{dt}{\sqrt{1 - k^2 \sin^2 t}}, \quad E(w) = \int_0^w \sqrt{1 - k^2 \sin^2 t} dt;$$

$F$  and  $E$  are called the *elliptic integrals* of Legendre of first and second type, respectively. The equation  $z = F(w)$  can be solved for  $w$  to yield a multiple-valued function  $w$  of  $z$ ,  $w = \operatorname{am} z$ ; finally

$$(52) \quad \begin{aligned} \operatorname{sn} z &= \sin(\operatorname{am} z), & \operatorname{cn} z &= \cos(\operatorname{am} z), \\ \operatorname{dn} z &= \sqrt{1 - k^2 \operatorname{sn}^2 z}. \end{aligned}$$

Despite the multiple-valued operations, the functions  $\operatorname{sn} z$ ,  $\operatorname{cn} z$ ,  $\operatorname{dn} z$  are defined as single-valued functions, analytic except for poles and with power series expansions about  $z = 0$ :

$$(53) \quad \begin{aligned} \operatorname{sn} z &= z - (1 + k^2) \frac{z^3}{3!} + \dots, \\ \operatorname{cn} z &= 1 - \frac{z^2}{2!} + (1 + 4k^2) \frac{z^4}{4!} + \dots, \\ \operatorname{dn} z &= 1 - \frac{k^2 z^2}{2!} + \dots. \end{aligned}$$

All these functions depend on the parameter  $k$ . They have periods  $\omega_1, \omega_2$  as follows; for  $\operatorname{sn} z$ ,  $\omega_1 = 4K$  and  $\omega_2 = 2iK'$ ; for  $\operatorname{cn} z$ ,  $\omega_1 = 4K$ ,  $\omega_2 = 2K + 2iK'$ ; for  $\operatorname{dn} z$ ,  $\omega_1 = 2K$ ,  $\omega_2 = 4iK'$ ; here

$$(54) \quad K' = \int_0^{\pi/2} \frac{dt}{\sqrt{\cos^2 t + k^2 \sin^2 t}}, \quad K = F\left(\frac{\pi}{2}\right).$$

Hence  $\operatorname{sn} z$ ,  $\operatorname{cn} z$ ,  $\operatorname{dn} z$  are elliptic functions. Tables of the functions  $F$ ,  $E$ ,  $\operatorname{sn} z$ ,  $\operatorname{cn} z$ ,  $\operatorname{dn} z$  are available; see, for example, G. W. and R. M. Spenceley: *Smithsonian Elliptic Function Tables* (The Smithsonian Institution, Washington, 1947), and L. M. Milne-Thompson: *Jacobian Elliptic Function Tables* (Dover, New York, 1950).

It can be shown that the most general elliptic function is expressible simply in terms of  $\operatorname{sn} z$  and  $\operatorname{cn} z$ . Furthermore, a large class of integrals, called *elliptic integrals*, can be expressed in terms of the functions  $F$  and  $E$  and the integral of third kind:

$$(55) \quad \pi(w) = \int_0^w \frac{dt}{(1 + \alpha^2 \sin^2 t) \sqrt{1 - k^2 \sin^2 t}},$$

which depends on  $k$  and the real parameter  $\alpha$ . The general elliptic integral has form  $\int R(z, \sqrt{P(z)}) dz$ , where  $R(z, w)$  is a rational function of  $z$  and  $w$ , and  $P$  is a polynomial of degree three or four with distinct roots.

In the theory of elliptic functions a number of additional functions appear, of which the following are important definitions and relations:

**The Weierstrass  $\wp$ -Function.**

$$(56) \quad \wp(z) = \frac{1}{z^2} + \sum'_{\Omega} \left[ \frac{1}{(z - \Omega)^2} - \frac{1}{\Omega^2} \right].$$

Here  $\Omega$  stands for  $n_1\omega_1 + n_2\omega_2$  and  $\Sigma'$  indicates a sum over all choices of  $\Omega$  other than  $0 = 0\omega_1 + 0\omega_2$ ;  $\omega_1$  and  $\omega_2$  are complex parameters, assumed to have nonreal ratio, and  $\wp(z)$  is elliptic with periods  $\omega_1, \omega_2$ .

**The Weierstrass Zeta Function and Sigma Function.**

$$(57) \quad \zeta(z) = - \int \wp(z) dz = \frac{1}{z} + \sum'_{\Omega} \left[ \frac{1}{z - \Omega} + \frac{z}{\Omega^2} + \frac{1}{\Omega} \right];$$

$$(58) \quad \sigma(z) = \exp \int \zeta(z) dz = z \prod'_{\Omega} \left( 1 - \frac{z}{\Omega} \right) \exp \left[ \frac{z}{\Omega} + \frac{1}{2} \left( \frac{z}{\Omega} \right)^2 \right].$$

The  $\zeta$ -function is meromorphic but not elliptic; the  $\sigma$ -function is an entire function.

**The Jacobi Theta Functions.**

$$(59) \quad \begin{aligned} \theta_0(z) &= 1 + 2 \sum_{n=1}^{\infty} (-1)^n q^{n^2} \cos 2\pi n z, \\ \theta_1(z) &= 2 \sum_{n=0}^{\infty} (-1)^n q^{(n+1/2)^2} \sin (2n+1)\pi z, \\ \theta_2(z) &= 2 \sum_{n=0}^{\infty} q^{(n+1/2)^2} \cos (2n+1)\pi z, \\ \theta_3(z) &= 1 + 2 \sum_{n=1}^{\infty} q^{n^2} \cos 2\pi n z. \end{aligned}$$

Here  $q$  is a complex parameter of form  $e^{i\pi\tau}$ , where  $\tau$  has positive imaginary part. If  $\tau$  is chosen as  $\omega_2/\omega_1$ , then one has the identity:

$$(60) \quad \sigma(z) = \frac{\omega_1}{\theta'_1(0)} \exp \left[ \frac{\zeta(\frac{1}{2}\omega_1)}{\omega_1} z^2 \right] \theta_1 \left( \frac{z}{\omega_1} \right).$$

There are other  $\sigma$ -functions  $\sigma_\alpha(z)$ ,  $\alpha = 1, 2, 3$ , defined by the equations

$$(61) \quad \sigma_\alpha(z) = \frac{1}{\theta_{\alpha+1}(0)} \exp \left[ \frac{\zeta(\frac{1}{2}\omega_1)}{\omega_1} z^2 \right] \theta_{\alpha+1} \left( \frac{z}{\omega_1} \right) \quad (\alpha = 1, 2, 3),$$

where  $\theta_4$  is interpreted as  $\theta_0$ .

The Jacobian elliptic functions are expressible in terms of the  $\theta$ -functions or in terms of the  $\sigma$ -functions. Let

$$(62) \quad k = \left[ \frac{\theta_2(0)}{\theta_3(0)} \right]^2, \quad k' = \left[ \frac{\theta_0(0)}{\theta_3(0)} \right]^2,$$

$$(63) \quad K = \frac{\pi}{2} [\theta_3(0)]^2, \quad K' = -i\tau K.$$

Then

$$(64) \quad \begin{aligned} \operatorname{sn} z &= \frac{1}{\sqrt{k}} \frac{\theta_1 \left( \frac{z}{2K} \right)}{\theta_0 \left( \frac{z}{2K} \right)} = \frac{\sigma(z)}{\sigma_3(z)}, \\ \operatorname{cn} z &= \sqrt{\frac{k'}{k}} \frac{\theta_2 \left( \frac{z}{2K} \right)}{\theta_0 \left( \frac{z}{2K} \right)} = \frac{\sigma_1(z)}{\sigma_3(z)}, \\ \operatorname{dn} z &= \sqrt{k'} \frac{\theta_3 \left( \frac{z}{2K} \right)}{\theta_0 \left( \frac{z}{2K} \right)} = \frac{\sigma_2(z)}{\sigma_3(z)}. \end{aligned}$$

Here  $k' = \sqrt{1 - k^2}$  and  $K, K'$  are related to  $k$  by eqs. (54). The functions  $\operatorname{sn} z, \operatorname{cn} z, \operatorname{dn} z$  depend on  $k$  as required.

For further theory of elliptic functions and integrals see Refs. 4, 5, 7, 10, 12.

### 9. FUNCTIONS DEFINED BY LINEAR DIFFERENTIAL EQUATIONS

Attention will be restricted to the equations of second order, although most of the results can be generalized to equations of higher order. From the theory of real differential equations one knows that the homogeneous linear equation

$$(65) \quad a_0(x)y'' + a_1(x)y' + a_2(x)y = 0$$

has, in general, two linearly independent solutions. When the coefficients  $a_j(x)$  are, for example, polynomials, the solutions can be obtained as power series (see Chap. 5, Sect. 7). Such power series are best studied from the point of view of complex variables. Accordingly, one drops the restriction to real variables and considers a differential equation

$$(66) \quad p_0(z) \frac{d^2 w}{dz^2} + p_1(z) \frac{dw}{dz} + p_2(z)w = 0 \quad (p_0(z) \neq 0).$$

The coefficients  $p_j(z)$  are assumed to be analytic in an open region  $D$  and the solutions  $w(z)$  sought are to be analytic at least in a portion of  $D$ . By dividing by  $p_0(z)$  one obtains an equation

$$(67) \quad \frac{d^2 w}{dz^2} + q_1(z) \frac{dw}{dz} + q_2(z)w = 0,$$

where the  $q_j(z)$  are analytic except for poles in  $D$ .

**Series Solution at Ordinary Points.** A point  $z_0$  of  $D$  at which all  $q_j(z)$  are analytic is called an *ordinary point* of eq. (67). If the  $q_j(z)$  are expanded in power series about an ordinary point  $z_0$  and  $w(z)$  is written as a power series with undetermined coefficients, substitution in eq. (67) yields a formal solution for  $w(z)$ . In this solution  $c_0 = w(z_0)$  and  $c_1 = w'(z_0)$  are arbitrary, and the later coefficients are expressible in terms of  $c_0, c_1$  by recursion formulas; one can write  $w(z) = c_0 w_1(z) + c_1 w_2(z)$ . One can prove that the general formal solution converges in a neighborhood of  $z_0$  and represents an analytic solution of eq. (67). Furthermore, the general solution can be continued analytically throughout  $D$ , minus the poles of  $q_1(z), q_2(z)$ , and it remains a solution under such continuation.

**Singular Points.** A point  $z_0$  at which  $q_1$  or  $q_2$  has an isolated singularity is called a *singular point* of eq. (67). One can study the solutions near a singular point  $z_0$  by selecting a nearby ordinary point  $z_1$ . If there are no singularities other than  $z_0$  in the circle  $|z - z_0| < 2|z_0 - z_1|$ , then each series solution about  $z_1$  will converge for  $|z - z_1| < |z_0 - z_1|$  and can be continued analytically in the ring domain  $0 < |z - z_0| < 2|z_0 - z_1|$ , as far as desired; the resulting function is then multiple-valued. One can also seek series solutions of form

$$(68) \quad (z - z_0)^\mu \sum_{n=-\infty}^{\infty} b_n (z - z_0)^n,$$

where  $\mu$  is allowed to be an arbitrary complex number; such a solution exhibits explicitly the multiple-valued behavior near  $z_0$ . One can show that a solution of form (68) does exist. A second linearly independent

solution can also be found, either of form (68) with a different choice of  $\mu$ , or of form

$$(69) \quad (z - z_0)^\mu \left\{ \log(z - z_0) \sum_{-\infty}^{\infty} c_n (z - z_0)^n + \sum_{-\infty}^{\infty} d_n (z - z_0)^n \right\}.$$

**Regular Singular Points.** If the Laurent series which appear in the two solutions have only a finite number of negative powers, the singular point  $z_0$  is termed *regular*; in this case (68) can be written in the form

$$(z - z_0)^\alpha \left[ 1 + \sum_{n=0}^{\infty} b_n (z - z_0)^n \right].$$

It can be shown that  $z_0$  is a regular singular point precisely when  $(z - z_0)q_1(z)$  and  $(z - z_0)^2q_2(z)$  are analytic at  $z_0$ . One can extend the definition to the case  $z_0 = \infty$ ; this is an ordinary point or regular singular point according as the substitution  $\zeta = 1/z$  leads to an equation in  $\zeta$  having an ordinary point or regular singular point at  $\zeta = 0$ .

If every value of  $z$ , including  $\infty$ , is either an ordinary point or a regular singular point, then the differential equation (67) is said to be of *Fuchsian type*. Since the regular singular points are isolated, there can be only a finite number of singular points in all, and  $q_1(z)$ ,  $q_2(z)$  must be rational functions of  $z$ .

**Hypergeometric Equation.** Let eq. (67) now be an equation of Fuchsian type. If there are at most two singular points, it can be verified that the solutions are elementary functions. If there are three singular points, then by a *linear fractional transformation*  $z' = (az + b)/(cz + d)$  these can be placed at 0, 1,  $\infty$  and by a substitution  $w = z^\lambda(1 - z)^\mu w'$ , the equation can be reduced to the form

$$(70) \quad z(1 - z) \frac{d^2w}{dz^2} + \{c - (a + b + 1)z\} \frac{dw}{dz} - abw = 0,$$

known as the *hypergeometric equation*. If  $c$  is not 0 or a negative integer, one solution valid for  $|z| < 1$  is the *hypergeometric series*

$$(71) \quad F(a, b, c; z) = 1 + \frac{a \cdot b}{1 \cdot c} z + \frac{a(a + 1)}{1 \cdot 2} \cdot \frac{b(b + 1)}{c(c + 1)} z^2 + \dots;$$

a second solution is found (for  $c$  not an integer) to be

$$(72) \quad z^{1-c} F(a - c + 1, b - c + 1, 2 - c; z).$$

**The Legendre Equation.**

$$(73) \quad (1 - z^2) \frac{d^2w}{dz^2} - 2z \frac{dw}{dz} + n(n + 1)w = 0.$$

This has regular singular points at  $\pm 1$  and  $\infty$ , and is reducible to the hypergeometric equation. Two linearly independent solutions are found to be

$$(74) \quad P_n(z) = F(n+1, -n, 1; \frac{1}{2} - \frac{1}{2}z)$$

$$Q_n(z) = \frac{\pi^{1/2}}{2^{n+1}} \frac{\Gamma(n+1)}{\Gamma(n+\frac{3}{2})} \frac{1}{z^{n+1}} F\left(\frac{n+1}{2}, \frac{n+2}{2}, \frac{2n+3}{2}; \frac{1}{z^2}\right).$$

$P_n(z)$  is known as the *Legendre function* of degree  $n$  of the first kind; when  $n$  is a positive integer or 0,  $P_n$  reduces to a polynomial, the *Legendre polynomial* of degree  $n$ ;  $Q_n(z)$  is called the *Legendre function of second kind*.

### The Lamé Equation.

$$(75) \quad \frac{d^2w}{dz^2} + \frac{1}{2} \left( \frac{1}{z-a} + \frac{1}{z-b} + \frac{1}{z-c} \right) \frac{dw}{dz} - \frac{h + n(n+1)z}{4(z-a)(z-b)(z-c)} w = 0.$$

This has regular singular points at  $a, b, c, \infty$ . For details concerning its solutions, in particular, the Lamé functions, one is referred to Chap. 23 of Ref. 12.

A number of equations possessing irregular singular points are of importance in applications:

### The Bessel Equation.

$$(76) \quad z^2 \frac{d^2w}{dz^2} + z \frac{dw}{dz} + (z^2 - n^2)w = 0.$$

This has a regular point at 0, an irregular point at  $\infty$ . One solution is the *Bessel function* of order  $n$ :

$$(77) \quad J_n(z) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{\Gamma(k+1)} \frac{1}{\Gamma(k+n+1)} \left(\frac{z}{2}\right)^{n+2k}.$$

The functions  $J_n(z)$  and  $J_{-n}(z)$  provide two linearly independent solutions, unless  $n$  is an integer. If  $n$  is a positive integer, two linearly independent solutions are  $J_n(z)$  and the *Hankel function*

$$(78) \quad Y_n(z) = \sum_{k=0}^{\infty} \frac{(-1)^k (z/2)^{n+2k}}{k!(n+k)!} \left\{ 2 \log \frac{z}{2} + 2\gamma - \sum_{m=k+1}^{n+k} \frac{1}{m} \right\} - \sum_{k=0}^{n-1} \frac{(n-k-1)!}{k!} \left(\frac{z}{2}\right)^{-n+2k},$$

where  $\gamma = 0.5722 \dots$  is Euler's constant (Sect. 10). When  $n$  is not an integer,  $Y_n(z)$  can be defined more simply:

$$(79) \quad Y_n(z) = 2\pi e^{n\pi i} \frac{J_n(z) \cos n\pi - J_{-n}(z)}{\sin 2n\pi}.$$

**The Mathieu Equation.**

$$(80) \quad \frac{d^2w}{dz^2} + (a + 16q \cos 2z)w = 0.$$

Here  $a$  and  $q$  are constants and there is one irregular singular point at  $\infty$ . The solutions are hence entire functions. For certain choices of  $a$  and  $q$ , eq. (80) has solutions which are periodic and either even or odd; these solutions are called *Mathieu functions*. For details on these functions one is referred to Chap. 19, Ref. 12.

**The Confluent Hypergeometric Equation.**

$$(81) \quad \frac{d^2w}{dz^2} + \left\{ -\frac{1}{4} + \frac{k}{z} + \frac{\frac{1}{4} - m^2}{z^2} \right\} w = 0.$$

This has a regular singular point at 0 and an irregular singular point at  $\infty$ . One solution is given by Whittaker's function

$$(82) \quad W_{k,m}(z) = -\frac{1}{2\pi i} \Gamma(k + \frac{1}{2} - m) e^{-\frac{1}{2}z} z^k \int_C (-t)^{-k-\frac{1}{2}+m} \left(1 + \frac{t}{z}\right)^{k-\frac{1}{2}+m} e^{-t} dt,$$

where  $C$  is a path from  $+\infty$  to  $\epsilon$  along the "upper edge" of the positive real axis, then around the circle  $|t| = \epsilon$  in the positive direction, then from  $t = \epsilon$  to  $+\infty$  along the lower edge of the positive real axis. The integrand is single-valued if one chooses  $\arg(-t)$  to vary from  $-\pi$  to  $+\pi$  on the path, if  $t = -z$  is outside  $C$ , and  $\arg(1 + t/z)$  is chosen as that branch which  $\rightarrow 0$  as  $t \rightarrow 0$  inside  $C$ . The definition fails when  $k + \frac{1}{2} - m$  is a negative integer, but can be modified to cover this case (see Chap. 16 of Ref. 12).

**10. OTHER TRANSCENDENTAL FUNCTIONS**

Other analytic functions of importance in applications are the following:

**The Gamma Function.**

$$(83) \quad \Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt.$$

This definition holds for  $\operatorname{Re}(z) > 0$ , but  $\Gamma$  can be continued analytically and becomes a meromorphic function with poles of order 1 at  $0, -1, -2, \dots$ . Identities satisfied by  $\Gamma(z)$  are the following:

$$(84) \quad \Gamma(z+1) = z\Gamma(z);$$

$$(85) \quad \Gamma(n+1) = n! \quad (n = 1, 2, 3, \dots);$$

$$(86) \quad \Gamma(z)\Gamma(-z) = -\frac{\pi}{z \sin \pi z};$$

$$(87) \quad \Gamma(z) = \lim_{n \rightarrow \infty} \frac{n! n^z}{z(z+1) \cdots (z+n)};$$

$$(88) \quad \frac{1}{\Gamma(z)} = ze^{\gamma z} \prod_{n=1}^{\infty} \left[ \left(1 + \frac{z}{n}\right) e^{-z/n} \right],$$

where

$$(89) \quad \gamma = \lim_{m \rightarrow \infty} \left( \sum_{n=1}^m \frac{1}{n} - \log m \right) = 0.5772 \, 1566 \, 49 \dots$$

is the Euler-Mascheroni constant.

### The Beta Function.

$$(90) \quad B(z, w) = \int_0^1 t^{z-1} (1-t)^{w-1} dt, \quad \operatorname{Re} z > 0, \operatorname{Re} w > 0.$$

This is expressible in terms of the  $\Gamma$ -function:

$$(91) \quad B(z, w) = \frac{\Gamma(z)\Gamma(w)}{\Gamma(z+w)}.$$

### The Incomplete Gamma Function.

$$(92) \quad \gamma(\alpha, z) = \int_0^z e^{-t} t^{\alpha-1} dt, \quad \operatorname{Re} \alpha > 0.$$

This is expressible in terms of the Whittaker function of the preceding section:

$$(93) \quad \gamma(\alpha, z) = \Gamma(\alpha) - z^{\frac{1}{2}(n-1)} e^{-\frac{1}{2}z} W_{\frac{1}{2}(n-1), \frac{1}{2}n}(z).$$

### The Error Functions.

$$(94) \quad \operatorname{Erf}(z) = \int_0^z e^{-t^2} dt;$$

$$(95) \quad \operatorname{Erfc}(z) = \int_z^{\infty} e^{-t^2} dt = \frac{\pi}{2} - \operatorname{Erf}(z).$$

These functions are also expressible in terms of the Whittaker function:

$$(96) \quad \operatorname{Erfc}(z) = \frac{1}{2} z^{-1/2} e^{-1/2 z^2} W_{-1/4, 1/4}(z^2).$$

**The Logarithmic Integral Function.**

$$(97) \quad \operatorname{li}(z) = \int_0^z \frac{dt}{\log t} = -(-\log z)^{1/2} z^{1/2} W_{-1/2, 0}(-\log z).$$

**The Exponential Integral Function.**

$$(98) \quad \operatorname{Ei}(z) = -\int_{-z}^{\infty} \frac{e^{-t}}{t} dt = \operatorname{li}(e^z).$$

**The Sine and Cosine Integral Functions.**

$$(99) \quad \operatorname{si} z = \int_{\infty}^z \frac{\sin t}{t} dt = \frac{1}{2i} [\operatorname{Ei}(iz) - \operatorname{Ei}(-iz)],$$

$$(100) \quad \operatorname{Si} z = \int_0^z \frac{\sin t}{t} dt = \frac{\pi}{2} + \operatorname{si} z;$$

$$(101) \quad \operatorname{Ci} z = -\int_z^{\infty} \frac{\cos t}{t} dt = \frac{1}{2} [\operatorname{Ei}(iz) + \operatorname{Ei}(-iz)].$$

In eq. (97)  $z$  is first taken as real and positive, but analytic continuation then gives meaning to the function, as a multiple-valued function, for all  $z \neq 0$ . Similarly, in eq. (98),  $z$  is first to be real and negative. The functions  $\operatorname{si} z$  and  $\operatorname{Si} z$  are entire functions;  $\operatorname{Ci}(z) - \log z$  is also entire.

**The Riemann Zeta Function.**

$$(102) \quad \zeta(z) = \sum_{n=1}^{\infty} \frac{1}{n^z}, \quad \operatorname{Re} z > 1.$$

This function can be continued analytically and becomes a function single-valued and analytic for all  $z$  except  $z = 1$ , where  $\zeta(z)$  has a pole of first order. One has the integral representation:

$$(103) \quad \zeta(z) = \frac{1}{\Gamma(z)} \int_0^{\infty} \frac{t^{z-1}}{e^t - 1} dz, \quad \operatorname{Re} z > 1.$$

REFERENCES

1. *Higher Transcendental Functions*, Vols. 1, 2, 3, prepared by the staff of the Bateman manuscript project, McGraw-Hill, New York, 1954.
2. L. V. Ahlfors, *Complex Analysis*, McGraw-Hill, New York, 1953.
3. R. V. Churchill, *Introduction to Complex Variables and Applications*, McGraw-Hill, New York, 1948.
4. H. Hancock, *Lectures on the Theory of Elliptic Functions*, Vol. 1, Wiley, New York, 1910.
5. H. Hancock, *Elliptic Integrals*, Wiley, New York, 1917.
6. A. Hurwitz and R. Courant, *Funktionentheorie*, 3rd edition, Springer, Berlin, 1929.
7. E. Jahnke and F. Emde, *Tables of Functions*, 3rd edition, Teubner, Berlin, 1938.
8. W. Kaplan, *A First Course in Functions of a Complex Variable*, Addison-Wesley, Cambridge, Mass., 1953.
9. K. Knopp, *Theory of Functions*, Vols. 1, 2, translated by F. Bagemihl, Dover, New York, 1945.
10. F. Oberhettinger and W. Magnus, *Anwendung der Elliptischen Funktionen in Physik und Technik*, Springer, Berlin, 1949.
11. E. C. Titchmarsh, *The Theory of Functions*, 2nd edition, Oxford University Press, Oxford, England, 1939.
12. E. T. Whittaker and G. M. Watson, *A Course of Modern Analysis*, 4th edition, Cambridge University Press, Cambridge, England, 1940.

# Operational Mathematics

W. Kaplan

1. Heaviside Operators	8-01
2. Application to Differential Equations	8-05
3. Superposition Principle. Response to Unit Function and Delta Function	8-06
4. Appraisal of the Heaviside Calculus	8-07
5. Operational Calculus Based on Integral Transforms	8-07
6. Fourier Series. Finite Fourier Transform	8-10
7. Fourier Integral. Fourier Transforms	8-15
8. Laplace Transforms	8-17
9. Other Transforms	8-18
References	8-19

REMARK. The purpose of this chapter is to provide a survey of operational methods. The method which has proved most useful is that of the Laplace transform, which is mentioned only briefly in this chapter and is discussed in full in Chap. 9.

## 1. HEAVISIDE OPERATORS

**Definitions.** For functions of a real variable  $t$  one writes:

$$(1) \quad D \equiv \frac{d}{dt}, \quad D^2 \equiv \frac{d^2}{dt^2}, \quad \dots, \quad D^n \equiv \frac{d^n}{dt^n}$$

and defines a *polynomial differential operator*

$$\phi(D) = a_0 D^n + a_1 D^{n-1} + \dots + a_{n-1} D + a_n$$

by the requirement that for every  $f(t)$

$$(2) \quad \phi(D)f = (a_0D^n + \cdots + a_n)f = a_0 \frac{d^n f}{dt^n} + \cdots + a_{n-1} \frac{df}{dt} + a_n f,$$

wherever  $f$  has derivatives through the  $n$ th order. The coefficients  $a_0$  ( $\neq 0$ ),  $a_1, \dots, a_n$  may depend on  $t$ , but for most applications they are constants. The integer  $n$  is the *order* of the operator.

The *sum* and *product* of polynomial operators are defined by the rules:

$$(3) \quad [\phi_1(D) + \phi_2(D)]f = \phi_1(D)f + \phi_2(D)f$$

$$(4) \quad [\phi_1(D) \cdot \phi_2(D)]f = \phi_1(D)[\phi_2(D)f].$$

It immediately follows that addition is the same as for ordinary polynomials; the same is true of multiplication *provided the coefficients are constants*. For example,

$$(D + 1)(D - 1) = D^2 - 1,$$

$$(D + t)(D - t) = D^2 - (t^2 + 1).$$

In general, the polynomial operator is *linear*:

$$(5) \quad \phi(D)(c_1f_1 + c_2f_2) = c_1\phi(f_1) + c_2\phi(f_2),$$

provided  $c_1$  and  $c_2$  are constants.

The *reciprocal* or *inverse* of a polynomial operator is defined by the condition:

$$(6) \quad g(t) = \frac{1}{\phi(D)}f(t) \quad \text{if} \quad \phi(D)g(t) = f(t)$$

and  $g(0) = 0$ ,  $g'(0) = 0$ ,  $\dots$ ,  $g^{(n-1)}(0) = 0$ , where  $n$  is the order of  $\phi$ . Thus  $[1/\phi(D)]f$  is the solution of the differential equation

$$(7) \quad a_0 \frac{d^n x}{dt^n} + \cdots + a_{n-1} \frac{dx}{dt} + a_n x = f(t)$$

satisfying the initial conditions  $x = 0$ ,  $dx/dt = 0$ ,  $\dots$ ,  $d^{n-1}x/dt^{n-1} = 0$  for  $t = 0$ . It is assumed the differential equation is such that there is a unique solution satisfying the initial conditions; this is surely so if  $a_0, \dots, a_n$  are constants and  $f(t)$  is continuous for all  $t$ . One defines the sums and products of reciprocal operators by the equations:

$$(8) \quad \left[ \frac{1}{\phi_1(D)} + \frac{1}{\phi_2(D)} \right] f = \frac{1}{\phi_1} f + \frac{1}{\phi_2} f,$$

$$(9) \quad \left[ \frac{1}{\phi_1(D)} \cdot \frac{1}{\phi_2(D)} \right] f = \frac{1}{\phi_1(D)} \left[ \frac{1}{\phi_2(D)} f \right].$$

Addition is commutative:  $(1/\phi_1) + (1/\phi_2) = (1/\phi_2) + (1/\phi_1)$ , as is multiplication, provided the coefficients are constant.

The *ratio* of two polynomial operators is defined by the equation:

$$(10) \quad \frac{\phi_1(D)}{\phi_2(D)} f = \phi_1(D) \left[ \frac{1}{\phi_2(D)} f \right].$$

Here the order chosen is essential: it is not true that

$$(11) \quad \phi_1(D) \left[ \frac{1}{\phi_2(D)} f \right] = \frac{1}{\phi_2(D)} [\phi_1(D)f],$$

even if the coefficients are constant. All operators defined thus far are linear.

**Integral Representation of Inverse Operators with Constant Coefficients.** One has the formulas:

$$(12) \quad \frac{1}{D} f(t) = \int_0^t f(u) du,$$

$$(13) \quad \frac{1}{D - a} f(t) = e^{at} \int_0^t e^{-au} f(u) du,$$

$$(14) \quad \frac{1}{(D - a)^k} f(t) = e^{at} \int_0^t \frac{e^{-au} (t - u)^{k-1}}{(k - 1)!} f(u) du,$$

where  $a$  is constant and  $k = 1, 2, \dots$ . Now  $\phi(D)$  can be factored as in algebra:

$$(15) \quad \phi(D) = a_0(D - r_1)(D - r_2) \cdots (D - r_n)$$

where  $r_1, \dots, r_n$  are the roots of the *characteristic equation*

$$(16) \quad \phi(r) = a_0 r^n + \cdots + a_{n-1} r + a_n = 0.$$

Correspondingly,

$$(17) \quad \frac{1}{\phi(D)} f = \frac{1}{a_0(D - r_1) \cdots (D - r_n)} f.$$

Thus if  $n = 2$

$$\frac{1}{\phi(D)} f = \frac{1}{a_0(D - r_1)} \left[ \frac{1}{(D - r_2)} f \right] = \frac{e^{r_1 t}}{a_0} \int_0^t e^{(-r_1 + r_2)u} \int_0^u e^{-r_2 v} f(v) dv du.$$

In general, computation of  $[1/\phi(D)]f$  is reduced to a repeated integration.

If  $\phi(r)$  has complex roots, quadratic factors appear in eq. (17); for these one has the rule:

$$(18) \quad \frac{1}{(D-a)^2 + b^2} f = \frac{e^{at}}{b} \int_0^t e^{-au} \sin b(t-u) f(u) du.$$

If  $1/\phi(D)$  is expanded in partial fractions as in algebra, then the corresponding operator identity is valid; for example,

$$\begin{aligned} \frac{1}{D^2-1} f &= \left( \frac{1}{2} \frac{1}{D-1} - \frac{1}{2} \frac{1}{D+1} \right) f \\ &= \frac{1}{2} e^t \int_0^t e^{-u} f(u) du - \frac{1}{2} e^{-t} \int_0^t e^u f(u) du. \end{aligned}$$

**HEAVISIDE EXPANSION THEOREM.** More generally, a ratio  $\phi_1(D)/\phi_2(D)$  can be replaced by its partial fraction expansion. If in particular the degree of  $\phi_2$  exceeds that of  $\phi_1$ , and  $\phi_2(r)$  has simple roots  $r_1, r_2, \dots, r_n$ , then by Chap. 7, Sect. 5,

$$(19) \quad \begin{aligned} \frac{\phi_1(r)}{\phi_2(r)} &= \sum_{k=1}^n \frac{\phi_1(r_k)}{\phi_2'(r_k)} \frac{1}{r-r_k}, \\ \frac{\phi_1(D)}{\phi_2(D)} &= \sum_{k=1}^n \frac{\phi_1(r_k)}{\phi_2'(r_k)} \frac{1}{D-r_k}. \end{aligned}$$

This is in essence the Heaviside expansion theorem.

**Power Series Operators.** The formal relation

$$\frac{1}{D-a} = \frac{1}{-a \left( 1 - \frac{D}{a} \right)} = -\frac{1}{a} - \frac{D}{a^2} - \frac{D^2}{a^3} - \dots$$

does not agree with the definition of  $1/(D-a)$ . However, if the operator is applied to a polynomial in  $t$ , one obtains a particular solution of the corresponding differential equation (with modified initial conditions). For example,

$$\left( -\frac{1}{a} - \frac{D}{a^2} - \frac{D^2}{a^3} - \dots \right) t^2 = -\frac{t^2}{a} - \frac{2t}{a^2} - \frac{2}{a^3}$$

is a solution of

$$\frac{dx}{dt} - ax = t^2$$

for which  $x(0) = -2a^{-3}$ .

One can also expand in inverse powers of  $D$ :

$$(20) \quad \frac{1}{D - a} = \frac{1}{D} + \frac{a}{D^2} + \frac{a^2}{D^3} - \dots$$

In this case the rule can be proved to be correct.

The power series  $\sum_0^{\infty} h^n D^n / n!$  can be interpreted as the operator  $e^{hD}$ . One then finds, under appropriate conditions,

$$(21) \quad e^{hD} f = \sum_0^{\infty} \frac{f^{(n)}(t)}{n!} h^n = f(t + h).$$

**2. APPLICATION TO DIFFERENTIAL EQUATIONS**

The general solution of a linear differential equation,

$$(22) \quad \phi(D)x = f(t), \quad \phi(D) = a_0 D^n + \dots + a_n,$$

is formed of the *complementary function*  $x_c(t)$ , which is the general solution of the homogeneous equation  $\phi(D)x = 0$  and of a *particular solution*  $x_p(t)$  of the given equation:

$$(23) \quad x = x_c(t) + x_p(t)$$

[cf. Chap. 5, Sect. 3]. The Heaviside operators provide simple ways of finding  $x_p(t)$ , namely as the function

$$(24) \quad x_p(t) = \frac{1}{\phi(D)} f(t);$$

this is the solution with zero initial conditions. If  $1/\phi(D)$  is expanded in partial fractions, one can then apply the integral formulas (12), (13), (14), (18).

The procedure can be extended to simultaneous equations. For *example*,

$$\begin{aligned} Dx + (D - 1)y &= F(t) \\ (D + 1)x + 2Dy &= G(t) \end{aligned}$$

can be solved formally:

$$x = \frac{2D}{D^2 + 1} F(t) - \frac{D - 1}{D^2 + 1} G(t), \quad y = \frac{D}{D^2 + 1} G(t) - \frac{D + 1}{D^2 + 1} F(t)$$

and it can be verified that these provide the solution for which  $x = 0$ ,  $y = 0$  when  $t = 0$ .

### 3. SUPERPOSITION PRINCIPLE. RESPONSE TO UNIT FUNCTION AND DELTA FUNCTION

The Heaviside *unit function*  $u(t)$  is defined to equal 0 for  $t < 0$  and to equal 1 for  $t \geq 0$ . The solution of the differential equation  $\phi(D)x = u(t)$  with zero initial values, i.e., the function  $(1/\phi(D))u(t) = A(t)$  is known as the *indicial admittance* or *step response*.

The *superposition principle* states that the response of a linear system to a linear combination  $c_1f_1(t) + c_2f_2(t)$  equals the corresponding linear combination  $c_1x_1(t) + c_2x_2(t)$  of the responses  $x_1(t)$  to  $f_1(t)$ ,  $x_2(t)$  to  $f_2(t)$ . In the typical case  $x(t)$  and  $f(t)$  are related by a differential equation  $\phi(D)x = f(t)$  and the superposition principle is equivalent to the statement that  $1/\phi(D)$  is a *linear operator*.

One can apply the superposition principle to show that (when  $\phi(D)$  has constant coefficients) the response to a general  $f(t)$  is deducible from the indicial admittance, i.e., the response to  $u(t)$ . Indeed, the response to  $u(t-h)$ , for  $h \geq 0$ , is  $A(t-h)$ ; one can approximate  $f(t)$  by a linear combination  $\sum_k c_k u(t-t_k)$ , where  $c_k = f(t_{k+1}) - f(t_k)$ . A passage to the limit gives the *Duhamel theorem*

$$(25) \quad x(t) = \int_0^t f(s)A'(t-s) ds.$$

[It is assumed that  $f(t)$  is 0 for  $t < 0$  and the solution  $x(t)$  has 0 initial values]. If  $f(t)$  is constant, equal to  $1/\epsilon$  for  $0 \leq t \leq \epsilon$ , and then equal to 0 for  $t > \epsilon$ , the response is  $[A(t) - A(t-\epsilon)]/\epsilon$ . The limiting case of such an  $f(t)$ , as  $\epsilon \rightarrow 0$ , is an "ideal function," the *delta function*  $\delta(t)$ , also termed the *unit impulse function*. The response to  $\delta(t)$  is interpreted as  $A'(t) = h(t)$ . Accordingly,

$$(26) \quad x(t) = \int_0^t f(s)h(t-s) ds.$$

For some linear systems the response to  $u(t)$  appears as  $[\phi(D)/\psi(D)]u$ , where  $\phi$  and  $\psi$  are polynomials. If  $\psi$  has simple roots  $b_\alpha$  ( $\alpha = 1, \dots, k$ ), then by eqs. (19)

$$\frac{\phi(D)}{\psi(D)} u = \sum_{\alpha=1}^k \frac{\phi(b_\alpha)}{\psi'(b_\alpha)} \frac{1}{D - b_\alpha} u = \sum_{\alpha=1}^k \frac{\phi(b_\alpha)}{\psi'(b_\alpha)} \frac{e^{b_\alpha t} - 1}{b_\alpha} u(t)$$

and hence

$$(27) \quad \frac{\phi(D)}{\psi(D)} u(t) = \frac{\phi(0)}{\psi(0)} u(t) + \sum_{\alpha=1}^k \frac{\phi(b_\alpha)}{\psi'(b_\alpha)} \frac{e^{b_\alpha t}}{b_\alpha} u(t).$$

## 4. APPRAISAL OF THE HEAVISIDE CALCULUS

The operational methods described in the preceding section provide a valuable tool for solution of linear differential equations. The method has two principal drawbacks: it is very awkward to obtain solutions with specified initial values other than 0; further development of the method leads to symbolic expressions whose meaning has to be studied afresh in each case. Great ingenuity has been employed to remedy these defects but a satisfactory general theory within the Heaviside framework has not been found.

On the other hand, it has been discovered that all the goals of the Heaviside calculus can be achieved without reference to differential operators or their inverses and, indeed, without any symbolic calculus. The means to this end is the *Laplace transform* (see Chap. 9); the closely related *Fourier transforms* can also serve the purpose. By means of these the questions about initial conditions are easily disposed of, and justification of formal rules becomes simple.

The transformations referred to do not merely serve as a substitute for the Heaviside calculus. Deeper study shows that they lie at the very basis of that calculus and must inevitably enter in a full justification of the operational rules.

## 5. OPERATIONAL CALCULUS BASED ON INTEGRAL TRANSFORMS

One considers equations of form

$$(28) \quad F(y) = \int_a^b f(t)K(t, y) dt.$$

Such an equation assigns a function  $F(y)$  to each function  $f(t)$ , whenever the integral has meaning. One calls  $F$  the *integral transform* of  $f$  with respect to the particular transformation (28) and writes:

$$(29) \quad F = T[f].$$

The relation between  $f$  and  $F$  is much like that between independent and dependent variables; here the variables are *functions*.

Because of the form of eq. (28), the transformation  $T$  is *linear*:

$$(30) \quad T[c_1f_1 + c_2f_2] = c_1T[f_1] + c_2T[f_2].$$

The transformation (28) is said to have a (single-valued) *inverse* if, for each  $F$  of a certain class, there is precisely one  $f$  such that  $T[f] = F$ . One writes:

$$(31) \quad f = T^{-1}[F]$$

and calls  $f$  the *inverse transform* of  $F$ . Because  $T$  is linear,  $T^{-1}$  must also be linear.

**Convolution.** If to each pair of functions  $f_1, f_2$  one can associate (in a unique manner) a third function  $f_3$  such that

$$(32) \quad T[f_3] = T[f_1] \cdot T[f_2],$$

then one calls  $f_3$  the *convolution* of  $f_1, f_2$  and writes:

$$(33) \quad f_3 = f_1 * f_2.$$

The convolution must then obey simple laws:

$$(34) \quad \begin{aligned} f_1 * f_2 &= f_2 * f_1; f_1 * (f_2 + f_3) = f_1 * f_2 + f_1 * f_3; \\ f_1 * cf_2 &= cf_1 * f_2 = c(f_1 * f_2); f_1 * (f_2 * f_3) = (f_1 * f_2) * f_3. \end{aligned}$$

**Solution of Differential Equations.** Suppose the transformation  $T$  has the property that, for a certain polynomial differential operator  $\phi(D)$  and for  $f(t)$  in a certain class of functions, one has an identity:

$$(35) \quad T[\phi(D)f] = H(y)T[f] = H(y)F(y),$$

where  $H(y)$  is a function of  $y$  associated with the operator  $\phi$ . Then to solve a differential equation

$$(36) \quad \phi(D)x = g(t)$$

for  $x = f(t)$  in the class referred to, one forms the transformed equation

$$T[\phi(D)x] = T[g]$$

or equivalently, by eqs. (35),

$$(37) \quad H(y)F(y) = G(y).$$

Accordingly,

$$(38) \quad F(y) = \frac{G(y)}{H(y)}, \quad f(t) = T^{-1} \left[ \frac{G(y)}{H(y)} \right].$$

One can try to find the inverse transform of  $G(y)/H(y)$  with the aid of *tables* of functions and their transforms. One can also seek

$$(39) \quad T^{-1} \left[ \frac{1}{H(y)} \right] = w(t).$$

Then eq. (38) gives

$$(40) \quad T[f(t)] = T[w(t)] \cdot T[g(t)] = T[w * g],$$

so that

$$(41) \quad f(t) = w(t) * g(t).$$

The crucial question is choice of the transformation  $T$  so that eqs. (35) hold. For differential equations with constant coefficients it is sufficient to choose  $T$  so that

$$(42) \quad T[Df(t)] = H(y)F(y).$$

For then

$$(43) \quad T(a_0D^n + \cdots + a_{n-1}D + a_n)f = (a_0H^n + \cdots + a_1H + a_n)F(y).$$

**Fourier Integral.** Now associated with the operator  $D$  are certain functions  $f$  such that  $Df$  is a constant times  $f$ ; these are precisely the functions  $ke^{at}$ . It is known that an "arbitrary" function  $f$  is expressible as a "sum" of functions of this form. For example, under appropriate conditions,

$$(44) \quad f(t) = \int_{-\infty}^{\infty} F(\omega)e^{i\omega t} d\omega;$$

this is the representation of  $f$  as a Fourier integral. One finds that

$$(45) \quad F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt,$$

so that  $F(\omega)$  can be considered as  $T[f]$ , a linear integral transform of  $T$ ; except for a constant multiplier, this is the Fourier transform of  $f$ . The fact that  $De^{i\omega t} = ie^{i\omega t}$  is reflected in the formula

$$(46) \quad Df = f'(t) = \int_{-\infty}^{\infty} i\omega F(\omega)e^{i\omega t} d\omega,$$

which follows from eq. (44). Hence

$$(47) \quad T[Df] = i\omega F(\omega).$$

Thus the transformation  $T$  defined by eq. (45) has the property desired.

The functions  $f$  representable as Fourier integrals must be small for large positive or negative  $t$  (see Sect. 9). For functions not satisfying such a condition other representations can be used. If  $f$  is defined only for  $t \geq 0$  and does not grow too rapidly as  $t \rightarrow \infty$ , one can use the *Laplace transform*. If  $f$  is defined for all  $t$  and has period  $2\pi$ , then  $f$  can be represented by a Fourier series; associated with this series is the *finite Fourier transform*.

If  $\phi(D)$  does not have constant coefficients, the transformation  $T$  must be related specially to the particular operator  $\phi$ . Associated with  $\phi$  are the "characteristic functions"  $f$  for which  $\phi D(f)$  is a constant times  $f$ . Representation of an arbitrary function as a series or integral of such characteristic functions leads to a corresponding integral transformation.

## 6. FOURIER SERIES. FINITE FOURIER TRANSFORM

Let  $f(t)$  be defined for all real  $t$ . One says that  $f(t)$  has period  $T \neq 0$  if  $f(t + T) = f(t)$  for all  $t$ . A function  $f(t)$  given only for  $a < t < b$  can always be defined outside this interval so as to have period  $T = b - a$  (periodic extension of  $f(t)$ ).

Let  $f(t)$  have period  $T$  and let  $\omega = 2\pi/T$ . The *Fourier series* of  $f(t)$  is defined as the series:

$$(48) \quad \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos n\omega t + b_n \sin n\omega t),$$

where

$$(49) \quad a_n = \frac{2}{T} \int_0^T f(t) \cos n\omega t \, dt, \quad b_n = \frac{2}{T} \int_0^T f(t) \sin n\omega t \, dt.$$

Because of the periodicity of  $f(t)$ , the interval of integration in eqs. (49) can be replaced by any other interval of length  $T$ , e.g., from  $-\frac{1}{2}T$  to  $\frac{1}{2}T$ .

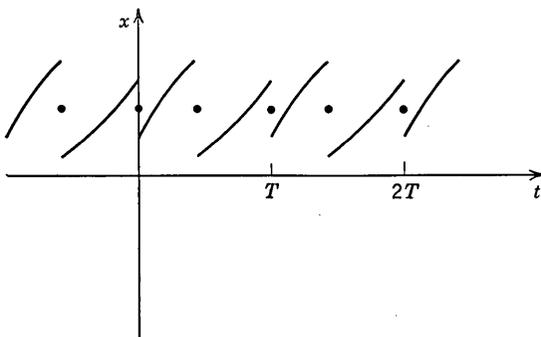


FIG. 1. Piecewise continuous function of period  $T$ .

It is assumed that the integrals in eqs. (49) have meaning. For this it is sufficient that  $f(t)$  be piecewise continuous, i.e., continuous except for jump discontinuities (Fig. 1).

**Convergence.** The Fourier series of  $f(t)$  converges to  $f(t)$  under very general conditions: for example, wherever  $f(t)$  is continuous and has a derivative to the left and to the right. At a jump discontinuity  $t_0$  the series converges to

$$\frac{1}{2}[f(t_0+) + f(t_0-)],$$

where

$$(50) \quad f(t_0+) = \lim_{t \rightarrow t_0+} f(t), \quad f(t_0-) = \lim_{t \rightarrow t_0-} f(t),$$

provided  $[f(t_0 + h) - f(t_0 +)]/h$  and  $[f(t_0 -) - f(t_0 - h)]/h$  have limits as  $h \rightarrow 0+$ . For example, if  $f(t) = t$  for  $-1 < t < 1$  and  $f(t)$  has period 2, then the corresponding Fourier series converges to 0 at  $t = 1, t = -1, t = 3, t = -3, \dots$ . It is common practice to redefine  $f(t)$  as  $\frac{1}{2}[f(t_0 +) + f(t_0 -)]$  at each jump discontinuity.

If  $f(t)$  is merely continuous, there is no general theorem on convergence. However, one has a "convergence in the mean," that is, if  $s_n(t)$  denotes the sum of the first  $n$  terms of the series (48), then the "mean square error"

$$\frac{1}{T} \int_0^T [f(t) - s_n(t)]^2 dt$$

tends to 0 as  $n \rightarrow \infty$ . This result holds considerably more generally, e.g., if  $f$  is merely piecewise continuous.

If  $f(t)$  has a continuous derivative over an interval  $t_0 \leq t \leq t_1$ , then the Fourier series of  $f(t)$  converges *uniformly* to  $f(t)$  over this interval; i.e.,

$$(51) \quad \max_{t_0 \leq t \leq t_1} |f(t) - s_n(t)| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

In general, if a series of form (48), i.e., a "trigonometric series," converges uniformly to  $f(t)$  for  $0 \leq t \leq T$ , then the series must be the Fourier series of  $f(t)$ .

A function is determined uniquely by its "Fourier coefficients"  $a_0, a_1, \dots, b_1, \dots$ ; that is, if  $f(t)$  and  $g(t)$  have the same Fourier coefficients, then  $f(t) = g(t)$  except perhaps at points of discontinuity.

**Fourier Cosine and Sine Series.** If  $f(t)$  is even [ $f(t) = f(-t)$ ], then all  $b_n$  are 0 and  $f(t)$  is represented by a *Fourier cosine series*; that is,

$$(52) \quad f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos n\omega t, \quad a_n = \frac{4}{T} \int_0^{T/2} f(t) \cos n\omega t dt,$$

provided the convergence conditions are satisfied. If  $f(t)$  is given merely between 0 and  $\frac{1}{2}T$ , eqs. (52) are still valid; for  $f(t)$  can be extended to all  $t$  to be even and have period  $T$ . Similar remarks apply to representation of an odd function [ $f(t) = -f(-t)$ ] by a *Fourier sine series*:

$$(53) \quad f(t) = \sum_{n=1}^{\infty} b_n \sin n\omega t, \quad b_n = \frac{4}{T} \int_0^{T/2} f(t) \sin n\omega t dt.$$

The identities:

$$(54) \quad \cos \alpha = \frac{1}{2}(e^{i\alpha} + e^{-i\alpha}), \quad \sin \alpha = \frac{1}{2i}(e^{i\alpha} - e^{-i\alpha}),$$

lead to a rewriting of the formulas (49) in complex form. Under conditions for convergence,

$$(55) \quad f(t) = \sum_{n=-\infty}^{\infty} c_n e^{in\omega t}, \quad c_n = \frac{1}{T} \int_0^T f(t) e^{-in\omega t} dt, \quad n = 0, \pm 1, \dots$$

**Finite Fourier Transform.** One can interpret the doubly infinite sequence of numbers

$$\int_0^T f(t) e^{-in\omega t} dt, \quad n = 0, \pm 1, \pm 2, \dots,$$

as a function of  $n$ ,  $\phi(n)$ , defined only when  $n$  is an integer. The equation

$$(56) \quad \phi(n) = \int_0^T f(t) e^{-in\omega t} dt$$

can then be regarded as a special case of the linear integral transformation (28); the variable  $y$  is replaced by  $n$  and is restricted to integer values. The notations:

$$(57) \quad \phi(n) = \Phi[f(t)] \quad \text{or} \quad \phi = \Phi[f]$$

will be used to denote the functional transformation  $\Phi$ , the finite Fourier transformation, which assigns the function  $\phi(n)$  to the function  $f(t)$ .  $\Phi$  is then defined at least for all  $f(t)$  which are piecewise continuous for  $0 \leq t \leq T$ .

As in Sect. 5,  $\Phi$  is *linear*:

$$(58) \quad \Phi[c_1 f_1 + c_2 f_2] = c_1 \Phi[f_1] + c_2 \Phi[f_2].$$

**Inverse Transform.** If  $\Phi[f] = \phi$ , then one writes:  $f = \Phi^{-1}[\phi]$ . The inverse transformation is then uniquely defined by the theorem stated above concerning functions having the same Fourier coefficients. It is a less simple matter to describe those functions  $\phi(n)$  for which  $\Phi^{-1}$  exists. One class of such functions  $\phi(n)$  consists of those for which the series  $\sum T^{-1} \phi(n) e^{in\omega t}$  converges uniformly for  $0 \leq t \leq T$ . The sum of the series is then a function  $f(t)$  which serves as  $\Phi^{-1}[\phi]$ ;

$$(59) \quad \Phi^{-1}[\phi] = \frac{1}{T} \sum_{n=-\infty}^{\infty} \phi(n) e^{in\omega t}.$$

**Convolution.** Given  $f_1(t)$ ,  $f_2(t)$  having period  $T$ , their *convolution* is defined as:

$$(60) \quad f_3(t) = \int_0^T f_1(s) f_2(t - s) ds;$$

one writes:  $f_3(t) = f_1(t) * f_2(t)$ . One can then prove the characteristic property:

$$(61) \quad \Phi[f_1 * f_2] = \Phi[f_1] \cdot \Phi[f_2].$$

**Transformation of Derivatives.** If  $f(t)$  has a continuous derivative for  $0 \leq t \leq T$ , then an integration by parts proves that

$$(62) \quad \Phi[f'(t)] = f(T) - f(0) + in\omega\Phi[f(t)].$$

This rule can be made the basis for application of the finite Fourier transform to *boundary value* problems. Interest will be concentrated here on the periodic case:  $f(T) = f(0)$ , for which the rule becomes

$$(63) \quad \Phi[f'(t)] = in\omega\Phi[f(t)].$$

Similarly, if  $f$  is periodic and has continuous derivatives through the  $k$ th order,

$$(64) \quad \Phi[f^{(k)}(t)] = (in\omega)^k\Phi[f(t)];$$

this relation remains true if  $f^{(k-1)}(t)$  is continuous and  $f^{(k)}$  is continuous except at a finite number of points at which left and right hand  $k$ th derivatives exist. From eq. (64) it follows that, for every polynomial operator  $\psi(D) = a_0D^n + \dots + a_{n-1}D + a_n$  with constant coefficients

$$(65) \quad \Phi\{\psi(D)[f(t)]\} = \psi(in\omega)\Phi[f(t)].$$

**Steady-State Solutions of Differential Equations.** Let  $f(t)$  be piecewise continuous and have period  $T$ . Let  $a_0, \dots, a_n$  be constants and let  $\psi(D) = a_0D^n + \dots + a_{n-1}D + a_n$ . It can then be shown that in general the differential equation

$$(66) \quad \psi(D)x = f(t)$$

has a solution  $x = X(t)$  having period  $T$ ;  $X(t)$  has continuous derivatives through the  $(n - 1)$ st order and an  $n$ th derivative which is continuous where  $F(t)$  is continuous. If  $\psi(p)$  has no root of the form  $in\omega$  for some  $n$ , there is precisely one such periodic solution; it will be assumed in the following that  $\psi(in\omega) \neq 0$  for every  $n$ . Applying the finite Fourier transformation to eq. (66), one finds by eq. (65)

$$(67) \quad \Phi[X] = \frac{\phi(n)}{\psi(in\omega)}, \quad \text{where } \phi(n) = \Phi[f].$$

This equation gives the Fourier coefficients of  $X(t)$ , from which one can write

$$(68) \quad X = \frac{1}{T} \sum_{n=-\infty}^{\infty} \frac{\phi(n)}{\psi(in\omega)} e^{in\omega t}.$$

One can attempt to reduce this to a simpler form by developing a table of finite Fourier transforms and inverse transforms. One can also apply the convolution formula to eq. (67):

$$(69) \quad X = g * f = \int_0^T f(s)g(t-s) ds,$$

where  $g = \Phi^{-1}[1/\psi(in\omega)]$ . To find  $g$ , decompose  $1/\psi(in\omega)$  into partial fractions and apply linearity. The problem is reduced to finding inverses of  $(in\omega - a)^{-k}$  ( $k = 1, 2, \dots$ ). One finds:

$$(70) \quad \begin{aligned} \Phi^{-1} \left[ \frac{1}{in\omega - a} \right] &= k_a e^{at}, \\ \Phi^{-1} \left[ \frac{1}{(in\omega - a)^2} \right] &= e^{at} [k_{at} + T k_a^2 e^{aT}], \\ \Phi^{-1} \left[ \frac{1}{(in\omega - a)^{m+1}} \right] &= \frac{1}{m!} \frac{\partial^m}{\partial a^m} (k_a e^{at}), \end{aligned}$$

where  $k_a = (1 - e^{aT})^{-1}$ . In particular, if  $\psi(p)$  has simple roots  $p_1, \dots, p_m$ , so that

$$(71) \quad \frac{1}{\psi(p)} = \sum_{j=1}^m \frac{A_j}{p - p_j},$$

one finds that

$$(72) \quad X(t) = \sum_{j=1}^m A_j H_j(t, p_j)$$

where

$$(73) \quad \begin{aligned} H_j(t, p) &= e^{pt} [Q_j(t, p) + k_p e^{pT} Q_j(T, p)], \\ Q_j(t, p) &= \int_0^t f(s) e^{-ps} ds, \quad 0 \leq t \leq T. \end{aligned}$$

The operators  $Q_j$  and  $H_j$  can be tabulated for various functions  $f(t)$  of interest, so that the corresponding periodic solutions can be found easily. For tables and illustrations of applications see Ref. 11.

## 7. FOURIER INTEGRAL. FOURIER TRANSFORMS

**Fourier Integral.** By allowing the period  $T$  to become infinite, one is led to the following integral analogue of the Fourier series expansion:

$$(74) \quad f(t) = \int_0^{\infty} [a(\omega) \cos \omega t + b(\omega) \sin \omega t] d\omega,$$

$$a(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(t) \cos \omega t dt, \quad b(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(t) \sin \omega t dt.$$

The "coefficients"  $a(\omega)$ ,  $b(\omega)$  exist if  $f(t)$  is, for example, piecewise continuous, and  $\int_{-\infty}^{\infty} |f(t)| dt$  exists. The representation of  $f(t)$  as a Fourier integral is then valid under the same conditions as for Fourier series, e.g., wherever  $f'(t)$  exists. Also, under the conditions described in Sect. 8, the integral equals  $\frac{1}{2}[f(t_0+) + f(t_0-)]$  at each jump discontinuity of  $f$ . One can write eqs. (74) in complex form:

$$(75) \quad f(t) = \int_{-\infty}^{\infty} A(\omega) e^{i\omega t} d\omega, \quad A(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt;$$

the first integral must, however, be treated as a *principal value*, i.e., as  $\lim_{b \rightarrow \infty} \int_{-b}^b \dots$  as  $b \rightarrow \infty$ .

Under conditions analogous to those for Fourier series one is led to representation of a function  $f(t)$  in the interval  $0 \leq t < \infty$  by a *Fourier cosine integral*  $\int_0^{\infty} a(\omega) \cos \omega t d\omega$ . It is customary to define the Fourier cosine transform of  $f(t)$  as

$$(76) \quad F_c(\omega) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(t) \cos \omega t dt$$

so that the Fourier cosine integral representation of  $f$  reads

$$(77) \quad f(t) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} F_c(\omega) \cos \omega t d\omega;$$

thus  $f$  is also the Fourier cosine transform of  $F_c$ . Similar formulas hold for the Fourier sine transform:

$$(78) \quad F_s(\omega) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(t) \sin \omega t dt, \quad f(t) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} F_s(\omega) \sin \omega t d\omega.$$

Similarly one defines the (exponential) Fourier transform of  $f$  as

$$(79) \quad F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt,$$

so that by eqs. (75)

$$(80) \quad f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\omega)e^{i\omega t} d\omega.$$

**Properties of the Fourier Transform.** For simplicity, the numerical factor is dropped and the Fourier transform is defined as

$$(81) \quad \Phi_{\infty}[f] = \int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt;$$

then  $\Phi_{\infty}[f]$  is a linear operator. If  $f$  has a continuous derivative  $f'(t)$  and  $f(t), f'(t)$  satisfy the conditions stated above, then

$$(82) \quad \Phi_{\infty}[f'] = i\omega\Phi_{\infty}[f].$$

A convolution is defined as follows:

$$(83) \quad f * g = \int_{-\infty}^{\infty} f(s)g(t-s) ds = h(t)$$

and one has the characteristic property:

$$(84) \quad \Phi_{\infty}[f * g] = \Phi_{\infty}[f]\Phi_{\infty}[g];$$

it is assumed here that  $f, g$  satisfy the conditions given above. An *inverse* operator is defined by the condition:  $\Phi_{\infty}^{-1}[F] = f$ , if  $\Phi_{\infty}[f] = F$ . The function  $f$  can be shown to be uniquely defined by its transform  $F$ .

The applications of the Fourier transform to differential equations parallel those for the finite Fourier transform, as described in Sect. 8 above; eq. (65) is replaced by

$$(85) \quad \Phi_{\infty}[\psi(D)f(t)] = \psi(i\omega)\Phi_{\infty}[f].$$

Application of the transform to the equation  $\psi(D)x = f(t)$  yields a solution in the form of a Fourier integral:

$$(86) \quad X(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{F(\omega)}{\psi(i\omega)} e^{i\omega t} d\omega, \quad F(\omega) = \Phi_{\infty}[f],$$

or as a convolution:

$$(87) \quad X(t) = f * g, \quad g = \Phi_{\infty}^{-1} \left[ \frac{1}{\psi(i\omega)} \right].$$

If  $f(t) = 0$  for  $t < 0$ , the same solution is obtainable by Laplace transforms; see Sect. 10 and Chap. 9 below.

References to tables of Fourier transforms are given at the end of this chapter (Refs. 1, 5, 6).

8. LAPLACE TRANSFORMS

The Laplace transform of  $f(t)$ ,  $t \geq 0$ , is defined as

$$(88) \quad F(s) = L[f] = \int_0^\infty f(t)e^{-st} dt.$$

It is convenient to allow  $s$  to be complex:  $s = \sigma + i\omega$ . Equation (88) then reads:

$$(89) \quad F(\sigma + i\omega) = L[f] = \int_0^\infty f(t)e^{-\sigma t}e^{-i\omega t} dt;$$

hence for each fixed  $\sigma$  the Laplace transform of  $f$  is the same as the Fourier transform of  $f(t)e^{-\sigma t}$ , where  $f$  is considered to be 0 for  $t < 0$ :

$$(90) \quad L[f] = \Phi_\infty[f(t)e^{-\sigma t}].$$

Accordingly, the Laplace transform is well defined if  $\sigma$  is chosen so that

$$(91) \quad \int_0^\infty |f(t)|e^{-\sigma t} dt$$

exists, and for such  $\sigma$  one can invert:

$$(92) \quad f(t)e^{-\sigma t} = \frac{1}{2\pi} \int_{-\infty}^\infty F(\sigma + i\omega)e^{i\omega t} d\omega,$$

$$f(t) = L^{-1}[F(s)] = \frac{1}{2\pi} \int_{-\infty}^\infty F(s)e^{st} d\omega, \quad t > 0;$$

in the last integral  $s = \sigma + i\omega$ ,  $\sigma$  has any value such that (91) exists, and the integral itself is a principal value. The integral can be interpreted as an integral in the complex  $s$ -plane along the line  $\sigma = \text{const.}$ ,  $\omega$  going from  $-\infty$  to  $+\infty$  (Fig. 2). Since  $ds = i d\omega$  on the path,

$$(93) \quad f(t) = \frac{1}{2\pi i} \int_C F(s)e^{st} ds,$$

$C$  being the line  $\sigma = \text{const.}$  The conditions for equality of left and right sides of (93) are the same as for Fourier integrals. At  $t = 0$ ,  $f(t)$  will

in general have a jump, because of the convention that  $f(t)$  be 0 for  $t < 0$ , and accordingly the right hand side gives  $\frac{1}{2}f(0+)$ .

The validity of eqs. (88) and (92) depends on choosing  $\sigma$  so that (91) exists. It can be shown that for each  $f(t)$  there is a value  $\sigma_0$ ,  $-\infty \leq \sigma_0 \leq$

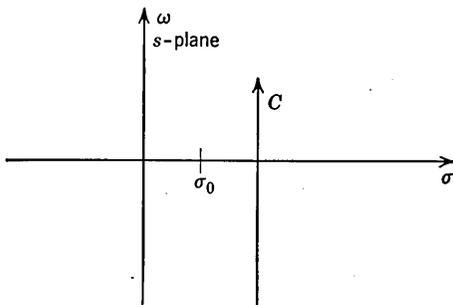


FIG. 2. Path of integration for inverse of Laplace transform.

$+\infty$ , called the *abscissa of absolute convergence*, such that the integral (91) exists for  $\sigma > \sigma_0$ . If  $\sigma_0 = -\infty$ , all values of  $\sigma$  are allowable; if  $\sigma_0 = +\infty$ , no values are allowed.

Further properties of the Laplace transform and its applications are discussed in Chap. 9.

## 9. OTHER TRANSFORMS

The *two-sided Laplace transform* is defined as

$$(94) \quad L_1[f] = F(s) = \int_{-\infty}^{\infty} f(t)e^{-st} dt.$$

Hence it differs from the (one-sided) Laplace transform only in the lower limit of integration; thus

$$(95) \quad L_1[f] = \Phi_{\infty}[f(t)e^{-\sigma t}],$$

with no requirement that  $f(t)$  be 0 for  $t < 0$ . The two-sided transform is thus a generalization of the one-sided transform.

The *Laplace-Stieltjes transform* of  $g(t)$  is defined as

$$(96) \quad G(s) = \int_0^{\infty} e^{-st} dg(t).$$

The integral on the right is an improper Stieltjes integral; it has meaning if  $g(t)$  is expressible as the difference of two monotone functions and if the limit as  $b \rightarrow +\infty$  of the integral from 0 to  $b$  exists. If  $g'(t) = f(t)$  exists,

then  $G(s)$  is the Laplace transform of  $f(t)$ . If  $g(t)$  is a step function with jumps at  $t_1, t_2, \dots$ , the integral reduces to a series  $\sum c_j e^{-t_j s}$ . For further information one is referred to the book of Widder (Ref. 10).

Other integral transforms have been defined and studied. These have found their main applications in the boundary value problems associated with partial differential equations; they could conceivably be applied to ordinary linear differential equations with variable coefficients, on the basis of the analysis of Sect. 5.

The *Legendre transform* is an example which assigns to each  $f(t)$ ,  $-1 \leq t \leq 1$ , the function

$$(97) \quad T[f] = \phi(n) = \int_{-1}^1 f(t) P_n(t) dt, \quad n = 0, 1, 2, \dots,$$

where  $P_n(t)$  is the  $n$ th Legendre polynomial. The transformation has the property

$$(98) \quad \begin{aligned} T[R\{f\}] &= -n(n+1)\phi(n), \\ R\{f\} &= \frac{d}{dt} \left[ (1-t^2) \frac{d}{dt} f(t) \right]. \end{aligned}$$

Hence the transform can be applied to differential equations of form

$$(99) \quad (a_0 R^m + a_1 R^{m-1} + \dots + a_{m-1} R + a_m)x = f(t), \quad -1 \leq t \leq 1,$$

where  $a_0, \dots, a_m$  are constants. For details on the Legendre transform see Ref. 12.

The Mellin transform, Bessel transforms, Hilbert transform, and others are defined and their properties are listed in the volumes of the Bateman project (Ref. 1).

## REFERENCES

1. *Tables of Integral Transforms*, Vols. 1, 2, prepared by the staff of the Bateman manuscript project, McGraw-Hill, New York, 1954.
2. R. V. Churchill, *Modern Operational Mathematics in Engineering*, McGraw-Hill, New York, 1944.
3. G. Doetsch, *Theorie und Anwendung der Laplace Transformation*, Springer, Berlin, 1937.
4. G. Doetsch, *Handbuch der Laplace Transformation*, Vol. I, Birkhäuser, Basel, 1950.
5. G. Doetsch, H. Kniess, and D. Voelker, *Tabellen zur Laplace Transformation*, Springer, Berlin, 1947.
6. M. F. Gardner and J. L. Barnes, *Transients in Linear Systems*, Vol. I, Chapman and Hall, London, 1942.
7. T. von Kármán and M. A. Biot, *Mathematical Methods in Engineering*, McGraw-Hill, New York, 1940.

8. D. F. Lawden, *Mathematics of Engineering Systems*, Wiley, New York, 1954.
9. B. van der Pol and H. Bremmer, *Operational Mathematics Based on the Two-sided Laplace Integral*, Cambridge University Press, Cambridge, England, 1950.
10. D. V. Widder, *The Laplace Transform*, Princeton University Press, Princeton, N. J., 1941.
11. W. Kaplan, *Operational Methods for Linear Systems*, Addison-Wesley, Cambridge, Mass., 1958.
12. R. V. Churchill, The Operational Calculus of Legendre Transforms, *J. Math. Phys.*, **33**, 165-178 (1954).

# Laplace Transforms

W. Kaplan

1. Fundamental Properties	9-01
2. Transforms of Derivatives and Integrals	9-03
3. Translation. Transform of Unit Function, Step Functions, Impulse Function (Delta Function)	9-06
4. Convolution	9-08
5. Inversion	9-09
6. Application to Differential Equations	9-10
7. Response to Impulse Functions	9-15
8. Equations Containing Integrals	9-18
9. Weighting Function	9-18
10. Difference-Differential Equations	9-20
11. Asymptotic Behavior of Transforms	9-21
References	9-21

## 1. FUNDAMENTAL PROPERTIES

Of the various operational methods described in Chap. 8 those based on the Laplace transform have proved to be the most fruitful.

**Basic Definitions and Properties.** Let  $f(t)$  be a function of the real variable  $t$ , defined for  $t \geq 0$ . The *Laplace transform* of  $f(t)$  is a function  $F(s)$  of the complex variable  $s = \sigma + i\omega$ :

$$(1) \quad L[f] = F(s) = \int_0^{\infty} f(t)e^{-st} dt.$$

It is convenient to allow  $f(t)$  itself to have complex values:  $f(t) = f_1(t) + if_2(t)$ , though for most applications  $f$  will be real.

It will be assumed that  $f(t)$  is piecewise continuous (Chap. 8), although the theory can be extended to more general cases. It can be shown that there is a number  $\sigma_0$ ,  $-\infty \leq \sigma_0 \leq +\infty$ , such that

$$(2) \quad \int_0^\infty |f(t)| e^{-\sigma t} dt$$

exists for  $\sigma > \sigma_0$  and does not exist for  $\sigma < \sigma_0$ . If  $\sigma_0 = -\infty$ , the integral exists for all  $\sigma$ ; if  $\sigma_0 = +\infty$ , it exists for no  $\sigma$ ;  $\sigma_0$  is called the *abscissa of absolute convergence* of  $L[f]$ . If  $\sigma > \sigma_0$ , then the Laplace transform of  $f$  does exist. Accordingly, there is a certain half-plane in the complex  $s$ -plane for which  $L[f] = F(s)$  is defined (Fig. 1). Furthermore,  $F(s)$  is an *analytic function of  $s$  in this half-plane* (Chap. 7, Sect. 2).

REMARK. For existence of  $F(s)$ , it is sufficient that the integral in (1)

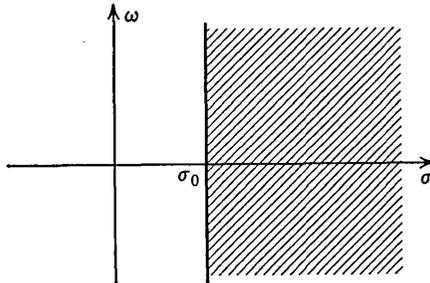


FIG. 1. Domain of definition of  $F(s) = L[f]$ .

have meaning. It can be shown that there is a number  $\sigma_1$ , the *abscissa of (conditional) convergence*, for which this integral exists, and  $\sigma_1 \leq \sigma_0$ . For most applications  $\sigma_1 = \sigma_0$  and for most operations on  $F(s)$  it is simpler to restrict  $\sigma$  to be greater than  $\sigma_0$ .

EXAMPLES OF LAPLACE TRANSFORMS. These are given in Table 1. For extensive tables one is referred to Refs. 1, 5, 6, Chap. 8.

**Existence.** For practical purposes the condition that the Laplace transform exist for some  $\sigma$  is that the function  $f(t)$  should not grow too rapidly as  $t \rightarrow +\infty$ . For example,  $e^{t^2}$ ,  $e^{e^t}$ , do not have Laplace transforms. In general, a function of *exponential type*, i.e., a function for which  $|f(t)| < e^{kt}$  for some  $k$  and for  $t$  sufficiently large, has a Laplace transform  $F(s)$ .

**Linearity.** The Laplace transform is a *linear operator*. More precisely, if  $L[f_1(t)] = F_1(s)$  exists for  $\sigma > \sigma_1$  and  $L[f_2(t)] = F_2(s)$  exists for  $\sigma > \sigma_2$ , then for every pair of constants  $c_1, c_2$   $L[c_1f_1 + c_2f_2]$  exists for  $\sigma > \max(\sigma_1, \sigma_2)$  and

$$(3) \quad L[c_1f_1 + c_2f_2] = c_1L[f_1] + c_2L[f_2].$$

## 2. TRANSFORMS OF DERIVATIVES AND INTEGRALS

### Rules.

$$(4) \quad L[f'(t)] = sL[f(t)] - f(0),$$

$$(5) \quad L[f''(t)] = s^2L[f(t)] - f'(0) - sf(0),$$

$$(6) \quad L[f^{(n)}(t)] = s^nL[f] - [f^{(n-1)}(0) + sf^{(n-2)}(0) + \cdots + s^{n-1}f(0)],$$

$$(7) \quad L\left[\int_0^t f(t) dt\right] = \frac{1}{s}L[f].$$

The first rule is basic here, the others being consequences of it; it is valid if (for some  $\sigma$ )  $f(t)$  and  $f'(t)$  have Laplace transforms and  $f(t)$ ,  $f'(t)$  are continuous for  $t \geq 0$ . More generally, eq. (4) is valid if only  $f(t)$  is continuous and  $f'(t)$  is continuous except for jump discontinuities. Similarly, eq. (6) is valid if the Laplace transforms exist and all derivatives concerned are continuous except perhaps the  $n$ th, which is allowed to have jump discontinuities. Rule (7) is valid if  $f$  is piecewise continuous and the transforms exist.

**EXAMPLE.** If  $f = \sin t$ ,  $f' = \cos t$ ,  $f(0) = 0$ , so that  $L[\cos t] = sL[\sin t] = s/(s^2 + 1)$ .

Of great importance is the special case of eq. (6):

$$(8) \quad L[f^{(n)}(t)] = s^nL[f], \quad \text{if } f(0) = f'(0) = \cdots = f^{(n-1)}(0) = 0.$$

Hence, if one restricts to functions with 0 initial values, *differentiation with respect to  $t$  corresponds to multiplication by  $s$ .*

TABLE 1. LAPLACE TRANSFORMS

	$f(t)$	$F(s) = L[f] = \int_0^{\infty} f(t)e^{-st} dt$	Range of $\sigma$
1	1	$1/s$	$\sigma > 0$
2	$e^{at}$	$1/(s - a)$	$\sigma > \operatorname{Re}(a)$
3	$t^n (n > -1)$	$\frac{\Gamma(n+1)}{s^{n+1}}$ or, if $n = 0, 1, 2, \dots, \frac{n!}{s^{n+1}}$	$\sigma > 0$
4	$t^n e^{at} (n > -1)$	$\frac{\Gamma(n+1)}{(s-a)^{n+1}}$ or, if $n = 0, 1, 2, \dots, \frac{n!}{(s-a)^{n+1}}$	$\sigma > \operatorname{Re}(a)$
5	$\cos at$	$s/(s^2 + a^2)$	$\sigma >  \operatorname{Im} a $
6	$\sin at$	$a/(s^2 + a^2)$	$\sigma >  \operatorname{Im} a $
7	$\cosh at$	$s/(s^2 - a^2)$	$\sigma >  \operatorname{Re} a $
8	$\sinh at$	$a/(s^2 - a^2)$	$\sigma >  \operatorname{Re} a $
9	$t^n \cos at (n > -1)$	$\frac{\Gamma(n+1)}{2} \frac{(s+ai)^{n+1} + (s-ai)^{n+1}}{(s^2 + a^2)^{n+1}}$	$\sigma >  \operatorname{Im} a $
10	$t^n \sin at (n > -1)$	$\frac{\Gamma(n+1)}{2i} \frac{(s+ai)^{n+1} - (s-ai)^{n+1}}{(s^2 + a^2)^{n+1}}$	$\sigma >  \operatorname{Im} a $
11	$\cos^2 t$	$\frac{1}{2} \left( \frac{1}{s} + \frac{s}{s^2 + 4} \right)$	$\sigma > 0$
12	$\sin^2 t$	$\frac{1}{2} \left( \frac{1}{s} - \frac{s}{s^2 + 4} \right)$	$\sigma > 0$
13	$\sin at \sin bt$	$\frac{2abs}{[s^2 + (a+b)^2][s^2 + (a-b)^2]}$	$\sigma > \operatorname{Max}(\alpha, \beta)$ $\alpha =  \operatorname{Im}(a+b) $ $\beta =  \operatorname{Im}(a-b) $

14	$e^{at} \sin(bt + c)$	$\frac{(s - a) \sin c + b \cos c}{(s - a)^2 + b^2}$	$\sigma > \text{Max}(\alpha, \beta)$ $\alpha = \text{Re}(a + bi)$ $\beta = \text{Re}(a - bi)$
15	1 for $2n \leq t < 2n + 1$ 0 for $2n + 1 \leq t < 2n + 2$ $n = 0, 1, 2, \dots$ (square wave)	$\frac{1}{s(1 + e^{-s})}$	$\sigma > 0$
16	1 for $a \leq t \leq b < \infty$ 0 for $0 \leq t < a$ and $t > b$	$\frac{e^{-as} - e^{-bs}}{s}$	all $\sigma$
17	0 for $0 \leq t < b$ , 1 for $t \geq b$	$\frac{e^{-bs}}{s}$	$\sigma > 0$
18	$t, 0 \leq t \leq 1$ $1, t \geq 1$	$\frac{1 - e^{-s}}{s^2}$	$\sigma > 0$
19	$t, 0 \leq t \leq 1$ $2 - t, 1 \leq t \leq 2$ $0, t \geq 2$	$\frac{(1 - e^{-s})^2}{s^2}$	all $\sigma$
20	$a - \frac{a}{b}  t - (2n + 1)b $ for $2nb \leq t \leq (2n + 2)b$ , $n = 0, 1, \dots, b > 0$ , $a$ real (triangular wave)	$\frac{a}{b} \frac{1 + e^{-bs}}{s^2}$	$\sigma > 0$
21	$a(t - nb)$ for $nb \leq t < (n + 1)b$ (sawtooth wave)	$\frac{a(1 + bs - e^{bs})}{s^2(1 - e^{bs})}$	$\sigma > 0$

### 3. TRANSLATION. TRANSFORM OF UNIT FUNCTION, STEP FUNCTIONS, IMPULSE FUNCTION (DELTA FUNCTION)

**Translation.** In Laplace transform theory it is convenient to consider each function  $f(t)$  to be defined as 0 for  $t < 0$ . Hence for  $c \geq 0$   $f(t - c)$  is 0 for  $t < c$  and coincides with a translated  $f(t)$  for  $t > c$  (Fig. 2). One finds

$$(9) \quad \begin{aligned} L[f(t - c)] &= \int_c^{\infty} f(t - c)e^{-st} dt \\ &= e^{-cs}L[f]. \end{aligned}$$

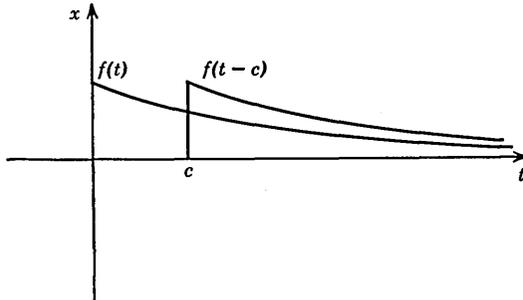


FIG. 2. Translated function.

**Unit Function.** Now let  $u(t) = 0$  for  $t \leq 0$ ,  $u(t) = 1$  for  $t > 0$ ;  $u(t)$  is called the unit function (of Heaviside). By entry 1 of Table 1,

$$(10) \quad L[u(t)] = \frac{1}{s}.$$

Hence for  $c \geq 0$

$$(11) \quad L[u(t - c)] = \frac{e^{-cs}}{s},$$

where  $u(t - c)$  is the translated unit function with jump at  $t = c$  (Fig. 3); cf. entry 17 of Table 1. A square pulse of height  $h$  (Fig. 4) can be repre-

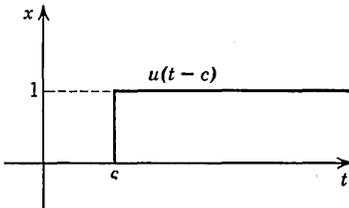


FIG. 3. Translated unit function.

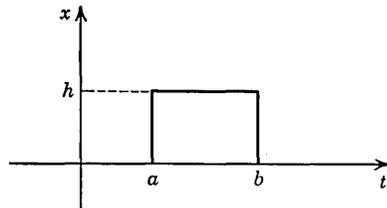


FIG. 4. Square pulse.

sented as a combination of two unit functions:

$$(12) \quad f(t) = h[u(t - a) - u(t - b)], \quad 0 \leq a < b;$$

hence its transform is

$$(13) \quad L[f] = \frac{h}{s} (e^{-as} - e^{-bs}).$$

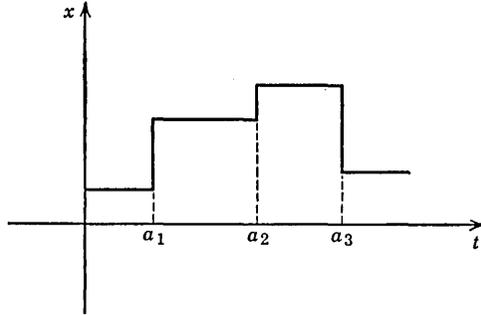


FIG. 5. Step function.

A *general step function* (Fig. 5) can be regarded as a superposition of such square pulses:

$$(14) \quad f = h_1[u(t) - u(t - a_1)] + h_2[u(t - a_1) - u(t - a_2)] + \dots;$$

hence (if the pulses do not grow too rapidly, so that  $L[f]$  exists)

$$(15) \quad L[f] = \frac{1}{s} [h_1(1 - e^{-a_1s}) + h_2(e^{-a_1s} - e^{-a_2s}) + \dots].$$

**Impulse Function.** The *unit impulse function* at  $t = 0$  is defined as the limit as  $\epsilon \rightarrow 0$  of a square pulse from  $t = 0$  to  $t = \epsilon$  and having unit area, i.e., the limit as  $\epsilon \rightarrow 0+$  of

$$(16) \quad \frac{1}{\epsilon} [u(t) - u(t - \epsilon)].$$

The limit does not exist in the ordinary sense; it can be considered as defining an "ideal" function, the delta function  $\delta(t)$ . One can consider  $\delta(t)$  to be 0 except near  $t = 0$  where  $\delta(t)$  is large and positive and has an integral equal to 1. Now

$$L \left[ \frac{u(t) - u(t - \epsilon)}{\epsilon} \right] = \frac{1 - e^{-\epsilon s}}{\epsilon s} \rightarrow 1 \text{ as } \epsilon \rightarrow 0,$$

and accordingly one defines:

$$(17) \quad L[\delta(t)] = 1.$$

The *unit impulse function* at  $t = c$  is defined as  $\delta(t - c)$  and one finds

$$(18) \quad L[\delta(t - c)] = e^{-cs}.$$

It should be noted that  $L[u(t)] = L[\delta(t)]/s$ , so that by eq. (7)  $u(t)$  can be thought of as an integral of  $\delta(t)$ :  $u(t) = \int_0^t \delta(t) dt$ . This in turn suggests interpretation of  $\delta(t)$  as  $u'(t)$ .

#### 4. CONVOLUTION

Let  $f(t)$  and  $g(t)$  be piecewise continuous for  $t \geq 0$ . Then the (*Laplace*) *convolution* of  $f$  and  $g$  is defined as

$$(19) \quad f * g = \int_0^t f(u)g(t - u) du = h(t).$$

It can be verified that  $h(t)$  is continuous for  $t \geq 0$ , also that

$$(20) \quad h(t) = \int_0^t g(u)f(t - u) du = g * f.$$

If now, for some  $\sigma$ ,  $\int_0^\infty |f(t)|e^{-\sigma t} dt$  and  $\int_0^\infty |g(t)|e^{-\sigma t} dt$  exist, then  $\int_0^\infty |h(t)|e^{-\sigma t} dt$  exists, so that  $L[f]$ ,  $L[g]$ ,  $L[h]$  exist and

$$(21) \quad L[h] = L[f * g] = L[f]L[g].$$

**Properties of the Convolution.** These are:

$$(22) \quad f * (g + h) = f * g + f * h;$$

$$(23) \quad f * (cg) = (cf) * g = c(f * g), \quad c = \text{const.};$$

$$(24) \quad f * (g * h) = (f * g) * h.$$

**Special Convolutions.** The following are useful:

$$(25) \quad e^{at} * e^{at} = te^{at};$$

$$(26) \quad e^{at} * e^{at} * \dots * e^{at} = \frac{t^{n-1}e^{at}}{(n-1)!} \quad (n \text{ factors});$$

$$(27) \quad t^m e^{at} * t^n e^{at} = \frac{m!n!}{(m+n+1)!} t^{m+n+1} e^{at};$$

$$(28) \quad e^{at} * e^{bt} = \frac{e^{at} - e^{bt}}{a - b} \quad (a \neq b).$$

## 5. INVERSION

If  $L[f] = F(s)$ , one writes  $f = L^{-1}[F]$ , thereby defining the *inverse Laplace transform*. The inverse is uniquely determined; more precisely, if  $L[f] = L[g]$  and  $f, g$  are piecewise continuous, then  $f = g$ , except perhaps at points of discontinuity.

If  $f = L^{-1}[F]$ , then as for Fourier series and integrals (Chap. 8, Sects. 8, 9),

$$(29) \quad f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(s)e^{st} d\omega = \lim_{b \rightarrow \infty} \frac{1}{2\pi} \int_{-b}^b F(s)e^{st} d\omega, \quad s = \sigma + i\omega,$$

at every  $t$  for which  $f$  has left- and right-handed derivatives; in the integrals  $\sigma$  is chosen greater than the abscissa of absolute convergence of  $L[f]$ . Under the conditions described in Chap. 8, Sect. 8, the integral represents  $\frac{1}{2}[f(t_0+) + f(t_0-)]$  at each jump discontinuity  $t_0$ . In general  $f(t)$  is defined to be 0 for  $t < 0$ , which will force a discontinuity at  $t = 0$  unless  $f(t) \rightarrow 0$  as  $t \rightarrow 0+$ ; the integral thus gives  $\frac{1}{2}f(0+)$  at  $t = 0$ .

**Conditions for Existence.** Given  $F(s)$  as a function of the complex variable  $s$ , one can ask whether  $L^{-1}[F]$  exists, i.e., whether  $F$  is the Laplace transform of some  $f(t)$ . For this to hold,  $F(s)$  must be analytic in some half-plane  $\sigma > \sigma_0$ , but this alone is not sufficient. If  $F(s)$  is analytic at  $s = \infty$  and has a zero there (Chap. 7, Sect. 5), so that

$$(30) \quad F(s) = \sum_{n=1}^{\infty} \frac{a_n}{s^n}, \quad |s| > R,$$

then  $F(s)$  is a Laplace transform:

$$(31) \quad L^{-1}[F(s)] = f(t) = \sum_{n=0}^{\infty} a_{n+1} \frac{t^n}{n!};$$

$f(t)$  is of exponential type and is an entire function of  $t$  (Chap. 7, Sect. 4). Furthermore,

$$(32) \quad f(t) = \frac{1}{2\pi i} \int_C e^{st} F(s) ds,$$

where  $C$  is a circle:  $|s| = R_0 > R$ . If in addition,  $F(s)$  is analytic for all finite  $s$  except at  $s_1, \dots, s_n$ , then  $f(t)$  equals the sum of the residues of  $F(s)e^{st}$  at  $s_1, \dots, s_n$  (Chap. 7, Sect. 5).

More general conditions that  $F(s)$  be a transform can be given. If, for example,  $F(s)$  is analytic for  $\sigma > \sigma_0 \geq 0$  and is representable in the form

$$(33) \quad F(s) = \frac{c}{s} + \frac{\mu(s)}{s^{1+\delta}} \quad (\delta > 0),$$

where  $|\mu(s)|$  is bounded for  $\sigma \geq \sigma_1 > \sigma_0$ , then  $F(s)$  is the Laplace transform of  $f(t)$ , where  $f(t)$  is given by eqs. (29), with  $\sigma = \sigma_1$ .

If  $F(s)$  is a proper rational function of  $s$ :  $F(s) = P(s)/Q(s)$ , then eq. (32) is applicable and the integral can be computed by residues. If in particular  $Q(s)$  has only simple roots  $s_1, \dots, s_n$ , then by Chap. 7, Sect. 5,  $e^{st}P/Q$  has residue  $\exp(s_k t)P(s_k)/Q'(s_k)$  at  $s_k$ , so that

$$(34) \quad L^{-1} \left[ \frac{P(s)}{Q(s)} \right] = \sum_{k=1}^n \frac{e^{s_k t} P(s_k)}{Q'(s_k)}.$$

This corresponds to the *Heaviside expansion formula* (Chap. 8, Sect. 1).

*Particular inverse transforms* can be read off Table 1 (Sect. 1) or the accompanying Table 2. Others can be deduced from these by linearity and the various rules such as (4)–(7), (9), and with the aid of convolutions. Extensive lists are given in Refs. 1, 5, 6 of Chap. 8.

**Rules for Finding Laplace Transforms and Their Inverses.** If  $f(t)$  has period  $T$ , then

$$(35) \quad L[f] = \frac{1}{1 + e^{-sT}} \int_0^T e^{-st} f(t) dt.$$

For general  $f(t)$  with transform  $F(s)$ ,

$$(36) \quad L[f(at)] = \frac{1}{a} F\left(\frac{s}{a}\right), \quad a > 0;$$

$$(37) \quad L[e^{-at}f] = F(s + a);$$

$$(38) \quad L[t^n f] = (-1)^n F^{(n)}(s), \quad n = 1, 2, \dots;$$

$$(39) \quad L[t^{-n}f] = \int_s^\infty \dots \int_s^\infty F(s) ds \dots ds \quad (n = 1, 2, \dots).$$

$n$  times

## 6. APPLICATION TO DIFFERENTIAL EQUATIONS

**Characteristic Function.** Let  $a_0, \dots, a_n$  be constants, with  $a_0 \neq 0$ . The function  $V(s) = a_0 s^n + \dots + a_n$  will be termed the characteristic function associated with the differential equation

$$(40) \quad a_0 \frac{d^n x}{dt^n} + \dots + a_{n-1} \frac{dx}{dt} + a_n x = f(t).$$

**Transfer Function.** The function

$$(41) \quad Y(s) = \frac{1}{V(s)} = \frac{1}{a_0 s^n + \dots + a_n}$$

will be termed the transfer function.

**Solutions.** Let  $f(t)$  be piecewise continuous for  $t \geq 0$  and have an absolutely convergent Laplace transform for  $\sigma > \sigma_0$ . A solution  $x(t)$  of eq. (40) satisfying given initial conditions

$$(42) \quad x(0) = \alpha_0, \quad x'(0) = \alpha_1, \dots, x^{(n-1)}(0) = \alpha_{n-1}$$

is obtained as follows. One forms the Laplace transform of both sides of eq. (40), applies the rule (6), and obtains the transformed equation

$$(43) \quad V(s)X(s) - Q(s) = F(s),$$

where  $X = L[x]$ ,  $F = L[f]$  and

$$(44) \quad Q(s) = \alpha_0 a_0 s^{n-1} + (\alpha_0 a_1 + \alpha_1 a_0) s^{n-2} + \dots + (\alpha_0 a_{n-1} + \dots + \alpha_{n-1} a_0).$$

Accordingly,

$$(45) \quad X(s) = \frac{Q(s)}{V(s)} + \frac{F(s)}{V(s)} = Y(s)Q(s) + Y(s)F(s),$$

$$(46) \quad x(t) = L^{-1}[Y(s)Q(s) + Y(s)F(s)] = L^{-1}[Y(s)Q(s)] + L^{-1}[Y(s)F(s)].$$

Since  $Y(s)Q(s)$  is a proper rational function, its inverse can be found by residues as in Chap. 9, Sect. 5. The inverse of  $Y(s)F(s)$  can be found in a variety of ways. In particular,  $Y(s)$  has an inverse transform  $y(t)$  and

$$(47) \quad L^{-1}[Y(s)F(s)] = y(t) * f(t) = \int_0^t y(u)f(t-u) du.$$

Thus both terms in eqs. (46) are well defined and it can be shown that  $x(t)$  is the solution sought;  $x(t)$  has continuous derivatives through the  $(n-1)$ st order and an  $n$ th derivative which is continuous except where  $f(t)$  is discontinuous.

The formula (47) defines  $y * f$  if  $f(t)$  is piecewise continuous for  $t \geq 0$ , even though  $f(t)$  may grow very rapidly as  $t \rightarrow +\infty$ . If  $V(s)$  has only simple roots  $s_1, \dots, s_n$ , so that

$$(48) \quad Y(s) = \sum_{j=1}^n \frac{A_j}{s - s_j}, \quad y(t) = \sum_{j=1}^n A_j e^{s_j t},$$

then

$$(49) \quad L^{-1}[YF] = \sum_{j=1}^n A_j e^{s_j t} \int_0^t f(u) e^{-s_j u} du.$$

If  $V$  has multiple roots, each multiple root  $s_j$  gives rise to terms of form

TABLE 2. INVERSE LAPLACE TRANSFORMS

$F(s)$	$L^{-1}[F(s)] = f(t)$
1 $\frac{c}{as + b}$	$\frac{c}{a} e^{(-b/a)t}$
2 $\frac{ps + q}{(s + \alpha)(s + \beta)}$	$\frac{(q - p\alpha)e^{-\alpha t} - (q - p\beta)e^{-\beta t}}{\beta - \alpha}, \quad \alpha \neq \beta$
3 $\frac{ps + q}{(s + \alpha)^2}$	$e^{-\alpha t}[p + (q - \alpha p)t]$
4 $\frac{ps + q}{as^2 + bs + c}, \quad b^2 - 4ac > 0$	$-\frac{1}{\mu} [(q - p\alpha)e^{-\alpha t} - (q - p\beta)e^{-\beta t}],$ $\alpha = \frac{b + \mu}{2a}, \quad \beta = \frac{b - \mu}{2a}, \quad \mu = \sqrt{b^2 - 4ac}$
5 $\frac{ps + q}{as^2 + bs + c}, \quad b^2 - 4ac < 0$	$e^{(-b/2a)t} \left[ \frac{p}{a} \cos \frac{\mu}{2a} t + \frac{2aq - pb}{a\mu} \sin \frac{\mu}{2a} t \right],$ $\mu = \sqrt{4ac - b^2}$
6 $\frac{ps^2 + qs + r}{(s + \alpha)(s + \beta)(s + \gamma)},$ $\alpha, \beta, \gamma$ distinct	$\frac{-1}{ABC} [A(p\alpha^2 - q\alpha + r)e^{-\alpha t}$ $+ B(p\beta^2 - q\beta + r)e^{-\beta t}$ $+ C(p\gamma^2 - q\gamma + r)e^{-\gamma t}],$ $A = \beta - \gamma, \quad B = \gamma - \alpha, \quad C = \alpha - \beta$
7 $\frac{ps^2 + qs + r}{(s + \alpha)^2(s + \beta)}, \quad \alpha \neq \beta$	$\frac{p\beta^2 - q\beta + r}{(\beta - \alpha)^2} e^{-\beta t} + \left[ \frac{p\alpha^2 - q\alpha + r}{(\beta - \alpha)} t \right.$ $\left. + \frac{p\alpha^2 - 2\alpha\beta p + q\beta - r}{(\beta - \alpha)^2} \right] e^{-\alpha t}$
8 $\frac{ps^2 + qs + r}{(s + \alpha)^3}$	$pe^{-\alpha t} + (q - 2p\alpha)te^{-\alpha t}$ $+ (p\alpha^2 - q\alpha + r) \frac{t^2}{2} e^{-\alpha t}$

TABLE 2. INVERSE LAPLACE TRANSFORMS (Continued)

	$F(s)$	$L^{-1}\{F(s)\} = f(t)$
9	$\frac{ps^2 + qs + r}{(s + \alpha)(as^2 + bs + c)},$ $a\alpha^2 - b\alpha + c \neq 0$	$\frac{M}{N} e^{-\alpha t} + \frac{1}{N} L^{-1} \left[ \frac{Bs + C}{as^2 + bs + c} \right],$ $M = p\alpha^2 - q\alpha + r, N = a\alpha^2 - b\alpha + c,$ $B = (aq - bp)\alpha + pc - ar,$ $C = (ar - pc)\alpha + qc - br$
10	$\frac{ps^3 + qs^2 + rs + u}{(as^2 + bs + c)(As^2 + Bs + C)}$ $as^2 + bs + c \text{ and}$ $As^2 + Bs + C \text{ having no}$ $\text{common roots}$	$L^{-1} \left[ \frac{p_0s + q_0}{as^2 + bs + c} \right] + L^{-1} \left[ \frac{p_1s + q_1}{As^2 + Bs + C} \right].$ <p>To find <math>p_0, q_0, p_1, q_1</math>, compute:</p> $\lambda_0 = a(ar - cp) + b(bp - aq),$ $\mu_0 = a(au - cq) + bcp,$ $\sigma_0 = a(bu - cr) + c^2p, \beta = aB - bA$ $\gamma = aC - cA, \delta_0 = a\gamma^2 - b\beta\gamma + c\beta^2,$ $\lambda_1 = A(Ar - Cp) + B(Bp - Aq),$ $\mu_1 = A(Au - Cq) + BCp,$ $\sigma_1 = A(Bu - Cr) + C^2p,$ $\delta_1 = A\gamma^2 - B\beta\gamma + C\beta^2.$ <p>Then</p> $p_0 = \frac{\lambda_0\gamma - \mu_0\beta}{\delta_0}, \quad q_0 = \frac{\mu_0\gamma - \sigma_0\beta}{\delta_0},$ $p_1 = \frac{-\lambda_1\gamma + \mu_1\beta}{\delta_1}, \quad q_1 = \frac{-\mu_1\gamma + \sigma_1\beta}{\delta_1}.$
11	$\frac{ps + q}{(as^2 + bs + c)^2}, \quad b^2 - 4ac < 0$	$\frac{e^{-\alpha t}}{2a^2\beta^3} [p\beta^2t \sin \beta t + (q - \alpha p)(\sin \beta t - \beta t \cos \beta t)]$ $\alpha = \frac{b}{2a}, \quad \beta = \frac{\sqrt{4ac - b^2}}{2a}$

$A(s - s_j)^{-k}$  in  $Y(s)$ . The corresponding term in  $y(t)$  is  $At^{k-1}e^{s_j t}/(k-1)!$  and in  $L^{-1}[YF]$  is

$$(50) \quad \frac{A}{(k-1)!} e^{s_j t} \int_0^t f(u)(t-u)^{k-1} e^{-s_j u} du.$$

**Particular Solutions.** If all the initial constants  $\alpha_0, \dots, \alpha_{n-1}$  are 0, then  $Q(s) = 0$  and  $x = L^{-1}[YF]$  is the solution sought. This particular solution can be found by eqs. (47), which requires knowledge of  $y(t)$  and hence of the roots of  $V(s)$ . This can cause difficulty. An alternative is to employ eqs. (29):

$$(51) \quad x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} Y(s)F(s)e^{st} d\omega, \quad \sigma = \text{const.} > \sigma_0.$$

It may be possible to simplify this by residues or series expansions.

If  $f(t)$  is of form  $e^{bt}p(t)$ , where  $p(t)$  is a polynomial of degree  $m$  in  $t$ , a particular solution can be found explicitly without finding the roots of  $V(s)$ . If  $V(b) \neq 0$ , the particular solution is

$$(52) \quad x(t) = e^{bt} \left[ Y(b)p(t) + \frac{Y'(b)}{1!} p'(t) + \frac{Y''(b)}{2!} p''(t) + \dots + \frac{Y^{(m)}(b)}{m!} p^{(m)}(t) \right].$$

If  $V(b) = 0$ , then  $V(s) = (s-b)^k W(s)$ ,  $W(b) \neq 0$ . Let  $Z(s) = 1/W(s)$  and let  $p_1(t)$  be the polynomial obtained by integrating

$$(53) \quad Z(b)p(t) + Z'(b)p'(t) + \dots + \frac{Z^{(m)}(b)}{m!} p^{(m)}(t)$$

$k$  times from 0 to  $t$ . Then  $x = e^{bt}p_1(t)$  is a particular solution of eq. (40). In both cases it can be verified that

$$L[x] = Y(s)L[e^{bt}p] + Y(s)R(s),$$

where  $R$  is a polynomial of degree less than that of  $V$  (in fact less than that of  $W(s)$  in the second case).

**Simultaneous Equations.** Similar methods are employed for simultaneous linear differential equations with constant coefficients in unknowns  $x_1, x_2, \dots$ . One applies the Laplace transformation to the equations, thereby obtaining equations for  $X_1(s), X_2(s), \dots$ ; in forming these new equations, certain initial conditions for  $x_1, x_2, \dots$  are assumed. The equations are simultaneous algebraic equations for  $X_1(s), X_2(s), \dots$  and can

be solved by elimination or determinants. When  $X_j(s)$  is known,  $x_j(t)$  can be found by forming the inverse transforms.

EXAMPLE.

$$\frac{d^2x}{dt^2} + 2\frac{dy}{dt} + y = 13e^{2t}, \quad \frac{dx}{dt} - 2x + \frac{d^2y}{dt^2} + 3\frac{dy}{dt} + 5y = 15e^{2t};$$

when  $t = 0$ ,  $x = 1$ ,  $y = 0$ ,  $dx/dt = 0$ ,  $dy/dt = 1$ . Hence

$$s^2X(s) + (2s + 1)Y(s) = s + \frac{13}{s - 2},$$

$$(s - 2)X(s) + (s^2 + 3s + 5)Y(s) = 2 + \frac{15}{s - 2},$$

$$X = \frac{s^4 + s^3 + 8s^2 + 5s + 54}{(s - 2)(s + 1)(s + 2)(s^2 + 1)},$$

$$Y = \frac{s^3 + 15s^2 - 17s + 26}{(s - 2)(s + 1)(s + 2)(s^2 + 1)},$$

$$X = \frac{2}{s - 2} - \frac{19}{2(s + 1)} + \frac{21}{5(s + 2)} + \frac{43s - 51}{10(s^2 + 1)},$$

$$Y = \frac{1}{s - 2} - \frac{19}{2(s + 1)} + \frac{28}{5(s + 2)} + \frac{29s + 7}{10(s^2 + 1)},$$

$$x = 2e^{2t} - \frac{19}{2}e^{-t} + \frac{21}{5}e^{-2t} + \frac{43 \cos t - 51 \sin t}{10},$$

$$y = e^{2t} - \frac{19}{2}e^{-t} + \frac{28}{5}e^{-2t} + \frac{29 \cos t + 7 \sin t}{10}.$$

## 7. RESPONSE TO IMPULSE FUNCTIONS

For many applications it is important to consider the response of a linear system to the impulse function  $\delta(t)$  or to other ideal functions such as  $\delta'(t)$ ,  $\delta''(t)$ ,  $\dots$ .

EXAMPLE 1. Consider the equation

$$\frac{dx}{dt} + x = \delta(t), \quad x(0) = \alpha_0.$$

If one applies the Laplace transform mechanically to both sides and em-

plys the rule:  $L[\delta(t)] = 1$  (Sect. 3), one finds:

$$X(s) = \frac{1 + \alpha_0}{s + 1}, \quad x(t) = (1 + \alpha_0)e^{-t}.$$

Hence  $x(t)$  has a discontinuity at  $t = 0$  (Fig. 6);  $x$  jumps from the assigned initial value  $\alpha_0$  to the value  $1 + \alpha_0$ .

EXAMPLE 2. Similarly,

$$\frac{d^2x}{dt^2} + \frac{dx}{dt} = \delta(t)$$

is found to have the solution (Fig. 7)  $x = 1 + \alpha_0 + \alpha_1 - (1 + \alpha_1)e^{-t}$ , with

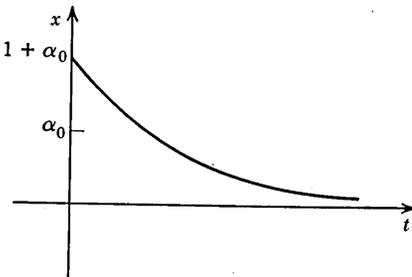


FIG. 6. Response of first order system to  $\delta$ -function.

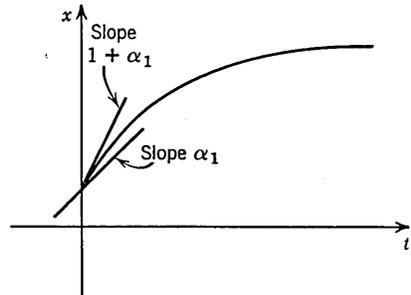


FIG. 7. Response of second order system to  $\delta$ -function.

$\alpha_0 = x(0)$ ,  $\alpha_1 = x'(0)$ . Here there is no discontinuity of  $x(t)$  at  $t = 0$ , but  $x'(t)$  has a discontinuity, jumping from the assigned initial slope of  $\alpha_1$  to the slope  $1 + \alpha_1$ .

It should be noted that the second example can be written as follows:

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} + y = \delta(t);$$

thus its solution is an *integral* of the solution of the first example. Each such integration reduces the type of discontinuity. In general,

$$V(D)x = \delta(t), \quad V(D) = a_0D^n + \dots, \quad a_0 \neq 0,$$

has a solution which has a jump in the  $(n - 1)$ st derivative at  $t = 0$ , but no jumps in the derivatives of lower order. For the equation

$$V(D)x = \delta(t - c), \quad c > 0,$$

a similar conclusion holds, with the discontinuity occurring at  $t = c$ .

One can interpret  $\delta(t - c)$  as  $\frac{d}{dt}u(t - c)$ , if one forms the transforms by the rule  $L[f'] = sL[f]$ , ignoring the discontinuity which would make the rule inapplicable. For then

$$L\left[\frac{d}{dt}u(t - c)\right] = sL[u(t - c)] = e^{-cs},$$

in accordance with eq. (18). This suggests the general procedure.

**General Procedure.** For the differential equation

$$V(D)x = \frac{df}{dt},$$

in which  $f$  has a jump discontinuity at  $t = c$  but has otherwise continuous derivatives, one should take transforms ignoring the discontinuity:

$$V(s)L[x] = sL[f] - f(0).$$

Under similar conditions on  $f$ , a similar procedure can be used for higher derivatives, and for the general equation

$$V(D)x = W(D)f.$$

If the order of  $V(D)$  is less than that of  $W(D)$ ,  $x$  will itself be an ideal function; otherwise  $x$  will merely show some discontinuity at  $t = c$ . Similar remarks apply when there are several jump discontinuities.

Let  $f$  have continuous derivatives of all orders except for  $t = c$ , at which the derivatives have limiting values to the left and to the right. Then  $f$  can be written as  $f_1(t) + k_1u(t - c)$ , where  $f_1(t)$  is continuous at  $t = c$ ; correspondingly,  $f'(t) = f'_1(t) + k_1\delta(t - c)$ , where  $f'_1(t)$  is discontinuous at  $t = c$ . Thus

$$\begin{aligned} f'(t) &= f'_2(t) + k_2u(t - c) + k_1\delta(t - c) \\ f''(t) &= f'_2(t) + k_2\delta(t - c) + k_1\delta'(t - c) \\ &= f_3(t) + k_3u(t - c) + k_2\delta(t - c) + k_1\delta'(t - c), \\ &\dots \end{aligned}$$

Computation of  $L[f']$ ,  $L[f'']$ ,  $\dots$ , as described above, is then equivalent to that obtained by writing

$$\begin{aligned} L[f'] &= L[f_2] + k_2L[u(t - c)] + k_1L[\delta(t - c)] \\ L[f''] &= L[f_3] + k_3L[u(t - c)] + k_2L[\delta(t - c)] + k_1L[\delta'(t - c)] \\ &\dots, \end{aligned}$$

if one agrees that

$$(54) \quad L[\delta^{(m)}(t - c)] = s^m e^{-cs} \quad (m = 1, 2, \dots).$$

The justification for the rules adopted lies in the fact that they give a reasonable limiting form for the response  $x(t)$ , and they meet the needs of the physical situations to which they are applied.

## 8. EQUATIONS CONTAINING INTEGRALS

The method of Sect. 6 is applicable to "integro-differential equations" such as the following:

$$(55) \quad a_0 \frac{dx}{dt} + a_1 x + a_2 \int_0^t x dt = f(t).$$

One need only apply the Laplace transformation to both sides and employ rule (7):

$$(56) \quad a_0(sX - \alpha_0) + a_1 X + a_2 \frac{X}{s} = F(s),$$

from which one can solve as before for  $X(s)$ .

One can also differentiate eq. (55) to obtain an equation of second order:

$$(57) \quad a_0 \frac{d^2x}{dt^2} + a_1 \frac{dx}{dt} + a_2 x = f'(t);$$

from eq. (55),  $a_0 x'(0) + a_1 x(0) = f(0)$ , so that one initial condition for eq. (57) is fixed. If  $f(t)$  has discontinuities,  $f'(t)$  has to be treated as an ideal function (Sect. 7); in such a case, it is simpler to use eq. (56).

It should be remarked that eq. (55) is equivalent to the system

$$a_0 \frac{dx}{dt} + a_1 x + a_2 y = f(t), \quad \frac{dy}{dt} = x,$$

with the initial conditions:  $x(0) = \alpha_0$ ,  $y(0) = 0$ . By similar devices integrals can be eliminated formally in most cases.

## 9. WEIGHTING FUNCTION

It has been seen that, for proper initial conditions, various problems lead to relations of form

$$(58) \quad X(s) = Y(s)F(s),$$

where  $F(s)$  is the Laplace transform of a driving function or "input"  $f(t)$  and  $X(s)$  is the Laplace transform of the "output"  $x(t)$ . In such cases

$Y(s)$  is termed the *transfer function*; i.e., in general, the transfer function is the ratio of the Laplace transforms of output and input.

If  $y(t)$  is the inverse Laplace transform of  $Y(s)$ , then as in Sect. 6

$$(59) \quad x(t) = y(t) * f(t) = \int_0^t f(t-u)y(u) du.$$

Accordingly,  $x(t)$  is a weighted average of  $f(t)$  over the interval from 0 to  $t$ , the value at  $t-u$  receiving weight  $y(u)$ . Since  $f(t) = 0$  for  $t < 0$ , one can also write

$$(60) \quad x(t) = \int_{-\infty}^t f(t-u)y(u) du,$$

so that the average is over the entire "past" of  $f(t)$ .

**Graphical Computation.** One can then compute  $x(t)$  at each  $t$  graphically as suggested in Fig. 8. Here  $y(u)$  is graphed against  $u$ , with

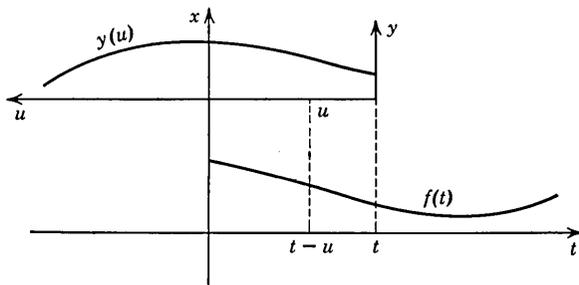


FIG. 8. Response as a weighted average.

the positive  $u$ -axis to the left and the origin above the point  $t$  on the  $t$ -axis. The value of  $f$  at  $t-u$  is multiplied by the value of  $y$  above  $t-u$  and the result is integrated to yield  $x(t)$  at the  $t$  chosen. As the graph of  $y(u)$  is moved parallel to the  $t$ -axis, the average at successive times  $t$  can be found.

**Weighting Function.** The function  $y(t) = L^{-1}[Y(s)]$  is termed the "weighting function." In view of the discussion given, this term would be justified only if  $\int_0^{\infty} y(t) dt = 1$ . But

$$(61) \quad Y(s) = \int_0^{\infty} y(t)e^{-st} dt, \quad Y(0) = \int_0^{\infty} y(t) dt,$$

provided  $Y(s)$  is defined for  $s = 0$ . Hence if  $Y(0) = 1$ , the "total weight" is 1, as desired. If  $Y(0) \neq \infty$ , one can redefine the input as a constant times  $x(t)$  and achieve the same result.

**Response to Unit Impulse.** If  $\epsilon$  is very small and  $f(t)$  is a square pulse of height  $1/\epsilon$  from  $t = 0$  to  $t = \epsilon$ , then eqs. (59) show that approximately

$$x(t) = \frac{1}{\epsilon} y(t) \cdot \epsilon = y(t);$$

as  $\epsilon \rightarrow 0$ , this can be shown to be the limiting relation. Thus *the weighting function is the response to the unit impulse function  $\delta(t)$* . This also follows from eq. (58), since if  $f(t) = \delta(t)$ ,  $L[f] = F(s) = 1$ .

One can also remark that, if  $f(t)$  is the unit function  $u(t)$ , then  $L[f] = F(s) = 1/s$ , so that by eq. (58)

$$X(s) = \frac{Y(s)}{s}, \quad Y(s) = sX(s), \quad y(t) = \frac{dx}{dt};$$

for, by eqs. (59),  $x(0) = 0$ . Thus the weighting function can be interpreted as the *derivative of the response to the unit function*. If one denotes by  $A(t)$  the response to the unit function, so that  $L[A] = Y(s)/s$ , then for an arbitrary driving function  $f(t)$ ,

$$(62) \quad X(s) = s \left( \frac{Y(s)}{s} F(s) \right), \quad x = \frac{d}{dt} \int_0^t f(t-u)A(u) du.$$

Equations (59) and (62) are equivalent to the eqs. (25), (26) of Chap. 8, Sect. 5.

## 10. DIFFERENCE-DIFFERENTIAL EQUATIONS

Because of the transformation rule:  $L[f(t-c)] = e^{-cs}L[f]$  (Sect. 3), Laplace transforms can be applied to solve linear difference-differential equations, i.e., equations of form

$$(63) \quad \sum_{k=0}^n \sum_{m=0}^M a_{mk} f^{(k)}(t-mT) = g(t);$$

it will be assumed that the coefficients  $a_{mk}$  are constants and that a solution  $f(t)$  is to be found which is equal to 0 for  $t \leq 0$  and satisfies eq. (63) for  $t > 0$ . Under these conditions

$$(64) \quad L[f^{(k)}(t-mT)] = s^k L[f(t-mT)] = s^k e^{-mTs} L[f]$$

and the transformed equation corresponding to eq. (63) is

$$(65) \quad \left( \sum_{k=0}^n \sum_{m=0}^M a_{mk} s^k e^{-mTs} \right) F(s) = G(s).$$

This can be solved for  $F(s)$  and the solution sought is  $L^{-1}[F(s)]$ . Validity of this process requires in particular that for some  $\sigma_0$  the term in parentheses in eq. (65) have no zeros in the complex  $s$ -plane for  $\sigma > \sigma_0$ . For discussion of the questions involved here see Ref. 1.

Instead of requiring that  $f(t)$  be  $\equiv 0$  for  $t < 0$  one can impose the condition that  $f(t)$  coincide with a given function  $f_0(t)$  in an "initial interval"  $-MT \leq t \leq 0$ . This case can be reduced to the previous one by first extending the definition of  $f_0(t)$  to the range  $t > 0$ , while preserving continuity, and introducing a new unknown function  $f_1(t) = f(t) - f_0(t)$ .

## 11. ASYMPTOTIC BEHAVIOR OF TRANSFORMS

In general the behavior of  $f(t)$  at  $t = 0$  is related to that of  $F(s) = L[f]$  as  $s \rightarrow \infty$  along the real axis, while the behavior of  $f(t)$  at  $t = +\infty$  is related to that of  $F(s)$  as  $s \rightarrow 0$  (or  $s \rightarrow \sigma_0$ ) along the real axis. A full discussion is given in the book of Doetsch (Ref. 3, Chap. 8), pp. 186-277.

If

$$(66) \quad F(s) = \frac{a_1}{s} + \frac{G(s)}{s^2},$$

where  $|G(s)| < M$  for  $\sigma > \sigma_0$ , then

$$(67) \quad \lim_{t \rightarrow 0^+} f(t) = \lim_{s \rightarrow \infty} sF(s) \quad (s \text{ real}).$$

If  $f(t)$  and  $f'(t)$  have convergent Laplace transforms for  $\sigma > 0$  and  $f(t)$  has a limit as  $t \rightarrow +\infty$ , then

$$(68) \quad \lim_{t \rightarrow \infty} f(t) = \lim_{s \rightarrow 0} sF(s) \quad (s \text{ real}).$$

## REFERENCES

1. R. Bellman and J. M. Danskin, The Stability Theory of Differential Difference Equations, *Proceedings of the Symposium on Nonlinear Circuit Analysis*, Vol. II, pp. 107-128, Polytechnic Institute of Brooklyn, New York, 1953.

See also the list following Chap. 8.



## Conformal Mapping

W. Kaplan

1. Definition of Conformal Mapping. General Properties	10-01
2. Linear Fractional Transformations	10-05
3. Mapping by Elementary Functions	10-06
4. Schwarz-Christoffel Mappings	10-08
5. Application of Conformal Mapping to Boundary Value Problems	10-09
References	10-11

### 1. DEFINITION OF CONFORMAL MAPPING. GENERAL PROPERTIES

**Definitions.** Let  $u = f(x, y)$ ,  $v = g(x, y)$  be two real functions of the real variables  $x, y$ , both defined in an open region  $D$  of the  $xy$ -plane. As  $(x, y)$  varies in  $D$  (Fig. 1), the corresponding point  $(u, v)$  varies in a set  $D_1$  and one says that the equations

$$(1) \quad u = f(x, y), \quad v = g(x, y)$$

define a *transformation* or *mapping*  $T$  of  $D$  onto  $D_1$  (Chap. 1, Sect. 3). If for each  $(u, v)$  in  $D_1$  there is precisely one  $(x, y)$  in  $D$  such that  $u = f(x, y)$ ,  $v = g(x, y)$ , then the transformation  $T$  is said to be one-to-one, and  $T$  has an inverse  $T^{-1}$ , defined by equations

$$(2) \quad x = \phi(u, v), \quad y = \psi(u, v),$$

obtained by solving eqs. (1) for  $x$  and  $y$  in terms of  $u$  and  $v$ .

Now let  $T$ , defined by eqs. (1), be a mapping of  $D$  onto  $D_1$ . In addition, let  $f(x, y)$  and  $g(x, y)$  have continuous first partial derivatives in  $D$ . The

mapping  $T$  is said to be *conformal* if, for each pair of curves  $C_1, C_2$  meeting at a point  $(x_0, y_0)$  of  $D$ , the corresponding curves  $C_1, C_2$  meeting at  $(u_0, v_0)$  form an angle  $\alpha$  at  $(u_0, v_0)$  equal to that formed by  $C_1^*, C_2^*$  at  $(x_0, y_0)$ . It is assumed that  $C_1, C_2$  are directed curves and have well-defined tangent vectors at  $(x_0, y_0)$  so that  $C_1^*, C_2^*$  also have tangent vectors at  $(u_0, v_0)$ . The angle  $\alpha$  is then measured between the tangent vectors. It is customarily a signed angle and measured, e.g., from  $C_1$  to  $C_2$  and, correspondingly, from

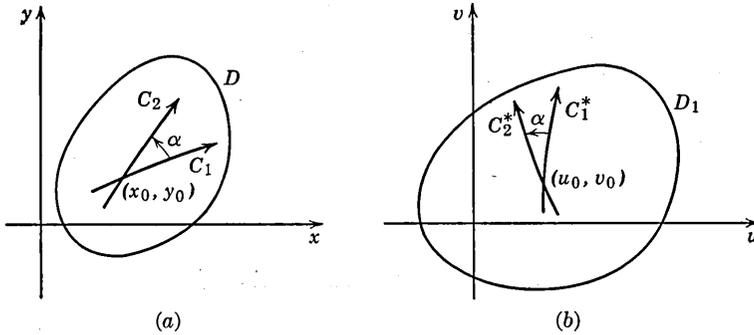


FIG. 1. Conformal mapping: (a)  $z$ -plane, (b)  $w$ -plane.

$C_1^*$  to  $C_2^*$ . Conformality then means that the corresponding angles are equal and have the same sense, as in Fig. 1. To emphasize this, one can write more explicitly that  $T$  is to be conformal and sense-preserving. For most applications  $T$  is assumed to be one-to-one. Conformality of  $T$  then implies conformality of  $T^{-1}$ .

**THEOREM 1.** Let (1) define a mapping  $T$  of  $D$  onto  $D_1$ . Let  $f(x, y)$  and  $g(x, y)$  have continuous first partial derivatives in  $D$ . Then  $T$  is conformal and sense-preserving if and only if the Cauchy-Riemann equations

$$(3) \quad \frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

hold in  $D$  and the Jacobian  $\partial(u, v)/\partial(x, y) \neq 0$  in  $D$ .

By virtue of this theorem, the theory of conformal mapping is related to the theory of analytic functions of a complex variable (Chap. 7). One can use complex notation:

$$(4) \quad z = x + iy, \quad w = u + iv, \quad i = \sqrt{-1},$$

and the transformation  $T$  is then simply a complex function  $w = F(z)$  defined in  $D$ . The mapping  $w = F(z)$  is conformal precisely when  $F$  is analytic in  $D$  and  $F'(z) \neq 0$  in  $D$ . (See Ref. 2.)

REMARK. If  $F'(z)$  is 0 at a point  $z_0$ , then  $z_0$  is termed a *critical point* of  $F(z)$ . A function  $w = F(z)$  cannot define a conformal mapping of any open region  $D$  containing a critical point  $z_0$ . The behavior of  $F(z)$  near a critical point is typified by the behavior of  $z^n$  near  $z = 0$ , for  $n = 2, 3, \dots$ ; except for  $w = 0$ , each  $w$  has  $n$  inverse values  $w^{1/n}$ . Curves meeting at angle  $\alpha$  at  $z = 0$  are transformed onto curves meeting at angle  $n\alpha$  at  $w = 0$ . The absence of critical points does not guarantee that  $F(z)$  describes a one-to-one mapping; all that can be said is that, if  $F'(z_0) \neq 0$ , then  $w = F(z)$  does define a one-to-one conformal mapping of some sufficiently small region containing  $z_0$ .

**Geometrical Meaning of Conformality.** Let  $w = F(z)$  define a one-to-one conformal mapping of  $D$  on  $D_1$ . Then each geometrical figure

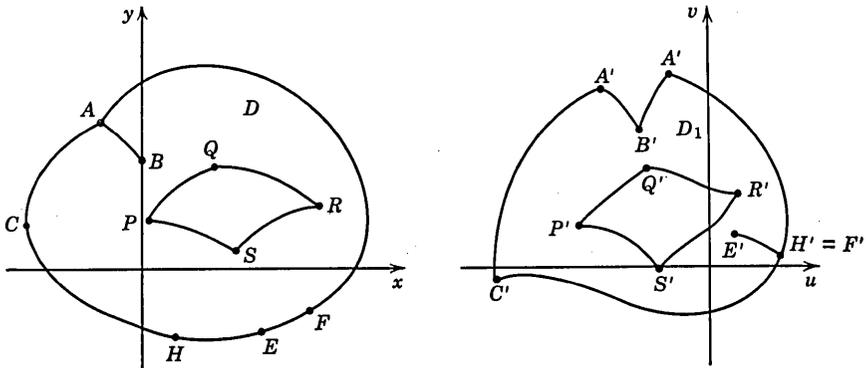


FIG. 2. Behavior of mapping in the interior and on the boundary.

in  $D$  will correspond to one in  $D_1$  which is similar in a certain sense; if the first figure is bounded by smooth arcs, the second will be bounded by similar arcs and corresponding pairs of arcs form the same angle (Fig. 2). The lines  $x = \text{const.}$ ,  $y = \text{const.}$  in  $D$  form two families of curves meeting at right angles; hence these correspond to curves in  $D_1$  formed of one family and of its family of orthogonal trajectories (Fig. 3). Similarly the curves  $u = \text{const.}$  form orthogonal trajectories of the curves  $v = \text{const.}$  On the boundary of  $D$  conformality may break down. In general there is some sort of continuous correspondence between boundary points of  $D$  and those of  $D_1$ . If  $D$  and  $D_1$  are each bounded by several simple closed curves, and  $F$  is one-to-one, then the mapping  $F$  and its inverse can indeed be extended continuously to the boundaries. Commonly there are points at which conformality is violated in that two boundary arcs of  $D$  meeting at angle  $\alpha$  correspond to boundary arcs of  $D_1$  meeting at angle  $\beta \neq \alpha$ ; in particular this can mean a folding together of the boundary, as suggested in Fig. 2.

As in Chap. 7, Sect. 5, one can adjoin the number  $\infty$  to the complex plane to form the extended plane. The mapping  $w = F(z)$  is said to be conformal in a region containing  $z = \infty$  if  $F(1/z)$  is conformal in a region containing  $z = 0$ . Similarly, one can discuss conformality in a neighborhood of a point  $z_0$  at which  $F(z_0) = \infty$ , so that  $F(z)$  has a pole, in terms of the conformality of  $1/F(z)$  near  $z_0$ .

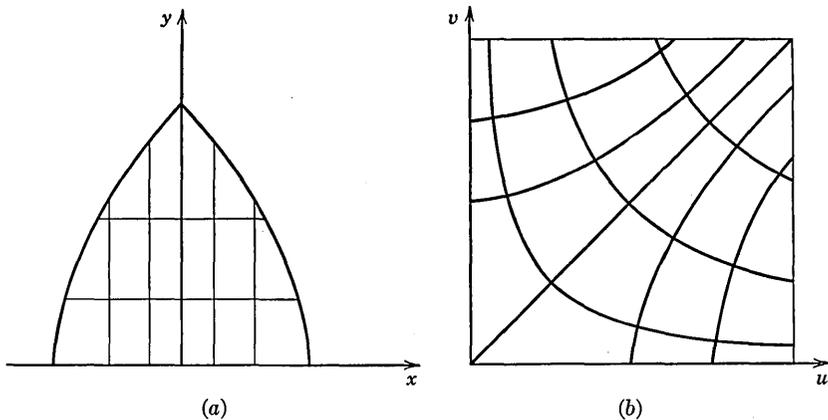


FIG. 3. Level curves of  $x$  and  $y$ : (a)  $z$ -plane, (b)  $w$ -plane.

**Conformal Equivalence.** Two regions  $D, D_1$  are said to be *conformally equivalent* if there is a one-to-one conformal mapping  $w = F(z)$  of  $D$  on  $D_1$  (so that the inverse function maps  $D_1$  conformally on  $D$ ). Conformally equivalent regions must have the same connectivity; i.e., if  $D$  is simply connected, then so is  $D_1$ ; if  $D$  is doubly connected, so is  $D_1$ . However, having the same connectivity does not guarantee conformal equivalence. If  $D$  is simply connected then  $D$  is conformally equivalent to one and only one of the following three: (a) the interior of a circle; (b) the finite plane; (c) the extended plane. In particular, one has the following theorem.

**THEOREM 2 (RIEMANN MAPPING THEOREM).** *Let  $D$  be a simply connected region of the finite  $z$ -plane, not the whole finite plane. Let  $z_0$  be a point of  $D$ , and let  $\alpha$  be a given real number. Then there exists a one-to-one conformal mapping  $w = F(z)$  of  $D$  onto the circle  $|w| < 1$  such that  $F(z_0) = 0$  and  $\arg F'(z_0) = \alpha$ . Furthermore,  $F(z)$  is uniquely determined.*

From this theorem it follows that the one-to-one conformal transformations of  $D$  onto  $|w| < 1$  depend on three real parameters:  $x_0 = \operatorname{Re}(z_0)$ ,  $y_0 = \operatorname{Im}(z_0)$  and  $\alpha$ . These parameters can be chosen in other ways. For example, three boundary points of  $D$  can be made to correspond to 3 points on  $|w| = 1$  (in the same "cyclic order").

2. LINEAR FRACTIONAL TRANSFORMATIONS

Each function

$$(5) \quad w = \frac{az + b}{cz + d}, \quad \begin{vmatrix} a & b \\ c & d \end{vmatrix} \neq 0,$$

where  $a, b, c, d$  are complex constants, defines a linear fractional transformation. Each such transformation is a one-to-one conformal mapping of the extended  $z$ -plane onto the extended  $w$ -plane. Special cases of eqs. (5) are the following:

**Translations.** The general form is

$$(6) \quad w = z + b.$$

Each point  $z$  is displaced through the vector  $b$ .

**Rotation Stretchings.** The general form is

$$(7) \quad w = az = Ae^{i\alpha}z.$$

The value of  $w$  is obtained by rotating  $z$  about the origin through angle  $\alpha$  and then increasing or decreasing the distance from the origin in the ratio  $A$  to 1.

**Linear Integral Transformations.**

$$(8) \quad w = az + b.$$

Each transformation (8) is equivalent to a rotation stretching followed by a translation.

**Reciprocal Transformation.**

$$(9) \quad w = \frac{1}{z}.$$

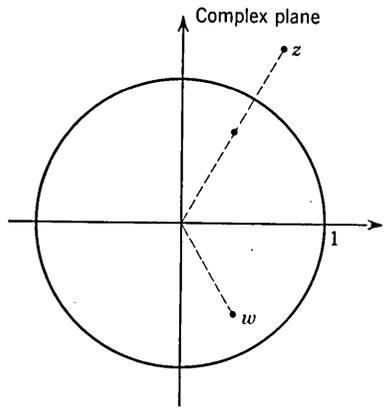


FIG. 4. The transformation  $w = 1/z$ .

Here  $|w| = 1/|z|$  and  $\arg w = -\arg z$ .

Hence  $w$  is obtained from  $z$  by "inversion" in the circle  $|z| = 1$  followed by reflection in the  $x$ -axis (Fig. 4).

**Important Conformal Mappings.** The general linear fractional transformation (5) can be composed of a succession of transformations of the special types:

$$(10) \quad w = \frac{a}{c} + \frac{bc - ad}{c} \zeta, \quad \zeta = \frac{1}{Z}, \quad Z = cz + d.$$

If one includes straight lines as "circles through  $\infty$ ," then each transformation (5) maps each circle onto a circle. By considering special regions bounded by circles and lines one obtains a variety of important conformal mappings, as illustrated in Table 1. The first three entries in the table depend on 3 real parameters and provide all conformal mappings of  $D$  on  $D_1$  in each case.

TABLE 1. IMPORTANT CONFORMAL MAPPINGS

$F(z)$	$D$	$D_1$
$e^{i\alpha} \frac{z - z_0}{1 - \bar{z}_0 z}$ $\alpha$ real, $ z_0  < 1$	$ z  < 1$	$ w  < 1$
$\frac{az + b}{cz + d}$ $a, b, c, d$ real, $ad - bc > 0$	$\text{Im}(z) > 0$	$\text{Im}(w) > 0$
$e^{i\alpha} \frac{z - z_0}{z - \bar{z}_0}$ $\alpha$ real, $\text{Im}(z_0) > 0$	$\text{Im}(z) > 0$	$ w  < 1$
$\frac{1}{z}$	region between circles $ z - a  = a,$ $ z - b  = b,$ $0 < a < b$	$\frac{1}{2b} < \text{Re}(w) < \frac{1}{2a}$

To find a one-to-one conformal transformation of a circular region  $D$  on a circular region  $D_1$ , one can choose three points  $z_1, z_2, z_3$  on the bounding circumference of  $D$  and three points  $w_1, w_2, w_3$  on the bounding circumference of  $D_1$ , in the same cyclic order as  $z_1, z_2, z_3$ . The equation

$$(11) \quad \frac{w - w_2}{w - w_3} \div \frac{w_1 - w_2}{w_1 - w_3} = \frac{z - z_2}{z - z_3} \div \frac{z_1 - z_2}{z_1 - z_3}$$

then defines a linear fractional transformation  $w = F(z)$ , mapping  $D$  on  $D_1$ , and moreover  $F(z_j) = w_j$  for  $j = 1, 2, 3$ . The left-hand side of eq. (11) is an expression formed from four complex numbers  $w, w_1, w_2, w_3$ , termed the *cross-ratio* of the four numbers in the order given, and denoted by  $[w, w_1, w_2, w_3]$ . Equation (11) states that  $[w, w_1, w_2, w_3] = [z, z_1, z_2, z_3]$ , or that the cross-ratio is invariant for a linear fractional transformation.

### 3. MAPPING BY ELEMENTARY FUNCTIONS

Except for the linear fractional transformations, the elementary analytic functions define one-to-one conformal transformations only in suitably restricted regions.

**The Function  $w = z^2$ .** This maps both  $z$  and  $-z$  on the same  $w$ , and hence it is one-to-one only if restricted to a region  $D$  which contains no pair  $z, -z$ . For example,  $D$  can be chosen as the upper half-plane  $\text{Im}(z) > 0$ . The corresponding region  $D_1$  consists of the  $w$ -plane minus the ray:  $u \geq 0, v = 0$ . The points  $(x, 0)$  on the boundary of  $D$  correspond to the points  $(u, 0)$  on the boundary of  $D_1$ , both  $(x, 0)$  and  $(-x, 0)$  corresponding to  $(u, 0)$ , with  $u = x^2$ . It should be noted that  $F'(z) = 2z$  is 0 at  $z = 0$ , so that this point is critical; conformality fails here, and in fact the edges of  $D$ , forming a  $180^\circ$  angle at  $z = 0$ , are transformed onto overlapping edges of  $D_1$  which form a  $360^\circ$  angle.

For  $w = z^2$  one can also choose  $D$  as a sector  $\alpha < \arg z < \beta$ , provided  $\beta - \alpha < \pi$ ; the region  $D_1$  is the sector:  $2\alpha < \arg w < 2\beta$ . A third choice of  $D$  is a hyperbolic region:  $xy > 1, x > 0$ ;  $D_1$  is then a half-plane,  $v > \frac{1}{2}$ . A fourth choice of  $D$  is a strip:  $a < x < b$ , where  $a > 0$ ;  $D_1$  is then a region bounded by two parabolas:  $4a^2u + v^2 = 4a^4, 4b^2u + v^2 = 4b^4$ .

**The Function  $w = z^n$ .** Analogous choices of regions can be made for  $w = z^n$  ( $n = 2, 3, 4, \dots$ ). The sector  $D: \alpha < \arg z < \beta$ , with  $\beta - \alpha < 2\pi/n$ , corresponds to the sector  $D_1: n\alpha < \arg w < n\beta$ . If  $n$  is allowed to be fractional or irrational,  $w = z^n$  becomes a multiple-valued analytic function (Chap. 7, Sects. 6 and 7) and one must select analytic branches. For such a branch the mapping of sectors is similar to that when  $n$  is an integer.

**The General Polynomial  $w = a_0z^n + \dots + a_{n-1}z + a_n$ .** Suitable regions can be obtained by means of the level curves of  $u = \text{Re}(w)$  and  $v = \text{Im}(w)$ . In particular the level curves of  $u$  and  $v$  which pass through the critical points of  $F(z)$  divide the  $z$ -plane into open regions each of which is mapped in one-to-one fashion on a region of the  $w$ -plane. This is illustrated in Fig. 5 for  $w = z^3 - 3z + 3$ . The critical points are at  $z = \pm 1$ , at which  $v = 0$ . The level curve  $v = 0$  divides the  $z$ -plane into six regions, in each of which  $w = F(z)$  describes a one-to-one conformal mapping of the region onto a half-plane. Adjacent regions, such as I and IV, can be merged along their common boundary to yield a region mapped by  $w = F(z)$  on the  $w$ -plane minus a single line.

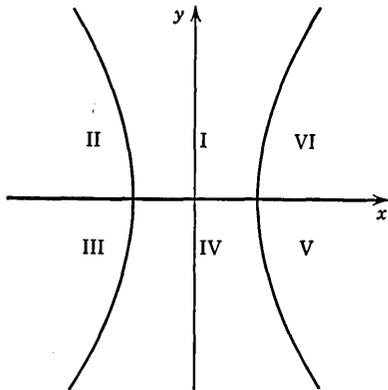


FIG. 5. Mapping by  $w = z^3 - 3z + 3$ .

**The Exponential Function  $w = e^z$ .** This maps each infinite strip  $a < y < b$  conformally onto a sector  $a < \arg w < b$ , provided  $b - a \leq 2\pi$ ; in particular each rectangle:  $c < x < d, a < y < b$  in the strip corresponds

to the part of the sector lying between the circles  $|w| = e^a$  and  $|w| = e^b$ . Similarly, the inverse of the exponential function,  $w = \log z$ , maps a sector on an infinite strip. When  $b - a = \pi/2$ , the sector is a *quadrant*; when  $b - a = \pi$ , the sector is a *half-plane*.

**The Trigonometric Function  $w = \sin z$ .** This maps the infinite strip  $-\pi/2 < x < \pi/2$  on the finite  $w$ -plane minus the portion  $|\operatorname{Re}(w)| \geq 1$  of the real axis.

**The Rational Function  $w = z + (1/z) = (z^2 + 1)/z$ .** This maps the exterior of the circle  $|z| = 1$  on the  $w$ -plane minus a slit from  $-2$  to  $+2$ . The same function maps the upper half-plane  $\operatorname{Im}(z) > 0$  on the  $w$ -plane minus the portion  $|\operatorname{Re}(w)| \geq 2$  of the real axis.

Let the real constants  $h_1, \dots, h_{n+1}, x_1, \dots, x_n$  satisfy the conditions

$$(12) \quad \begin{aligned} h_1 < h_2 < \dots < h_m, & \quad h_m > h_{m+1} > \dots > h_{n+1}, \\ x_1 < x_2 < \dots < x_n, & \end{aligned}$$

for some  $m$ ,  $1 \leq m \leq n + 1$ . Then

$$(13) \quad f(z) = h_1 \log(z - x_1) - h_{n+1} \log(z - x_n) + \sum_{k=2}^n h_k \log \frac{z - x_k}{z - x_{k-1}}$$

maps the half-plane  $\operatorname{Im}(z) > 0$  one-to-one conformally on a region  $D_1$  consisting of a strip between two lines  $v = \text{const.}$  minus several rays of form  $v = \text{const.}$  If the strip  $D_1$  has width  $\leq 2\pi$ , the function  $F(z) = \exp[f(z)]$  maps the upper half-plane conformally and one-to-one on a sector minus certain rays and segments on which  $\arg w = \text{const.}$  (See Chap. 7, Ref. 7, pp. 605-606.)

#### 4. SCHWARZ-CHRISTOFFEL MAPPINGS

These are defined by the equation

$$(14) \quad w = f(z) = A \int_{x_0}^z \frac{dz}{(z - x_1)^{k_1} \dots (z - x_n)^{k_n}} + B,$$

where  $A, B$  are complex constants,  $x_0, x_1, \dots, x_n, k_1, \dots, k_n$  are real constants, and  $-1 \leq k_j \leq 1$ . The function  $f(z)$  is analytic for  $\operatorname{Im}(z) > 0$ , with  $(z - x_j)^{k_j}$  interpreted as the principal value:  $\exp[k_j \log(z - x_j)]$ . Every one-to-one conformal mapping of the half-plane  $D$  onto the interior of a polygon can be represented in the form (14); this applies more generally to every one-to-one conformal mapping of the half-plane onto a simply connected region whose boundary consists of a finite number of lines, line segments, and rays.

**Polygon.** When the function maps  $D$  onto a polygon, the points  $x_1, \dots, x_n$  (and possibly  $\infty$ ) on the  $x$ -axis correspond to vertices of the polygon,

and the corresponding exterior angles are  $k_1\pi, \dots, k_n\pi$ . If there is an  $(n + 1)$ st vertex, corresponding to  $z = \infty$ , then necessarily  $k_1 + \dots + k_n \neq 2$ ; in general,  $1 < k_1 + \dots + k_n < 3$ .

**Convex Polygon.** When the function (14) maps  $D$  onto a convex polygon, all exterior angles are between 0 and  $\pi$  and the sum of the exterior angles is  $2\pi$ ; accordingly,

$$(15) \quad 0 < k_j < 1 \quad \text{and} \quad k_1 + \dots + k_n \leq 2.$$

When  $k_1 + \dots + k_n < 2$ , there is an  $(n + 1)$ st vertex corresponding to  $z = \infty$ . In general, for every choice of the numbers  $k_1, \dots, k_n$  such that (15) holds, eq. (14) describes a one-to-one conformal mapping of the half-plane  $\text{Im}(z) > 0$  onto the interior of a convex polygon.

**Rectangle.** For the special case

$$(16) \quad w = \int_0^z \frac{dz}{\sqrt{(1 - z^2)(1 - k^2z^2)}}, \quad 0 < k < 1,$$

the mapping is onto a rectangle with vertices  $\pm K, \pm K + iK'$ , where

$$(17) \quad K = \int_0^1 \frac{dz}{\sqrt{(1 - z^2)(1 - k^2z^2)}}, \quad iK' = \int_1^{1/k} \frac{dz}{\sqrt{(1 - z^2)(1 - k^2z^2)}}.$$

In this case  $F(z)$  is an elliptic integral of the first kind (Chap. 7, Sect. 8), and its inverse is the elliptic function  $z = \text{sn } w$ .

A great variety of conformal mappings have been studied and classified. See Ref. 1 for an extensive survey.

### 5. APPLICATION OF CONFORMAL MAPPING TO BOUNDARY VALUE PROBLEMS

The applications depend primarily on the following formal rule. If  $U(x, y)$  is given in a region  $D$  and  $w = f(z)$  is a one-to-one conformal mapping of  $D$  on a region  $D_1$ , then

$$(18) \quad \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = \left[ \frac{\partial^2 U}{\partial u^2} + \frac{\partial^2 U}{\partial v^2} \right] |f'(z)|^2.$$

In particular,  $U$  is harmonic in terms of  $x$  and  $y$ :

$$(19) \quad \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0,$$

if and only if  $U$  is harmonic when expressed in terms of  $u$  and  $v$ .

The boundary value problems considered require determination of  $U$  in  $D$  when  $U$  is required to satisfy some conditions on the boundary of  $D$  and to satisfy an equation

$$(20) \quad \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = h(x, y),$$

for given  $h(x, y)$ , in  $D$ . It follows from eq. (18) that a conformal mapping  $w = f(z)$  amounts to a change of variable reducing the problem to one of similar form in the region  $D_1$ . It is in general simpler to solve the problem for a special region such as a circle or a half-plane. Hence one tries to find a conformal mapping of  $D$  onto such a special region  $D_1$ . Once the problem has been solved for  $U$  in  $D_1$ ,  $U$  can be expressed in terms of  $(x, y)$  in  $D$  and the problem has been solved for  $D$ .

For most cases  $D$  has a boundary  $B$  consisting of a finite number of smooth closed curves  $C_1, \dots, C_n$ , the case  $n = 1$  being most common. The most important boundary value problems are then the following.

**I. Dirichlet Problem.** The values of  $U$  on  $B$  are given;  $U$  is required to be harmonic in  $D$  and to approach these values as limits as  $z$  approaches the boundary.

**II. Neumann Problem.** Again  $U$  is harmonic in  $D$  but on  $B$  the values of  $\partial U / \partial n$  are given, where  $n$  is an exterior normal vector on  $B$ .

Both problems can be generalized by requiring that  $U$  satisfy a Poisson eq. (20) in  $D$ . In general this case can be reduced to the previous one by introducing a new variable  $W$ , where

$$(21) \quad W = U + \frac{1}{4\pi} \int_D \int h(\xi, \eta) \log [(x - \xi)^2 + (y - \eta)^2] d\xi d\eta.$$

Furthermore, the Neumann problem can be reduced to the Dirichlet problem by consideration of the harmonic function  $V(x, y)$  conjugate to  $U$  (Chap. 7, Sect. 2).

To solve the Dirichlet problem for a simply connected region  $D$ , one seeks a one-to-one conformal mapping of  $D$  on the circular region  $|w| < 1$ . This reduces the problem to a Dirichlet problem for the circular region. If  $p(u, v)$  are the new boundary values, its solution is given by

$$(22) \quad U = \frac{1}{2\pi} \int_0^{2\pi} p(\cos \phi, \sin \phi) \frac{1 - r^2}{1 + r^2 - 2r \cos(\phi - \theta)} d\phi,$$

where  $r, \theta$  are polar coordinates in the  $wv$ -plane.

If  $D$  is multiply connected, it is also possible to map  $D$  conformally on a standard type of domain, for which solution of the Dirichlet problem is known. For details, see Ref. 2.

#### REFERENCES

1. H. Kober, *Dictionary of Conformal Representations*, Dover, New York, 1952.
  2. Z. Nehari, *Conformal Mapping*, McGraw-Hill, New York, 1952.
- See also the list at the end of Chap. 7.



# Boolean Algebra

A. H. Copeland, Sr.

1. Table of Notations	11-01
2. Definitions of Boolean Algebra	11-01
3. Boolean Algebra and Logic	11-05
4. Canonical Form of Boolean Functions	11-08
5. Stone Representation	11-09
6. Sheffer Stroke Operation	11-10
References	11-11

## 1. TABLE OF NOTATIONS

Table 1 lists notations in current use. There are some inconsistencies between different systems, and care is needed to ensure proper understanding. The list is not exhaustive and there are other notations even for the crucial relations; for example, “ $a$  and  $b$ ” is sometimes denoted by “ $ab$ .” The grouping under mathematics, engineering, and logic is somewhat arbitrary.

## 2. DEFINITIONS OF BOOLEAN ALGEBRA

**First Definition.** A study of the rules governing the operations on sets (Chap. 1) leads to a type of algebraic system, in which the basic operations are  $\cup$  and  $\cap$ , frequently called “or” and “and,” corresponding to union and intersection of sets. In addition, the system can be partially ordered (the relation of set inclusion) and each object of the system has a complement.

TABLE 1. TABLE OF SYMBOLS, BOOLEAN ALGEBRA

Operation Name			Symbols		
Mathematics (Set Theory)	Engineering	Logic	Mathematics (Set Theory)	Engi- neer- ing	Logic
Union	“or”	“or”	$\cup$	$+$	$\vee$
Intersection	“and”	“and”	$\cap$	$\cdot$	$\wedge$
Symmetric difference	Exclusive “or”	“or”	$\oplus$ or $+$	None	$\Delta$
Complement	Complement	Negation	$'$ or $\mathcal{C}$	$-$	$\sim$
Order	Order	Material implication	$\leq$	$\Rightarrow$	$\supset$
Sheffer stroke	Sheffer stroke	Sheffer stroke	$ $	$ $	$ $
Existential quantifier	Existential quantifier	Existential quantifier	$\exists$ or $\cup$ <small><math>\iota</math></small>	$\vee$ <small><math>\iota</math></small>	$\exists$ or $\sum$ <small><math>\iota</math>   <math>\iota</math></small>
Universal quantifier	Universal quantifier	Universal quantifier	$\cap$ <small><math>\iota</math></small>	$\wedge$ <small><math>\iota</math></small>	$\forall$ or $\prod$ <small><math>\iota</math>   <math>\iota</math></small>

A Boolean algebra  $B$  is a set of elements  $x, y, z, \dots$  with two binary operations  $\cup$  and  $\cap$ , an order relation  $\leq$ , and operation  $'$  of forming the complement such that:

- (1)  $x \cup x = x, \quad x \cap x = x,$
- (2)  $x \cup y = y \cup x, \quad x \cap y = y \cap x,$
- (3)  $x \cap (y \cap z) = (x \cap y) \cap z, \quad x \cup (y \cup z) = (x \cup y) \cup z,$
- (4)  $x \cap (y \cup z) = (x \cap y) \cup (x \cap z),$   
 $x \cup (y \cap z) = (x \cup y) \cap (x \cup z),$
- (5)  $x \leq x,$
- (6)  $x \leq y$  and  $y \leq z$  imply  $x \leq z,$
- (7)  $x \leq y$  and  $y \leq x$  imply  $x = y,$
- (8)  $B$  contains two elements  $0$  and  $1$  such that  $0 \leq x \leq 1$  for all  $x$  in  $B,$
- (9)  $0 \cap x = 0, \quad 1 \cap x = x,$
- (10)  $0 \cup x = x, \quad 1 \cup x = 1,$
- (11)  $x \cap x' = 0, \quad x \cup x' = 1,$
- (12)  $(x \cap y)' = x' \cup y', \quad (x \cup y)' = x' \cap y',$
- (13)  $(x')' = x.$

The properties (1) to (13) can be regarded as a set of postulates, from which all other properties are to be deduced. Some of the postulates are consequences of others, so that the list could be considerably reduced (Refs. 1, 3, 5).

The definition given here is easily verified to be equivalent to that given in Chap. 1, Sect. 7, in terms of lattices (Ref. 3).

**Second Definition.** An alternative definition is based upon the set operation of *symmetric difference*, also known as “exclusive or.” The symmetric difference of two sets  $X, Y$ , denoted by  $X \oplus Y$ , is the set of all elements in  $X$ , or in  $Y$ , but not in both. In symbols,

$$(14) \quad X \oplus Y = \{s | s \in X \cup Y \text{ and } s \notin X \cap Y\}.$$

This is pictured in Fig. 1.

From the definition, a number of properties can be verified. For example,  $X \oplus X = \emptyset$  (here the empty set plays the role of the 0 of a Boolean

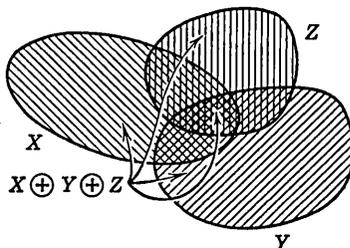
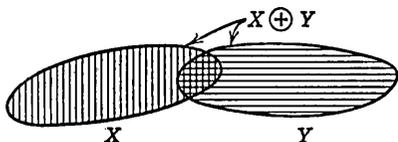


FIG. 1. Symmetric difference.      FIG. 2. Three-term symmetric difference.

algebra),  $(X \oplus Y) \oplus Z = X \oplus (Y \oplus Z)$ . The proof of the second rule is suggested in Fig. 2.

In an arbitrary Boolean algebra one can define  $x \oplus y$  in terms of the other operations

$$(15) \quad x \oplus y = (x \cap y') \cup (x' \cap y).$$

From (1),  $\dots$ , (13) and (15) a number of rules can then be deduced by algebraic means alone.

It is possible to consider  $\oplus$  and  $\cap$  as the basic operations and express  $\cup$ ,  $'$ , and  $\leq$  in terms of these two:

$$(16) \quad x \cup y = (x \oplus y) \oplus (x \cap y),$$

$$(17) \quad x' = 1 \oplus x,$$

$$(18) \quad x \leq y \text{ if } x \cap y = x.$$

Pursuing this point of view further, one is led to a second definition of a Boolean algebra.

**Alternative Definition.** A Boolean algebra  $B$  is a set of elements  $x, y, z, \dots$  with two binary operations  $\oplus, \cap$  satisfying the laws:

$$(19) \quad x \oplus y = y \oplus x, \quad x \cap y = y \cap x,$$

$$(20) \quad (x \oplus y) \oplus z = x \oplus (y \oplus z), \quad (x \cap y) \cap z = x \cap (y \cap z),$$

$$(21) \quad x \cap (y \oplus z) = (x \cap y) \oplus (x \cap z),$$

$$(22) \quad x \cap x = x,$$

$$(23) \quad B \text{ contains two elements } 0 \text{ and } 1 \text{ such that for all } x \text{ in } B, x \oplus 0 = x, \\ x \oplus x = 0 \text{ and } x \cap 1 = x.$$

If the rules (19) to (23) are regarded as postulates and the relations (16), (17), (18) as definitions of  $\cup, ', \leq$ , then one can prove all the laws (1) to (13). Conversely, from (1) to (13) and the definition (15), one can prove (19) to (23). Hence the two definitions of a Boolean algebra are equivalent.

**Relation to Set Theory.** Although Boolean algebras arise naturally in set theory, that is not the only source of such systems. They arise in *logic* and in other mathematical contexts. It is natural to ask whether every Boolean algebra can be interpreted as an algebra of all subsets of a given set. This is not true as stated, but *there is a close relationship between each Boolean algebra and an algebra of sets* (Sect. 5).

**EXAMPLE 1.** A very simple but nevertheless useful Boolean algebra is one in which  $B$  contains only 0 and 1. The properties are given in Tables 2 and 3. This Boolean algebra is used in switching circuits:  $x = 1$  means

TABLE 2.  $x \cup y$ 

		$y$	
		0	1
$x$	0	0	1
	1	1	1

TABLE 3.  $x \cap y$ 

		$y$	
		0	1
$x$	0	0	0
	1	0	1

that a certain switch is closed and  $x = 0$  means that the switch is open. Two switches in parallel correspond to  $x \cup y$ ; two switches in series correspond to  $x \cap y$ .

**EXAMPLE 2.** A somewhat more general Boolean algebra is used in the design of electronic digital computers. This can be described as follows. The elements of  $B$  are all ordered  $n$ -tuples  $x = (x_1, x_2, \dots, x_n)$ , where

each  $x_k$  is 0 or 1; if  $y = (y_1, \dots, y_n)$ , then  $x \cup y, x \cap y$  are defined as follows:

$$x \cup y = (x_1 \cup y_1, x_2 \cup y_2, \dots, x_n \cup y_n),$$

$$x \cap y = (x_1 \cap y_1, x_2 \cap y_2, \dots, x_n \cap y_n),$$

where  $x_k \cup y_k, x_k \cap y_k$  are evaluated as in Tables 1 and 2. The 0 and 1 of  $B$  are defined as follows:

$$0 = (0, 0, \dots, 0), \quad 1 = (1, 1, \dots, 1).$$

Electronic devices can be constructed to perform the Boolean operations on the  $n$ -tuples  $x$ . The operation of ordinary arithmetic can be defined in terms of the Boolean operations together with the operation of shifting the decimal point.

### 3. BOOLEAN ALGEBRA AND LOGIC

**Algebra of Sentences.** Let  $x, y, \dots$  stand for *declaratory sentences*. For *example*  $x$  might stand for "greed is evil" and  $y$  for "lead is heavy." From two sentences  $x, y$  one can form the new sentence " $x$  and  $y$ "; this is denoted by  $x \cap y$ . In the example given,  $x \cap y$  is the sentence "greed is evil and lead is heavy." From  $x$  and  $y$  one can also form the sentence " $x$  or  $y$ "; this is understood to mean:  $x$  or  $y$ , but not both; the new sentence is denoted by  $x \oplus y$ . One can also form the statement " $x$  and/or  $y$ ," meaning:  $x$  or  $y$  or both; this is denoted by  $x \cup y$ . Finally, one can form the *negation* of a sentence  $x$ : "lead is heavy" when negated becomes "lead is not heavy." The negation of  $x$  is denoted by  $x'$ .

One can now verify that in the normal logical procedures for manipulating sentences, the operations  $\cup, \cap, \oplus, '$  obey all the rules of a Boolean algebra. Two sentences are regarded as equal if they are logically equivalent. In this sense, all false sentences can be considered equal and identified with the  $\emptyset$  of the Boolean algebra; a *universal truth* ("tautology") can serve as the 1. In logic a table showing the Boolean algebra relationship of variables is called a *truth table*. The order relation  $x \leq y$  can be interpreted to mean: " $x$  implies  $y$ ." For *example*, if  $x$  is the sentence " $t$  is an even integer" and  $y$  is the sentence " $2t$  is an even integer," then  $x \leq y$ , but  $y \leq x$  is false, so that  $x < y$ . The implication defined here is essentially *strict implication* (see below) (Refs. 4, 5).

**Propositional Functions.** The sentence " $t$  is an even integer" contains a variable,  $t$ . Accordingly, the sentence can be regarded as a *function* of  $t$ . For each value of  $t$ , the function becomes a definite sentence or proposition. Hence the function is termed a propositional function. It can be denoted by  $f$ , with  $f(t)$  denoting the value for each  $t$ . For *example*,

$f(4)$  is the true sentence "4 is an even integer," while  $f(3)$  is the false sentence "3 is an even integer." The  $t$ 's for which  $f(t)$  is true form a *set*.

Similarly, the sentence " $t$  is a human being" is a propositional function which in turn determines a set; namely, the set of all human beings. If  $f, g, \dots$  are propositional functions, then one can form new propositional functions  $f \cup g, f \cap g, f \oplus g, f', \dots$  as above. If  $X_f, X_g, \dots$  are the sets corresponding to these propositional functions, then the operations on the functions correspond precisely to the operations  $\cup, \cap, \oplus, '$  on sets. For example,  $f \cap g$  is true when  $f$  and  $g$  are true; therefore an object belongs to  $X_{f \cap g}$  when it belongs to  $X_f$  and to  $X_g$ , that is, to  $X_f \cap X_g$ . Thus the calculus of propositional functions can be interpreted as a Boolean algebra of sets. The zero element represents a propositional function which is false for all values of the variable; the set 1 corresponds to a function which is true for all values.

Conversely, each set  $X$  gives rise to a propositional function:  $t$  is an element of  $X$ . This function is true precisely when  $t$  belongs to  $X$ . A Boolean algebra of sets thus leads to a Boolean algebra of propositional functions.

Because of the parallel between propositional functions and sets, one can employ geometric set diagrams, as in Figs. 1 and 2, to reason about propositional functions. In logic they are called "Venn diagrams."

**Quantifiers.** The operation of forming the intersection of many sets has an analogue for sentences or propositional functions. As for sets

(Chap. 1, Sect. 1)  $\bigcap_{i=1}^n x_i$  denotes  $x_1 \cap x_2 \cap \dots \cap x_n$ . When the  $x$ 's are

sentences, this is the new sentence: "every one of the  $x$ 's" or "for every  $t, x_t$ ." The range of  $t$  may be over an infinite set. When the range is understood, one writes simply  $\bigcap_t x_t$ . Similarly, if  $f(t)$  is a propositional function and  $t$  ranges over all values for which  $f(t)$  has meaning, then  $\bigcap_t f(t)$  is

read: "for every  $t, f(t)$ ." Alternative notations for  $\bigcap_t f(t)$  are  $\forall_t f(t)$  and

$\prod_t f(t)$ . One terms  $\bigcap_t$  a *quantifier*. There is an analogous interpretation

of  $\bigcup_t x_t$  and  $\bigcup_t f(t)$ ; the first is read "for some  $t, x_t$ " and the second "there exists a  $t$  such that  $f(t)$ ." An alternative notation for  $\bigcup_t$  is  $\exists_t$ ;  $\bigcup_t$  is also called a *quantifier*.

**Implication.** The statement " $x$  implies  $y$ " is capable of various interpretations, of which three will be discussed here: material implication, conditional implication, and strict implication. Throughout,  $x, y, \dots$  denote sentences forming a Boolean algebra  $B$ .

*Material implication.* From the sentences  $x, y$  one forms the new sentence: " $x$  implies  $y$ " as the sentence  $x' \cup y$ . This is called material impli-

cation. One often writes  $x \supset y$  or  $x \Rightarrow y$  for this implication:

$$(24) \quad x \supset y = x' \cup y = x \Rightarrow y.$$

If  $x$  and  $y$  are propositional functions  $x(t), y(t)$ , then they can be represented as sets  $X, Y$ . The sentence is then a propositional function which is true for all  $t$  if  $X' \cup Y = I$ ; that is, if  $X \subset Y$ . The notation  $x \supset y$  is therefore unfortunate. *Material implication is the basis for most mathematical arguments*, but it is criticized as permitting such statements as "if Iceland is an island, then fish can swim" to be judged true.

*Conditional implication.* For each pair of sentences  $x, y$  a new sentence  $y/x$  is formed and is read "if  $x$  then  $y$ " or " $y$  if  $x$ ." It will be assumed that  $x \neq 0$ . This is called conditional implication. The significance of the new sentence is indicated by certain postulates:

$$(25) \quad x/x = 1,$$

$$(26) \quad y/x = 0 \text{ implies } y \cap x = 0,$$

$$(27) \quad (y \cap z)/x = (y/x) \cap (z/x),$$

$$(28) \quad z/(x \cap y) = (z/x)/(y/x),$$

$$(29) \quad (1 \oplus y)/x = 1 \oplus (y/x),$$

$$(30) \quad \text{for every } x, y \text{ there is a } z \text{ such that } z/x = y, \text{ if } x \neq 0.$$

*Conditional implication is designed to fit the needs of the theory of probability.* When  $x$  is false, it may happen that  $y/x$  is true or that  $y/x$  is neither true nor false.

One can verify that postulates (25) to (28) are satisfied by material implication, but that (29), (30) are not. However, (29) is a reasonable demand to make on an implication, and it is valuable in theory of probability. Postulate (30) requires that  $B$  contain sufficiently many sentences so that one can always solve the equation  $z/x = y$  for the sentence  $z$ . A Boolean algebra which has an operation  $x/y$  satisfying postulates (25) to (30) is called an *implicative Boolean algebra*. It can be shown that an implicative Boolean algebra cannot be atomic (Sect. 4) but that one can always construct an implicative Boolean algebra containing any given Boolean algebra.

*Strict implication* is defined as follows. The strict implication  $x$  *implies*  $y$  holds if and only if the material implication is a tautology (i.e.,  $x \supset y = 1$ ) and this is true if and only if  $y/x = 1$ . When  $x$  and  $y$  are interpreted as sets then the equation  $x \supset y = 1$  can be interpreted as stating that  $x$  is contained in  $y$ . This relation has the following alternative notations:

$x \leq y, y \geq x, x \subseteq y, y \supseteq x, x \subset y, y \supset x$ . The last two notations are unfortunate since they almost reverse the interpretation of the implication symbol.

#### 4. CANONICAL FORM OF BOOLEAN FUNCTIONS

Let a Boolean algebra  $B$  be given, with operations  $\cup, \cap,$  and  $'$  as in the first definition of Sect. 1. By a *Boolean function* or *Boolean polynomial* in  $n$  variables  $x_1, \dots, x_n$  is meant an expression constructed from the  $n$  variable elements  $x_1, \dots, x_n$  by the three operations  $\cup, \cap, '.$  For example,

$$(x \cup y) \cap (x' \cup z')$$

is a Boolean polynomial in three variables. It would appear at first that such expressions can be made arbitrarily long and hence that, for fixed  $n$ , there are infinitely many polynomials. However, by the rules of the algebra, each polynomial can be simplified, and *there are precisely  $2^{2^n}$  polynomials for each  $n$ .* For example, there are four polynomials in one variable  $x$ :  $x, x', 0 = x \cap x', 1 = x \cup x'.$

If two Boolean polynomials in  $x_1, \dots, x_n$  are given, one may wish to determine whether they are the same; that is, whether one can be reduced to the other by applying the algebraic rules. In order to decide this, one reduces both polynomials to a *canonical form*, as described below. If both have the same canonical form, they are the same; otherwise, they are unequal polynomials.

**Definitions of Canonical Form and Minimal Polynomials.** By a *minimal polynomial* in  $x_1, \dots, x_n$  is meant an intersection of  $n$  letters in which the  $i$ th letter is either  $x_i$  or  $x'_i.$

EXAMPLES. There are four minimal polynomials in  $x, y$ :

$$x \cap y, \quad x' \cap y, \quad x \cap y', \quad x' \cap y'.$$

Similarly, there are eight minimal polynomials in  $x, y, z$ :

$$\begin{array}{cccc} x \cap y \cap z, & x \cap y' \cap z, & x' \cap y \cap z, & x' \cap y' \cap z \\ x \cap y \cap z', & x \cap y' \cap z', & x' \cap y \cap z', & x' \cap y' \cap z'. \end{array}$$

There are  $2^n$  such minimal polynomials in  $x_1, \dots, x_n.$

By a polynomial in *canonical form* is meant a polynomial which is either 0 or else is a union of distinct minimal polynomials. (The order of the terms can be specified, but this is of no importance since  $\cup$  is commutative.) For example,

$$(x \cap y) \cup (x' \cap y'), \quad (x \cap y) \cup (x \cap y') \cup (x' \cap y')$$

are in canonical form. Every polynomial can be written in a unique canonical form, so that equality of two polynomials holds if and only if they have the same canonical form (Ref. 3).

**Reduction to Canonical Form.** A given polynomial can be reduced to canonical form by the following steps:

- (i) Moving all primes inside parentheses by (12);
- (ii) Moving all caps ( $\cap$ 's) to the inside of parentheses by the first rule (4);
- (iii) Simplification of terms by rules (1), (2), (9), (10), (11), (13), so that one finally obtains a union of terms, each of which is a minimal polynomial in some of the  $x$ 's;
- (iv) Adjoining missing  $x$ 's to the minimal polynomials by inserting  $x \cup x' = 1$  for each such  $x$ ;
- (v) Applying steps (ii) and (iii) again.

EXAMPLE.

$$\begin{aligned}
 & [x \cap (y \cup z)] \cup [(x \cup y) \cap (y' \cup z)'] \\
 &= [(x \cap y) \cup (x \cap z)] \cup [(x \cup y) \cap (y \cap z')] \\
 &= (x \cap y) \cup (x \cap z) \cup (x \cap y \cap z') \cup (y \cap y \cap z') \\
 &= [(x \cap y) \cap (z' \cup z)] \cup [(x \cap z) \cap (y' \cup y)] \cup (x \cap y \cap z') \\
 &\quad \cup [(y \cap z') \cap (x' \cup x)] \\
 &= (x \cap y \cap z') \cup (x \cap y \cap z) \cup (x \cap y' \cap z) \cup (x \cap y \cap z) \\
 &\quad \cup (x \cap y \cap z') \cup (x' \cap y \cap z') \cup (x \cap y \cap z') \\
 &= (x \cap y \cap z') \cup (x \cap y \cap z) \cup (x \cap y' \cap z) \cup (x' \cap y \cap z')
 \end{aligned}$$

## 5. STONE REPRESENTATION

Let a Boolean algebra  $B$  be given. Then it is possible to find a set  $S$  and to define a one-to-one correspondence between the elements  $x, y, \dots$  of  $B$  and the certain subsets  $X, Y, \dots$  of  $S$  in such a fashion that if  $x$  corresponds to  $X$  and  $y$  to  $Y$ , then  $x \cup y$  corresponds to  $X \cup Y$ ,  $x \cap y$  to  $X \cap Y$ ,  $x'$  to  $X'$ ,  $0$  to the empty subset  $\emptyset$ , and  $1$  to  $S$  itself. Thus every Boolean algebra can be represented as (is *isomorphic* to) the Boolean algebra of certain subsets of a set  $S$ . This is the Stone representation.

If  $B$  has only a finite number  $m$  of elements, then  $B$  can always be represented as the Boolean algebra of *all* subsets of a given set  $S$ . Furthermore,  $m$  must be of the form  $2^n$ , where  $n$  is the number of elements in  $S$ . If  $B_1$  and  $B_2$  are Boolean algebras both having  $m$  elements, where  $m$  is finite, then  $B_1$  and  $B_2$  are isomorphic.

**STONE REPRESENTATION THEOREM.** *An infinite Boolean algebra  $B$  can be represented as the Boolean algebra of all subsets of a set  $S$  if and only if  $B$  is atomic, complete, and distributive. These properties are defined as follows:*

An element  $a$  of  $B$  is called an *atom* if the intersection  $x \cap a$  of  $a$  with an arbitrary element  $x$  of  $B$  is either  $a$  or  $0$ . If, for each  $x$  other than  $0$  in  $B$ , there is an atom  $a$  such that  $x \cap a = a$ , then  $B$  is said to be *atomic*. In the representation of  $B$  as a class of sets, the atoms correspond to sets each containing one point.

A Boolean algebra  $B$  is said to be *complete* if every subset  $A$  of  $B$  has a least upper bound (Chap. 1, Sect. 7). The least upper bound is then unique; it can be denoted by

$$\bigcup(A) \quad \text{or} \quad \bigcup_{\alpha \in A} \alpha,$$

and is also called the *union* of  $A$ .

A Boolean algebra  $B$  is said to be *distributive* if, whenever  $\bigcup(A)$  exists,

$$(31) \quad \beta \cap \bigcup(A) = \bigcup_{\alpha \in A} (\beta \cap \alpha)$$

for every  $\beta$  in  $B$ .

## 6. SHEFFER STROKE OPERATION

In a Boolean algebra  $B$  let

$$(32) \quad x|y = x' \cup y'.$$

If  $x$  and  $y$  are sentences,  $x|y$  is the sentence "either not  $x$  or not  $y$ ." One can then prove that

$$(33) \quad x|x = x' = 1 \oplus x,$$

$$(34) \quad (x|y)|(x|y) = x \cap y,$$

$$(35) \quad (x|x)|(y|y) = x \cup y,$$

$$(36) \quad x|(y|y) = x' \cup y,$$

$$(37) \quad x|(x|x) = 1.$$

*Accordingly, all the operations of the Boolean algebra can be expressed in terms of the Sheffer stroke operation. This proves to be of value in the design of electronic digital computing machines, which compute in the scale of two (see Ref. 6).*

REFERENCES

1. *Digital Computers and Data Processing*, J. W. Carr and N. R. Scott, Editors, University of Michigan, Ann Arbor, 1955, especially Article III. 4.1.
2. *High-Speed Computing Devices*, Engineering Research Associates, McGraw-Hill, New York, 1950.
3. G. Birkhoff, *Lattice Theory*, American Mathematical Society, New York, 1940.
4. I. M. Copi, *Symbolic Logic*, Macmillan, New York, 1951.
5. P. C. Rosenbloom, *The Elements of Mathematical Logic*, Dover, New York, 1950.
6. M. Phister, Jr., *Logical Design of Digital Computers*, Wiley, New York, 1958.



## Probability

A. H. Copeland, Sr.

1. Fundamental Concepts and Related Probabilities	12-01
2. Random Variables and Distribution Functions	12-04
3. Expected Value	12-06
4. Variance	12-11
5. Central Limit Theorem	12-13
6. Random Processes	12-18
References	12-20

### 1. FUNDAMENTAL CONCEPTS AND RELATED PROBABILITIES

**Postulates.** The *probability* that an event will occur is a real number between 0 and 1. If  $x$  denotes the sentence, the event will occur, then  $Pr(x)$  denotes the probability that *the event will occur*. Thus  $Pr(x)$  is the probability associated with the sentence  $x$ . Consider a Boolean algebra  $B$  of sentences (see Chap. 11) in which 0 is interpreted as the sentence associated with an impossible event and 1 is interpreted as the sentence associated with a certain event. This treatment will (a) show how some probabilities can be computed from others; (b) study the relations between probabilities of sentences connected by the words *and*, *or*, *not*, *if* (denoted respectively by  $\cap$ ,  $\cup$ ,  $'$ ,  $/$ ).

Assume that the following *postulates* hold:

(1)  $Pr(x)$  is a non-negative real number if  $x$  is in  $B$ .

(2) If  $x_1, x_2, \dots$ , are in  $B$  and  $x_i \cap x_j = 0$  when  $i \neq j$  where  $i, j = 1, 2,$

$\dots$  then  $\bigcup_{k=1}^{\infty} x_k$  is in  $B$  and

$$Pr\left(\bigcup_{k=1}^{\infty} x_k\right) = \sum_{k=1}^{\infty} Pr(x_k).$$

(3)  $Pr(1) = 1$ .

(4)  $Pr(x \cap y) = Pr(x) Pr(y/x)$ .

If  $x_i \cap x_j = 0$ , i.e., if  $x_i, x_j$  cannot both occur, then  $x_i, x_j$  are said to be *mutually exclusive* and the events associated with them are also said to be mutually exclusive. Thus postulate (2) states that the probability that at least one of a set mutually exclusive events will occur is the sum of their probabilities. The following theorems are consequences of the above postulates.

THEOREM 1.  $0 \leq Pr(x) \leq 1$ .

THEOREM 2.  $Pr(0) = 0$ .

THEOREM 3.  $Pr\left(\bigcup_{k=1}^n x_k\right) = \sum_{k=1}^n Pr(x_k)$  if  $x_i \cap x_j = 0$  whenever  $i \neq j$ .

THEOREM 4.  $Pr(x \cup y) = Pr(x) + Pr(y) - Pr(x \cap y)$ .

THEOREM 5.  $Pr(x') = 1 - Pr(x)$ .

THEOREM 6.  $Pr(x \cap y') = Pr(x) - Pr(x \cap y)$ .

THEOREM 7. If  $x_1, x_2, \dots, x_n$  are *mutually exclusive* (i.e.,  $x_i \cap x_j = 0$  when  $i \neq j$ ) and *exhaustive* (i.e.,  $x_1 \cup x_2 \cup \dots \cup x_n = 1$ ) and *equally likely* (i.e., all  $Pr(x_k)$  are equal) then  $Pr(x_k) = 1/n$  for  $k = 1, 2, \dots, n$ .

EXAMPLE 1. As an illustration of Theorem 7 consider a coin which is about to be tossed, and let  $x_1$  be the sentence "the coin will turn up heads" and  $x_2$  be the sentence "the coin will turn up tails." If the coin is not loaded, one says that it is *honest* and assumes that the hypotheses of Theorem 7 hold. Then  $Pr(x_1) = Pr(x_2) = \frac{1}{2}$ .

EXAMPLE 2. Next consider an honest die which is about to be thrown and let  $x_k$  be the sentence "the face numbered  $k$  will turn up" where  $k = 1, 2, \dots, 6$ . Again one assumes that the hypotheses of Theorem 7 hold and concludes that  $Pr(x_k) = \frac{1}{6}$  for  $k = 1, 2, \dots, 6$ . The probability that the die will turn up an odd number is given by Theorem 3. Thus

$$Pr(x_1 \cup x_3 \cup x_5) = Pr(x_1) + Pr(x_3) + Pr(x_5) = \frac{3}{6} = \frac{1}{2}.$$

EXAMPLE 3. Next let  $x = x_1 \cup x_3 \cup x_5$ ,  $y = x_1 \cup x_2 \cup x_3$  and note that  $Pr(x)$  is the probability that the die will turn up an odd number and  $Pr(y)$  is the probability that it will turn up a number less than 4. It will

be instructive for the reader to check that

$$\begin{aligned}x' &= x_2 \cup x_4 \cup x_6, & y' &= x_4 \cup x_5 \cup x_6 \\x \cap y &= x_1 \cup x_3, & x \cup y &= x_1 \cup x_2 \cup x_3 \cup x_5\end{aligned}$$

and also to check Theorems 4, 5, and 6 for this  $x$  and  $y$ . To compute the *conditional probability*  $Pr(y/x)$ , i.e., the probability that the die will turn a number less than 4 if it turns up an odd number, use postulate (4). Thus

$$Pr(x \cap y) = \frac{1}{3} = Pr(x)Pr(y/x) = \frac{1}{2}Pr(y/x),$$

and hence

$$Pr(y/x) = \frac{2}{3}.$$

EXAMPLE 4. Next consider three boxes and let  $x_k$  denote the sentence "the  $k$ th box will be selected" where  $k = 1, 2, 3$ . If one of the boxes is selected at random, this is interpreted to mean that the hypotheses of Theorem 7 hold and hence that

$$Pr(x_1) = Pr(x_2) = Pr(x_3) = \frac{1}{3}.$$

Suppose further that the first box contains two silver coins, the second contains one silver coin and one gold coin, the third contains two gold coins, and that a coin is drawn at random from the box which has been selected. Let  $y$  denote the sentence "a gold coin will be drawn from the box which has been selected." Then

$$Pr(y/x_1) = 0, \quad Pr(y/x_2) = \frac{1}{2}, \quad Pr(y/x_3) = 1.$$

Now suppose that this experiment has been performed and that the coin has been examined and found to be gold. On the basis of this information what is the probability that the coin came from the third box containing the two gold coins? One interprets the answer to this question as the conditional probability  $Pr(x_3/y)$ , i.e., the probability that the third box was drawn if the coin was observed to be gold. It will be instructive for the reader to verify that  $Pr(x_3/y) = \frac{2}{3}$  with the aid of the following theorem which is called *Bayes's theorem* and which is a consequence of the above postulates.

THEOREM 8. BAYES'S THEOREM. *If  $x_1, x_2, \dots, x_n$  are mutually exclusive, exhaustive and distinct from 0, then for any  $y$  one has*

$$Pr(x_i/y) = Pr(x_i)Pr(y/x_i) / \sum_{k=1}^n Pr(x_k)Pr(y/x_k)$$

*if the denominator is not 0.*

**Independence.** The sentences  $x_1, x_2, \dots, x_n$  are said to be *independent* if

$$Pr\left(\bigcap_{k=1}^n x_k\right) = Pr(x_1) \cdots Pr(x_n)$$

and if a similar equation holds for every subset of  $x_1, x_2, \dots, x_n$ .

Thus when  $n = 3$  one has

$$Pr(x_1 \cap x_2 \cap x_3) = Pr(x_1)Pr(x_2)Pr(x_3),$$

$$Pr(x_1 \cap x_2) = Pr(x_1)Pr(x_2),$$

$$Pr(x_2 \cap x_3) = Pr(x_2)Pr(x_3),$$

$$Pr(x_1 \cap x_3) = Pr(x_1)Pr(x_3).$$

If  $x_1, x_2$  are independent and  $Pr(x_1) \neq 0, Pr(x_2) \neq 0$  then

$$\begin{aligned} Pr(x_1 \cap x_2) &= Pr(x_1)Pr(x_2) \\ &= Pr(x_1)Pr(x_2/x_1) \\ &= Pr(x_2)Pr(x_1/x_2), \end{aligned}$$

and hence

$$Pr(x_2/x_1) = Pr(x_2) \quad \text{and} \quad Pr(x_1/x_2) = Pr(x_1).$$

## 2. RANDOM VARIABLES AND DISTRIBUTION FUNCTIONS

Consider a physical experiment which is designed to result in a real number. This number is subject to certain random fluctuations since in all physical experiments one expects experimental errors to be present. The result of the experiment is interpreted as a *random variable*  $X$ . For a mathematical definition of a random variable, see below. Let  $x_\lambda$  (for any real number  $\lambda$ ) denote the sentence *the experiment will produce a number less than  $\lambda$* , i.e., the sentence  $X$  is less than  $\lambda$ . Then the probability  $Pr(x_\lambda)$  is a function  $F$  of the real variable  $\lambda$  called the *distribution function* of  $X$ . Thus  $Pr(x_\lambda) = F(\lambda)$ .

If  $\lambda_1 < \lambda_2$  then

$$Pr(x_{\lambda_2} \cap x'_{\lambda_1}) = Pr(x_{\lambda_2}) - Pr(x_{\lambda_1} \cap x_{\lambda_2}) = F(\lambda_2) - F(\lambda_1)$$

is the probability that  $X$  is greater than or equal to  $\lambda_1$ , but less than  $\lambda_2$ . Thus when  $F$  is known, one can find the probability that  $X$  lies in a given interval.

In Chap. 11 it was noted that the elements of a Boolean algebra can be interpreted as sets of points of some space. Thus one interprets  $x_{\lambda_1} \cap x'_{\lambda_2}$  as a set and  $Pr(x_{\lambda_2} \cap x'_{\lambda_1})$  as the probability of obtaining a point of this set, that is, the probability that the experiment will select a point  $\xi$  of

this set. Imagine that the number which the experiment produces is determined by the point  $\xi$  selected and hence that  $X$  is a function of  $\xi$ . Then  $x_\lambda$  is the set of all points  $\xi$  for which  $X(\xi) < \lambda$ . The only restrictions placed on the function  $X$  are that it is real valued and that each of the sets  $x_\lambda$  shall belong to  $B$ . Such a function is said to be measurable with respect to  $B$ . The *measure* of a set  $x_\lambda$  is defined as the probability  $Pr(x_\lambda)$ . A *random variable*  $X$  is a function which is measurable with respect to  $B$ .

Let  $X$  be a random variable, let  $\bar{x}_\lambda$  be the set of points  $\xi$  for which  $X(\xi) \leq \lambda$ , and denote  $Pr(\bar{x}_\lambda)$  by  $F(\lambda+)$ . Then it can be proved (using postulate 2) that  $\bar{x}_\lambda$  is in  $B$  for all real  $\lambda$ . Moreover  $F(\lambda+)$  is the limit of  $F(\mu)$  as  $\mu$  approaches  $\lambda$  through values greater than  $\lambda$ . If  $\mu$  approaches  $\lambda$  through values less than  $\lambda$  then the limit of  $F(\mu)$  is  $F(\lambda)$ . Furthermore,  $F$  is a nondecreasing function for which

$$\lim_{\lambda \rightarrow -\infty} F(\lambda) = F(-\infty) = 0, \quad \lim_{\lambda \rightarrow +\infty} F(\lambda) = F(+\infty) = 1.$$

The above properties characterize the distribution function of a random variable.

EXAMPLE 1. As an illustration of a random variable let  $x$  be any element of  $B$  and let

$$\psi_x(\xi) = \begin{cases} 1 & \text{if } \xi \text{ is in the set } x \\ 0 & \text{if } \xi \text{ is not in the set } x. \end{cases}$$

Then  $\psi_x$  is called the *characteristic function* of the set  $x$  and is interpreted as the random variable which takes on the value 1 when  $x$  succeeds and the

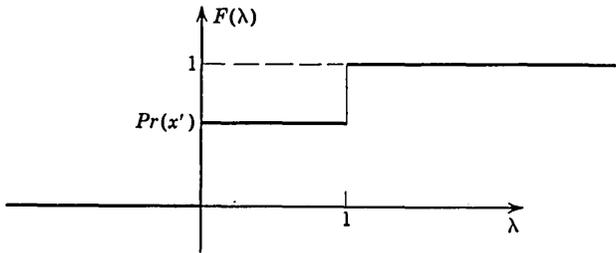


FIG. 1. Distribution function for Example 1.

value 0 when  $x$  fails. The distribution function  $F$  of the random variable  $\psi_x$  is the following (see Fig. 1):

$$F(\lambda) = \begin{cases} 0 & \text{if } \lambda \leq 0 \\ Pr(x') & \text{if } 0 < \lambda \leq 1 \\ 1 & \text{if } 1 < \lambda. \end{cases}$$

EXAMPLE 2. Consider a die and let  $x_k$  denote the sentence *the face numbered  $k$  will turn up*. Let

$$X = \psi_{x_1} + 2\psi_{x_2} + 3\psi_{x_3} + 4\psi_{x_4} + 5\psi_{x_5} + 6\psi_{x_6}.$$

If the face numbered  $k$  does turn up then this will assign the value 1 to  $\psi_{x_k}$  and the value 0 to the remaining characteristic functions and hence  $X$

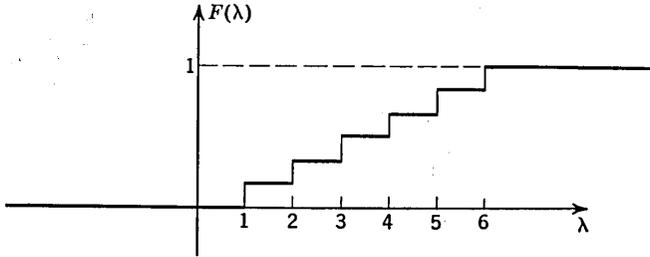


FIG. 2. Distribution function for Example 2, random tossing of a die.

will take on the value  $k$ . Thus  $X$  is the random variable which takes on the value which the die turns up (see Fig. 2).

It can be proved that *sums, products, and differences* of random variables are again random variables. Furthermore *any real number is a random variable*.

EXAMPLE 3. The number  $\sqrt{2}$  is the random variable whose distribution function  $F$  is given by (see Fig. 3):

$$F(\lambda) = \begin{cases} 0 & \text{if } \lambda < \sqrt{2} \\ 1 & \text{if } \sqrt{2} \leq \lambda. \end{cases}$$

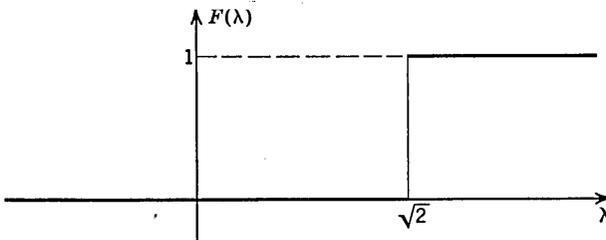


FIG. 3. Distribution function for Example 3.

### 3. EXPECTED VALUE

If  $X$  is a random variable associated with some experiment and if the experiment is repeated a large number of times, then one should expect the average of the numbers obtained to be very close to some fixed number

$E(X)$  which is called the *expected value* of  $X$ . In order to make this idea more precise the following definition is introduced. The random variables  $X_1, X_2, \dots, X_n$  are said to be *independent* provided  $x_{1,\lambda_1}, x_{2,\lambda_2}, \dots, x_{n,\lambda_n}$  are independent for all  $\lambda_1, \lambda_2, \dots, \lambda_n$  where  $x_{k,\lambda_k}$  is the set of points for which  $X_k(\xi) < \lambda_k$ .

As an illustration of independent random variables, consider a pair of honest dice. Let  $X_1$  denote the random variable which takes on the value resulting from the throw of the first die and  $X_2$  denote the random variable corresponding to the second die. It is reasonable to assume that  $X_1$  and  $X_2$  are independent. Thus we assume that the occurrence of a number less than  $\lambda_1 = 3$  on the first die and the occurrence of a number less than  $\lambda_2 = 5$  on the second die are independent events; similarly, for other choices of  $\lambda_1$  and  $\lambda_2$ . Next let  $X_3 = X_1 + X_2$ . Then  $X_1$  and  $X_3$  are dependent random variables.

**Weak Law of Large Numbers.** Now let  $X$  be an arbitrary random variable and let  $X_1, X_2, \dots, X_n$  be independent random variables all having the same distribution function as  $X$ . Let  $x_{\epsilon,n}$  be the set of points  $\xi$  for which

$$|(X_1(\xi) + X_2(\xi) + \dots + X_n(\xi))/n - E(X)| < \epsilon.$$

Then  $x_{\epsilon,n}$  is interpreted as the sentence the *average* of  $X_1, X_2, \dots, X_n$  will differ from  $E(X)$  by less than  $\epsilon$ . One might expect that

$$\lim_{n \rightarrow \infty} Pr(x_{\epsilon,n}) = 1 \text{ for every } \epsilon > 0$$

and that there is only one choice of  $E(X)$  for which this limiting probability is 1. This, as a matter of fact, is the case and this result is called the *weak law of large numbers*. Roughly the *weak law of large numbers* states that if an experiment is repeated a large number of times then it is very likely that the average of the results will differ only slightly from the expected value. The expected value  $E(X)$  exists for a large class of random variables but not for all random variables.

**Properties of Expected Value.**

**THEOREM 9.**  $E(\lambda X + \mu Y) = \lambda E(X) + \mu E(Y)$  if  $\lambda, \mu$  are real numbers and  $X, Y$  are random variables for which  $E(X), E(Y)$  exist.

**THEOREM 10.** If  $E(X), E(Y)$  exist and  $X(\xi) \leq Y(\xi)$  for all  $\xi$  then  $E(X) \leq E(Y)$ .

**THEOREM 11.** If  $\psi_x$  is the characteristic function of the set  $x$  then  $E(\psi_x) = Pr(x)$ .

With the aid of Theorems 9 and 11 one can compute expected value for certain random variables called *simple random variables*.

A random variable  $X$  is *simple* if it has the form

$$X = \sum_{k=1}^n \lambda_k \psi_{x_k}$$

where each  $\lambda_k$  is a real number and each  $\psi_{x_k}$  is the characteristic function of the set  $x_k$ .

**THEOREM 12.**

$$E(X) = E\left(\sum_{k=1}^n \lambda_k \psi_{x_k}\right) = \sum_{k=1}^n \lambda_k Pr(x_k).$$

Theorem 10 is used to approximate expected value for a much larger class of random variables called *bounded random variables*. The real numbers  $\lambda$ ,  $\mu$  are called *bounds* for a random variable  $X$  if  $\lambda \leq X(\xi) \leq \mu$  for all  $\xi$ . When the bounds exist  $X$  is said to be *bounded*.

**THEOREM 13.** *If  $\lambda$ ,  $\mu$  are bounds for  $X$  and if  $\lambda = \lambda_0 < \lambda_1 < \dots < \lambda_n = \mu$  then  $E(X)$  exists and*

$$\sum_{k=1}^n \lambda_{k-1}(F(\lambda_k) - F(\lambda_{k-1})) \leq E(X) \leq \sum_{k=1}^n \lambda_k(F(\lambda_k) - F(\lambda_{k-1}))$$

where  $F$  is the distribution function of  $X$ . If each  $\lambda_k - \lambda_{k-1} < \epsilon$ , then the extreme members of the inequalities differ by at most  $\epsilon$ .

Theorem 13 is readily established as follows. Let

$$\phi_k = \psi_{x_{\lambda_k}} \cap x'_{\lambda_{k-1}}$$

then

$$\sum_{k=1}^n \phi_k = 1, \quad \sum_{k=1}^n X\phi_k = X,$$

$$\lambda_{k-1}\phi_k \leq X\phi_k < \lambda_k\phi_k$$

and

$$E(\phi_k) = Pr(x_{\lambda_k} \cap x'_{\lambda_{k-1}}) = F(\lambda_k) - F(\lambda_{k-1})$$

and hence the inequalities follow from Theorems 9 to 11. The difference between the extreme members of the inequalities is:

$$\begin{aligned} \sum_{k=1}^n (\lambda_k - \lambda_{k-1})(F(\lambda_k) - F(\lambda_{k-1})) &< \sum_{k=1}^n \epsilon(F(\lambda_k) - F(\lambda_{k-1})) \\ &= \epsilon(F(\mu) - F(\lambda)) = \epsilon. \end{aligned}$$

If  $F$  has a continuous derivative  $f$  (i.e.,  $dF(\lambda)/d\lambda = f(\lambda)$ ) then

$$F(\lambda_k) - F(\lambda_{k-1}) = f(\mu_k)(\lambda_k - \lambda_{k-1})$$

where  $\lambda_{k-1} < \mu_k < \lambda_k$  and

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} \sum_{k=1}^n \lambda_k (F(\lambda_k) - F(\lambda_{k-1})) &= \lim_{\epsilon \rightarrow 0} \sum_{k=1}^n \lambda_k f(\mu_k) (\lambda_k - \lambda_{k-1}) \\ &= \int_{\lambda}^{\mu} u f(u) du. \end{aligned}$$

Thus

**THEOREM 14.** *If the distribution function  $F$  of a random variable  $X$  has a continuous derivative  $f$  and  $\lambda, \mu$  are bounds of  $X$  then  $E(X)$  exists and*

$$E(X) = \int_{\lambda}^{\mu} u f(u) du = \int_{\lambda}^{\mu} u dF(u).$$

**THEOREM 15.** *If the distribution function  $F$  of a random variable  $X$  has a derivative  $f$  then  $E(X)$  exists and*

$$E(X) = \int_{-\infty}^{\infty} u f(u) du = \int_{-\infty}^{\infty} u dF(u)$$

*whenever the integral exists.*

**Stieltjes and Lebesgue Integrals.** The two cases which arise most frequently in practice are the simple random variables and the random variables whose distribution functions have continuous derivatives. In the first case the expected value is computed by means of Theorem 12 and in the second case by Theorem 15. The integral on the right of the equation of Theorem 15 can be assigned a meaning even when  $f$  does not exist. In the case of a bounded variable this integral is defined to be the limit of the approximations given in Theorem 13. A meaning can also be assigned in certain unbounded cases. This integral is called a *Stieltjes integral*. Another integral expression for  $E(X)$  is

$$E(X) = \int X dPr.$$

This is called a *Lebesgue integral* and it is also defined in terms of the approximations of Theorem 13.

The terms *expectation* and *mean* are often used as synonyms for expected value.

**Probability Density and Joint Distribution.** The derivative  $f$  of  $F$  is called the *probability density*. When the density is given the distribu-

tion function can be computed by the formula

$$F(\lambda) = \int_{-\infty}^{\lambda} f(u) du.$$

See Figs. 4 and 5, Sect. 5.

The *joint distribution* of two random variables  $X_1, X_2$  is a function  $F$  such that  $F(\lambda_1, \lambda_2)$  is the probability that  $X_1 < \lambda_1$ , and  $X_2 < \lambda_2$ . If the joint distribution has a density  $f$  then

$$F(\lambda_1, \lambda_2) = \int_{-\infty}^{\lambda_1} \int_{-\infty}^{\lambda_2} f(u_1, u_2) du_1 du_2.$$

if  $F_1, F_2$  are the distribution functions of  $X_1, X_2$ , and  $f_1, f_2$  are the corresponding densities then

$$\begin{aligned} F_1(\lambda_1) &= \int_{-\infty}^{\lambda_1} \int_{-\infty}^{\infty} f(u_1, u_2) du_1 du_2 = \int_{-\infty}^{\lambda_1} f_1(u_1) du_1, \\ F_2(\lambda_2) &= \int_{-\infty}^{\lambda_2} \int_{-\infty}^{\infty} f(u_1, u_2) du_2 du_1 = \int_{-\infty}^{\lambda_2} f_2(u_2) du_2, \\ f_1(u_1) &= \int_{-\infty}^{\infty} f(u_1, u_2) du_2, \quad f_2(u_2) = \int_{-\infty}^{\infty} f(u_1, u_2) du_1. \end{aligned}$$

The expected value of the product  $X_1X_2$  is

$$E(X_1X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u_1u_2f(u_1, u_2) du_1 du_2.$$

If  $X_1, X_2$  are independent then

$$F(\lambda_1, \lambda_2) = F(\lambda_1)F(\lambda_2)$$

and

$$f(u_1, u_2) = f(u_1)f(u_2).$$

Furthermore:

**THEOREM 16.** *If  $X_1, X_2$  are independent then  $E(X_1X_2) = E(X_1)E(X_2)$ . Two random variables  $X_1, X_2$  for which  $E(X_1X_2) = E(X_1)E(X_2)$  are said to be *uncorrelated*. Thus Theorem 16 states that *independent random variables are uncorrelated*. This result holds even when there is no joint probability density. The converse is not true. That is, random variables may be uncorrelated, but not independent.*

As an illustration of a pair of random variables which are dependent but uncorrelated, consider an honest die whose faces are numbered respectively  $-3, -2, -1, 1, 2, 3$ . Let  $X_1$  denote the random variable which takes on the value resulting from the throw of this die and let  $X_2 = X_1^2$ . Then

$$E(X_1) = 0, \quad E(X_2) = \frac{14}{3},$$

$$E(X_1X_2) = E(X_1^3) = 0 = E(X_1)E(X_2)$$

Hence  $X_1$  and  $X_2$  are uncorrelated but they are clearly dependent.

4. VARIANCE

**THEOREM 17.** *If  $f$  is a function of a real variable  $\lambda$  with at most a finite number of discontinuities and if  $X$  is a random variable with distribution function  $F$ , then  $f(X)$  is a random variable and*

$$E(f(X)) = \int_{-\infty}^{\infty} f(u) dF(u)$$

whenever the integral exists.

A special case of this formula is the following:

If

$$E(X) = \mu$$

and

$$E((X - \mu)^2) = \int_{-\infty}^{\infty} (u - \mu)^2 dF(u) = E(X^2) - E^2(X) = \sigma^2(X),$$

then  $\sigma^2(X)$  is called the *variance* of  $X$  and the positive square root of the variance,  $\sigma(X)$ , is called the *standard deviation* of  $X$ .

**The Properties of Variance.**

**THEOREM 18.**  $\sigma^2(X + \lambda) = \sigma^2(X)$ ,  $\sigma^2(\lambda X) = \lambda^2\sigma^2(X)$ .

**THEOREM 19.** *If  $X_1, X_2, \dots, X_n$  are independent random variables, then*

$$\sigma^2(X_1 + X_2 + \dots + X_n) = \sigma^2(X_1) + \sigma^2(X_2) + \dots + \sigma^2(X_n),$$

$$\sigma^2\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n} \frac{\sigma^2(X_1) + \sigma^2(X_2) + \dots + \sigma^2(X_n)}{n}.$$

If  $x_\epsilon$  is the set of all points  $\xi$  such that  $|X(\xi) - \mu| < \epsilon$  where  $\mu = E(X)$  then  $x'_\epsilon$  is the set of all  $\xi$  such that  $|X(\xi) - \mu| \geq \epsilon$ . Moreover the inequality

$$\epsilon^2 \psi_{x'_\epsilon}(\xi) \leq (X(\xi) - \mu)^2$$

can readily be verified when  $\xi$  is in  $x_\epsilon$  and when  $\xi$  is in  $x'_\epsilon$  and hence this inequality holds for all  $\xi$ . Thus by Theorems 9 to 11 it follows that

$$E(\epsilon^2 \psi_{x'_\epsilon}) = \epsilon^2 Pr(x'_\epsilon) \leq E((X - \mu)^2) = \sigma^2(X),$$

and therefore

$$Pr(x_\epsilon) \geq 1 - \sigma^2(X)/\epsilon^2.$$

This inequality is called *Tchebysheff's inequality*. By combining Tchebysheff's inequality with Theorem 19, one obtains:

**THEOREM 20.** *If  $X_1, X_2, \dots, X_n$  are independent random variables with common mean  $\mu$  and common variance  $\sigma^2$  and if  $x_{\epsilon,n}$  is the set of points  $\xi$  for which*

$$|(X_1(\xi) + X_2(\xi) + \dots + X_n(\xi))/n - \mu| < \epsilon$$

then

$$Pr(x_{\epsilon,n}) \geq 1 - \sigma^2/n\epsilon^2$$

and

$$\lim_{n \rightarrow \infty} Pr(x_{\epsilon,n}) = 1 \text{ for every } \epsilon > 0.$$

**The Strong Law of Large Numbers.** The first part of Theorem 20 gives a crude approximation for the probability that the average will differ from the common mean by less than  $\epsilon$ . Recall that the second part of this theorem is the weak law of large numbers. The reasoning by which one arrived at this result is of course circular but this circularity can be avoided. The *strong law of large numbers* is the following:

**THEOREM 21.** *If  $X_1, X_2, \dots, X_n$  are independent random variables with common expected value  $\mu$  and common variance  $\sigma^2$  and if  $x$  is the set of points  $\xi$  for which*

$$\lim_{n \rightarrow \infty} \frac{X_1(\xi) + X_2(\xi) + \dots + X_n(\xi)}{n} = \mu.$$

Then  $Pr(x) = 1$ .

Even though  $Pr(x) = 1$  it is not in general true that  $x = 1$ . If an element  $x$  of  $B$  is such that  $Pr(x) = 1$ , then  $x$  is said to be *almost certain*. The strong law of large numbers states that it is almost certain that the limit of the average is the *common expected value*.

The following example will help one understand the distinction between certain and almost certain. Let  $X$  be a random variable with distribution function  $F$  defined as follows:

$$F(\lambda) = \begin{cases} 0 & \text{if } \lambda \leq 0 \\ \lambda & \text{if } 0 \leq \lambda \leq 1 \\ 1 & \text{if } 1 \leq \lambda. \end{cases}$$

Then it is almost certain, but not entirely certain, that  $X$  will take on a value distinct from  $\frac{1}{2}$ .

5. CENTRAL LIMIT THEOREM

**Distribution of Sums and Averages of Independent Random Variables.** Consider

$$e^{iXt} = \cos Xt + i \sin Xt$$

where  $i^2 = -1$  and  $t$  is a parameter. This exponential converts the real random variable  $X$  into a complex valued random variable. The expected value of the latter random variable is defined in a natural way to be

$$E(e^{iXt}) = E(\cos Xt) + iE(\sin Xt) = \phi_X(t).$$

The advantage of the exponential is that it converts a sum into a product and hence enables one to make use of the condition of independence. Thus if  $X, Y$  are independent then it can be shown that  $e^{iXt}, e^{iYt}$  are independent and hence by Theorem 16

$$\phi_{X+Y}(t) = E(e^{iXt}e^{iYt}) = \phi_X(t)\phi_Y(t).$$

The advantage of the factor  $i$  is that it produces a bounded random variable and insures the existence of the expected value for all real values of  $t$ . The advantage of the parameter  $t$  is that it produces a function in terms of which one can compute the distribution function. Thus  $\phi_X$  is a function of the parameter  $t$  called the *characteristic function of the random variable  $X$* . Unfortunately the phrase, characteristic function, has two distinct meanings in the theory of probability, namely, *characteristic function of a set of points* and *characteristic function of a random variable*.

**Computation of the Characteristic Function of a Simple Random Variable.** Let

$$X = \sum_{k=1}^n \lambda_k \psi_{x_k},$$

where  $x_1, x_2, \dots, x_n$  are mutually exclusive and exhaustive. Then

$$e^{iX(\xi)t} = \sum_{k=1}^n e^{i\lambda_k t} \psi_{x_k}(\xi)$$

for all  $\xi$ , since if  $\xi$  lies in  $x_k$ , then  $\psi_{x_k}(\xi) = 1$  and the remaining characteristic functions have the value 0 and hence both sides of the equation become  $e^{i\lambda_k t}$ . From this it follows that

$$E(e^{iXt}) = \sum_{k=1}^n Pr(x_k) e^{i\lambda_k t} = \phi_X(t).$$

As a special case consider the simple random variable  $\psi_x$ . One can write

$$\psi_x = 0 \cdot \psi_{x'} + 1 \cdot \psi_x$$

and hence

$$\phi_{\psi_x}(t) = q + pe^{it}$$

where  $p = Pr(x)$ ,  $q = Pr(x')$ .

Next compute the characteristic function for the sum  $\psi_{x_1} + \psi_{x_2} + \cdots + \psi_{x_n}$  where  $x_1, x_2, \cdots, x_n$  are independent and  $Pr(x_k) = p$ ,  $Pr(x'_k) = q$  for each  $k$ . Then

$$\phi_{\psi_{x_1} + \psi_{x_2} + \cdots + \psi_{x_n}}(t) = \prod_{k=1}^n \phi_{\psi_{x_k}}(t) = (q + pe^{it})^n.$$

If  $X$  has the distribution function  $F$  then the characteristic function is

$$\phi_X(t) = E(e^{iXt}) = \int_{-\infty}^{\infty} e^{i\lambda t} dF(\lambda).$$

This transforms the function  $F$  into the function  $\phi_X$  (essentially the Laplace-Fourier transform). The inverse transform is

$$\frac{1}{2}(F(\lambda) + F(\lambda+)) = \frac{1}{2\pi i} \lim_{h \rightarrow \infty} \int_{-h}^h \frac{e^{it} - \phi_X(t)e^{-\lambda t}}{t} dt.$$

To see why this is the case note that

$$\begin{aligned} \frac{1}{2\pi i} \lim_{h \rightarrow \infty} \int_{-h}^h \frac{e^{it} - e^{i\mu t}e^{-i\lambda t}}{t} dt &= \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{e^{it} - e^{i\mu t}e^{-i\lambda t}}{t} dt \\ &= \begin{cases} 1 & \text{if } \mu < \lambda, \\ \frac{1}{2} & \text{if } \mu = \lambda, \\ 0 & \text{if } \lambda < \mu. \end{cases} \end{aligned}$$

This formula is verified by converting the integral into integrals of the form

$$\int_{-\infty}^{\infty} \frac{\sin mt}{t} dt$$

by means of the relation  $e^{imt} = \cos mt + i \sin mt$ . Now compute the inverse transform for a simple random variable

$$X = \sum_{k=1}^n \lambda_k \psi_{x_k},$$

where  $x_1, x_2, \cdots, x_n$  are mutually exclusive and exhaustive. Since

$$\sum_{k=1}^n Pr(x_k) = 1,$$

then

$$\begin{aligned} \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{e^{it} - \phi_X(t)e^{-i\lambda t}}{t} dt &= \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{\sum_{k=1}^n Pr(x_k)e^{it} - \sum_{k=1}^n Pr(x_k)e^{i\lambda_k t}e^{-i\lambda t}}{t} dt \\ &= \sum_{k=1}^n \frac{Pr(x_k)}{2\pi i} \int_{-\infty}^{\infty} \frac{e^{it} - e^{i\lambda_k t}e^{-i\lambda t}}{t} dt \\ &= \sum_{\lambda_k < \lambda} Pr(x_k), \end{aligned}$$

if  $\lambda \neq$  any  $\lambda_k$ . The final sum is the probability that  $X$  will take on a value less than  $\lambda$  and hence this sum is equal to

$$F(\lambda) = F(\lambda+) = \frac{1}{2}(F(\lambda) + F(\lambda+)).$$

If  $\lambda$  equals some  $\lambda_k$ , then the corresponding term  $Pr(x_k)/2$  must be added and again the result is  $\frac{1}{2}(F(\lambda) + F(\lambda+))$ . The proof for an arbitrary random variable  $X$  consists in approximating  $X$  by a simple random variable. In the general case the integral from  $-\infty$  to  $\infty$  may not exist, and one has to resort to integrating from  $-h$  to  $h$  and then passing to the limit.

**Binomial and Poisson Distributions.** If  $F_n$  is the distribution function of

$$X = \psi_{x_1} + \psi_{x_2} + \cdots + \psi_{x_n}$$

where  $x_1, x_2, \dots, x_n$  are independent and  $Pr(x_k) = p, Pr(x'_k) = q$  for each  $k$  then  $F_n(\lambda)$  is the probability that less than  $\lambda$  of the events  $x_1, \dots, x_n$  will succeed. If  $\lambda \neq$  any  $k$  then

$$\begin{aligned} F_n(\lambda) &= \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{e^{it} - \phi_X(t)e^{-i\lambda t}}{t} dt \\ &= \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{(q+p)^n e^{it} - (q+pe^{it})^n e^{-i\lambda t}}{t} dt \\ &= \sum_{k=0}^n \binom{n}{k} q^k p^{n-k} \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{e^{it} - e^{ikt}e^{-i\lambda t}}{t} dt \\ &= \sum_{k < \lambda} \binom{n}{k} p^k q^{n-k}, \end{aligned}$$

where  $\binom{n}{k}$  is the number of combinations of  $n$  things taken  $k$  at a time. If  $\lambda =$  some  $k$ , then the corresponding term  $\binom{n}{k} p^k q^{n-k}/2$  must be added.

This is called the *binomial distribution*. To obtain an approximation to this distribution for small  $p$  and large  $n$  set  $p = \mu/n$  and let  $n$  become infinite. The limiting distribution  $F$  is given by

$$F(\lambda) = \sum_{k < \lambda} \frac{\mu^k}{k!} e^{-\mu}.$$

This is called the *Poisson distribution*.

**Normal or Gaussian Distribution.** Next consider a sum  $U_n = X_1 + X_2 + \dots + X_n$  of independent random variables with a common distribution function  $F$ . Denote the common expected value by  $\mu$  and common variance by  $\sigma^2$ . If  $Y_k = (X_k - \mu)/\sigma$  and  $V_n = (U_n - n\mu)/\sigma\sqrt{n}$  then (see Theorems 18 and 19),

$$V_n = (Y_1 + Y_2 + \dots + Y_n)/\sqrt{n},$$

$$E(Y_k) = 0 = E(V_n), \quad \sigma^2(Y_k) = 1 = \sigma^2(V_n).$$

Moreover

$$\begin{aligned} \phi_{V_n}(t) &= \prod_{k=1}^n E \exp(iY_k t/\sqrt{n}) = E^n \exp(iY_1 t/\sqrt{n}) \\ &= E^n(1 + iY_n t/\sqrt{n} - Y_n^2 t^2/2n + \dots), \\ &= (1 + 0 - t^2/2n + \dots)^n, \\ &= ((1 - t^2/2n + \dots)^{-2n/t^2})^{-t^2/2}. \end{aligned}$$

Hence

$$\lim_{n \rightarrow \infty} \phi_{V_n}(t) = e^{-t^2/2}$$

and the limiting distribution is

$$\Phi(\lambda) = \frac{1}{2\pi i} \int_{-\infty}^{+\infty} \frac{e^{it} - e^{-t^2/2} e^{-i\lambda t}}{t} dt.$$

To obtain a simpler form for  $\Phi$  compute its derivative. Thus

$$\frac{d\Phi}{d\lambda}(\lambda) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-t^2/2} e^{-i\lambda t} dt.$$

Since  $t^2 + 2i\lambda t = (t + i\lambda)^2 + \lambda^2$ , set  $u = t + i\lambda$ ,  $du = dt$  and obtain

$$\frac{d\Phi}{d\lambda}(\lambda) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-(t+i\lambda)^2/2} e^{-\lambda^2/2} dt = \frac{e^{-\lambda^2/2}}{2\pi} \int_{-\infty}^{+\infty} e^{-u^2/2} du = \frac{A e^{-\lambda^2/2}}{2\pi},$$

where

$$A = \int_{-\infty}^{\infty} e^{-u^2/2} du.$$

Hence

$$\Phi(\lambda) = \frac{A}{2\pi} \int e^{-\lambda^2/2} du = \frac{A}{2\pi} \int_{-\infty}^{\lambda} e^{-u^2/2} du + C.$$

Since  $\Phi(-\infty) = 0$  this means that  $C = 0$ . Moreover

$$\Phi(+\infty) = \frac{A}{2\pi} \int_{-\infty}^{\infty} e^{-u^2/2} du = \frac{A^2}{2\pi}.$$

Therefore  $A = \sqrt{2\pi}$  and

$$\Phi(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda} e^{-u^2/2} du.$$

This limiting distribution is called the *normal distribution* or *gaussian distribution* and is shown in Fig. 4.

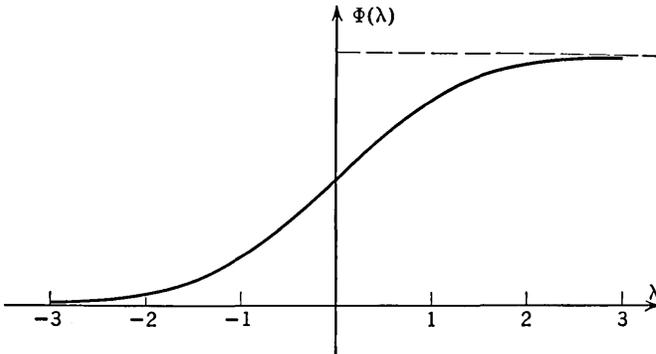


FIG. 4. Normal distribution function.

**THEOREM 22. CENTRAL LIMIT THEOREM.** *If  $X_1, X_2, \dots$  are independent random variables with a common distribution function  $F$  common expected value  $\mu$  and common variance  $\sigma^2$  then the distribution function of  $(X_1 + X_2 + \dots + X_n - n\mu)/\sigma\sqrt{n}$  converges to the normal distribution function as  $n$  becomes infinite.*

If  $s = \lambda\sigma/\sqrt{n}$ , i.e.,  $\lambda = s\sqrt{n}/\sigma$  then the probability that  $(X_1 + X_2 + \dots + X_n - n\mu)/\sigma\sqrt{n} < \lambda$  is equal to the probability that  $(X_1 + X_2 + \dots + X_n)/n - \mu < s$  and approximately equal to

$$\Phi(\lambda) = \Phi(s\sqrt{n}/\sigma).$$

Figure 5 shows the probability density corresponding to the normal distribution function.

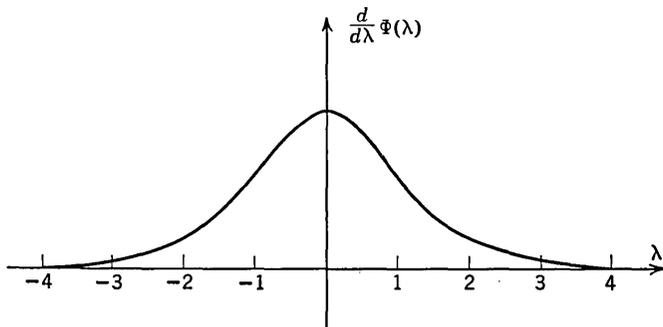


FIG. 5. Probability density for the normal distribution function.

## 6. RANDOM PROCESSES

A *continuous random process* is a function  $X$  which assigns to every real number  $t$ , a random variable  $X_t$ . If  $t$  ranges only over the integers then the process  $X$  is said to be *discrete* and if  $t$  ranges only over the positive integers then the process  $X$  is simply a *sequence* of random variables. Consider complex valued random variables, i.e., random variables of the form  $X = X_1 + iX_2$  where  $X_1, X_2$  are real. The *complex conjugate* of  $X$  is  $X_1 - iX_2$  and is denoted by  $\bar{X}$ . The *inner product* of two random variables  $X, Y$  is denoted by  $(X, Y)$  and defined by the equation

$$(X, Y) = E(X\bar{Y}).$$

The *covariance function*  $R$  of a process  $X$  is defined by the equation

$$R(t, \tau) = (X_{t+\tau}, X_t).$$

If  $R$  depends only on  $\tau$  and not on  $t$ , then the process is said to be *stationary in the wide sense*. A physical *example* of such a process is the phenomenon of *noise*. In the mathematical model (i.e., the process  $X$ ) the variable  $t$  is interpreted as time. The process can be envisioned as being composed of simple harmonic oscillations in which the amplitudes associated with the various frequencies are selected in accordance with a certain random procedure. A simple harmonic oscillation of frequency  $\lambda$  is represented by  $e^{2\pi i\lambda t}$  and the (complex) amplitude associated with the frequencies between  $\lambda$  and  $\lambda + d\lambda$  is denoted by  $dY_\lambda$ , and hence the contribution of such frequencies to the process is

$$e^{2\pi i\lambda t} dY_\lambda.$$

Here  $Y$  is a process which assigns to each real number  $\lambda$  a random variable

$Y_\lambda$ . The process  $X$  is obtained by adding the contributions associated with the various frequencies. Hence

$$X_t = \int_{-\infty}^{\infty} e^{2\pi i \lambda t} dY_\lambda.$$

Thus the *spectrum* of the process  $X$  is described by the process  $Y$ .

The expected value of the square of the amplitude associated with the frequencies between  $\lambda$  and  $\lambda + d\lambda$  is denoted by  $dF(\lambda)$  and is defined by

$$dF(\lambda) = (dY_\lambda, dY_\lambda) \geq 0.$$

Thus  $F$  is a monotone nondecreasing function. A property of the process  $Y$ , called the property of orthogonal increments, is the following

$$(dY_\lambda, dY_\mu) = 0,$$

if the intervals  $d\lambda, d\mu$  have no common points. Hence

$$\begin{aligned} R(\tau) &= (X_{t+\tau}, X_t) = \left( \int_{-\infty}^{\infty} e^{2\pi i \lambda (t+\tau)} dY_\lambda, \int_{-\infty}^{\infty} e^{2\pi i \mu t} dY_\mu \right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{2\pi i \lambda (t+\tau)} \overline{e^{2\pi i \mu t}} (dY_\lambda, dY_\mu) \\ &= \int_{-\infty}^{\infty} e^{2\pi i \lambda (t+\tau)} e^{-2\pi i \lambda t} dF(\lambda) = \int_{-\infty}^{\infty} e^{2\pi i \lambda \tau} dF(\lambda). \end{aligned}$$

As a special case of this formula

$$R(0) = \int_{-\infty}^{\infty} dF(\lambda).$$

The following example of a one-dimensional Brownian motion will aid in visualizing a random process. A tiny mirror is suspended by a fiber. Particles of air bombard the mirror and cause it to turn through an angle. A beam of light is reflected by the mirror and the position of the reflection enables the observer to measure the angle  $X(t)$  through which the mirror has turned at time  $t$ . Since  $X(t)$  is produced by the average effect of a number of bombardments, one might expect  $X(t)$  to have a normal distribution. That is, the probability that  $X(t) < \lambda$  is

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\lambda} e^{-X^2/2\sigma^2} dx,$$

where  $\sigma^2$  is the variance and is assumed to be independent of  $t$ . From this formula one can readily show that  $E(X(t)) = 0$ . The zero angle is the angle

in which the fiber is untwisted. If  $t$  and  $t + \tau$  are two times at which the mirror is observed, the joint probability that  $X(t) < \lambda_1$  and  $X(t + \tau) < \lambda_2$  is

$$\frac{1}{2\pi\sigma^2\sqrt{1-r^2(\tau)}} \int_{-\infty}^{\lambda_1} \int_{-\infty}^{\lambda_2} e^{-[x^2-2r(\tau)xy+y^2]/2\sigma^2[1-r^2(\tau)]} dx dy.$$

This is called the bivariate normal distribution. From this formula one can show that the covariance is

$$(X(t), X(t + \tau)) = \sigma^2 r(\tau),$$

and hence that the process is stationary in the wide sense. If it is known that  $X(t) = a$  then the probability that  $X(t + \tau) < \lambda$  is

$$\frac{1}{\sigma\sqrt{2\pi(1-r^2(\tau))}} \int_{-\infty}^{\lambda} e^{-[x-ar(\tau)]^2/2\sigma^2[1-r^2(\tau)]} dx.$$

Any information concerning the motion previous to time  $t$  is irrelevant to this probability. A process having this property is said to be Markovian. The assumption that the above process is Markovian implies that

$$r(\tau) = e^{-k\tau}, \text{ where } k > 0.$$

#### REFERENCES

1. H. Cramer, *The Elements of Probability Theory*, Wiley, New York, 1950.
2. J. L. Doob, *Stochastic Processes*, Wiley, New York, 1953.
3. W. Feller, *Probability Theory and Its Applications*, Wiley, New York, 1950.
4. A. N. Kolmogoroff, *Foundations of the Theory of Probability*, Chelsea, New York, 1950.
5. P. Lévy, *Théorie de l'addition des variables aléatoires*, Gautier-Villars, Paris, 1937.
6. J. V. Uspenski, *Introduction to Mathematical Probability*, McGraw-Hill, New York, 1937.
7. Ming Chen Wang and G. E. Uhlenbeck, On the theory of Brownian motion. II. *Revs. Mod. Phys.*, 17, 323-342 (1945).

## Statistics

A. B. Clarke

1. Nature of Statistics	13-01
2. Probability Background	13-02
3. Important Probability Distributions	13-04
4. Sampling	13-06
5. Bivariate Distributions	13-13
6. Tests for Goodness of Fit	13-16
7. Sequential Analysis	13-16
8. Monte Carlo Method	13-17
9. Statistical Tables	13-18
References	13-21

### 1. NATURE OF STATISTICS

The basic assumption underlying the application of the mathematical theory of probability and statistics to physical situations is the following: If a physical "experiment" is repeated under "identical" conditions and "without bias," the observed relative frequency of success of any physical "event" approaches as a limit the probability assigned to this event by some underlying probability distribution.

*Probability theory* is the study of probability distributions as mathematical entities. *Statistics* is the analysis of probability distributions on the basis of a number of experimental observations; the distribution is in general not fully known to start with, and one seeks properties of the distribution on the basis of the observations. Since an infinite number

of experiments would usually be required to determine a distribution with precision, it is only rarely possible to answer a statistical question with 100 per cent surety. Accordingly *the answer to each statistical question* should consist of two parts: (a) the best possible answer to the question and (b) the amount of confidence that can be placed in the correctness of this answer. The omission of (b) greatly diminishes the value of the conclusion.

## 2. PROBABILITY BACKGROUND

The basic probability theory required for statistics is reviewed in Chap. 12. For the sake of convenience the principal definitions are recalled here. (See Refs. 2, 6.)

**Sample Space.** The sample space  $S$  is the collection of all possible outcomes of a physical experiment; the individual outcomes are *sample points*. By an *event* is meant a certain type of outcome; in other words, a certain set  $A$  of sample points. A class  $\mathcal{G}$  of events is assumed specified. To each event  $A$  of class  $\mathcal{G}$  is assigned a *probability*,  $Pr(A)$ , which is a real number between 0 and 1. One has  $Pr(\emptyset) = 0$ ,  $Pr(S) = 1$ , and  $Pr(A \cup B) = Pr(A) + Pr(B)$ , provided  $A, B$  have no points in common (are mutually exclusive events).

A sample space  $S$  is *discrete* if its points form a finite or infinite sequence  $\xi_1, \xi_2, \dots$ . For discrete spaces a probability is usually defined for each point, and then for each subset  $A$  as the sum of the probabilities of the points in  $A$ .

**Random Variables.** A random variable is a function  $X = X(\xi)$  which assigns to each sample point  $\xi$  a real number  $x$  in such a fashion that, for each  $a$ , the set  $A$  for which  $x \leq a$  has a probability; thus  $Pr(X \leq a)$  is well defined. With each random variable  $X$  is associated a *distribution*  $F(x)$ ;  $F(a) = Pr(X \leq a)$ .  $F(x)$  is nondecreasing,  $F(-\infty) = 0$ ,  $F(+\infty) = 1$ . If  $X_1, \dots, X_n$  are random variables associated with the same experiment, then their *joint distribution* is  $F(x_1, \dots, x_n)$ , where  $F(a_1, \dots, a_n)$  is the probability assigned to the set where  $x_1 \leq a_1, \dots, x_n \leq a_n$ . The random variable  $X$  has a *density*  $f$ , if

$$(1) \quad F(x) = \int_{-\infty}^x f(t) dt;$$

the random variables  $X_1, \dots, X_n$  have a *joint density*  $f$  if

$$(2) \quad F(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_n} f(t_1, \dots, t_n) dt_n \dots dt_1.$$

When the range or collection of values of  $X$  forms a discrete sequence  $x_1, x_2, \dots$ , then

$$(3) \quad F(a) = \sum_{x_i \leq a} Pr(X = x_i) = \sum_{x_i \leq a} f(x_i),$$

where  $Pr(X = x_i) = f(x_i)$  is the probability assigned to the set of sample points for which  $X = x_i$ . This can be generalized to joint distributions.

Random variables  $X_1, \dots, X_n$  are mutually *independent* if

$$(4) \quad F(x_1, \dots, x_n) = F_1(x_1) \cdots F_n(x_n),$$

where  $F$  is the joint distribution and  $F_i(x_i)$  is the distribution of  $X_i$ .

Throughout the following it will be assumed that either the range of each random variable is discrete or else each distribution has a density (continuous case).

The *expectation* or *mean* of a random variable  $X$  is

$$(5) \quad E(X) = \begin{cases} \int_{-\infty}^{\infty} xf(x) dx & \text{(continuous case),} \\ \sum_i x_i f(x_i) & \text{(discrete case).} \end{cases}$$

If  $\phi$  is a continuous function of  $x$ , then

$$(6) \quad E(\phi(X)) = \begin{cases} \int_{-\infty}^{\infty} \phi(x)f(x) dx & \text{(continuous case),} \\ \sum_i \phi(x_i)f(x_i) & \text{(discrete case).} \end{cases}$$

**Moments.** The moments of  $X$  about the origin are the numbers

$$(7) \quad \mu'_k = E(X^k), \quad k = 1, 2, \dots.$$

The *moments of  $X$  about the mean* are defined by

$$(8) \quad \mu_k = E((X - \mu)^k), \quad k = 2, 3, \dots,$$

where  $\mu = E(X) = \mu'_1$ . The quantity  $\sigma^2 = \mu_2$  is the *variance* of  $X$ , while  $\sigma = \sqrt{\sigma^2}$  is the *standard deviation* of  $X$ .

By expanding the quantity  $(x - \mu)^k$  by the binomial formula and applying eq. (8), one obtains an expression for the  $\mu_k$  in terms of  $\mu'_1, \dots, \mu'_k$ :

$$\mu_k = \mu'_k - k\mu'_{k-1}\mu + \dots.$$

In particular,

$$(9) \quad \sigma^2 = \mu'_2 - 2\mu'_1\mu'_1 + (\mu'_1)^2 = \mu'_2 - \mu^2.$$

The *mean*  $\mu$  is a measure of the location of the "center" of the distribution, while the *variance*  $\sigma^2$  is a measure of the "spread" of the distribution.

Other possible measures of central tendency are:

*Median:* a point  $x_0$  such that  $Pr(X \leq x_0) = Pr(X \geq x_0)$ ,

*Mode:* a point  $x_0$  where  $f(x)$  is a maximum,

*Midrange:*  $\frac{1}{2}(a + b)$ , if  $a \leq x \leq b$  is the smallest interval containing all  $x$  for which  $f(x) > 0$ .

Other measures of the spread of the distribution are:

*Mean deviation from the mean* =  $E(|X - \mu|)$ ,

*Probable error:* a number  $\alpha$  such that  $Pr(|X - \mu| \leq \alpha) = \frac{1}{2}$ .

For comparison and tabulating purposes it is useful to describe a random variable in a manner independent of origin and scale. These requirements are met by the *standardized variable*  $X^* = (X - \mu)/\sigma$ , which has mean 0, has standard deviation 1, is dimensionless, and is invariant under any linear change of variable:  $X' = aX + b$ .

### 3. IMPORTANT PROBABILITY DISTRIBUTIONS

**Binomial or Bernoulli Distribution.** If  $X$  represents the number of "successes" in  $n$  independent trials of an experiment, with probability  $p$  of "success" each time, then  $X$  takes on the values 0, 1, 2,  $\dots$ ,  $n$  with probabilities

$$(10) \quad f(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}, \quad q = 1 - p.$$

Hence the sample space  $S$  has  $2^n$  points  $\xi$ , each representing one particular succession of successes and failures. The random variable  $X$  assigns to each  $\xi$  the number of successes in  $\xi$ . The mean and standard deviation are found to be

$$(11) \quad \mu = np, \quad \sigma = \sqrt{npq}.$$

**Poisson Distribution.** A discrete random variable  $X$  with values 0, 1, 2,  $\dots$ , is said to have a Poisson distribution if the corresponding function  $f(x)$  has form

$$(12) \quad f(x) = e^{-\alpha} \frac{\alpha^x}{x!} \quad (x = 0, 1, 2, \dots),$$

where  $\alpha$  is a positive constant. One finds

$$(13) \quad \mu = \alpha, \quad \sigma = \sqrt{\alpha}.$$

For large  $n$  and small  $p$  the binomial distribution (10) is well approximated by the distribution (12), with  $\alpha = np$ .

If a number of events occur independently in space or time and if  $X$  represents the number of these events occurring in any given space or time interval, then the Poisson distribution is a good model for the distribution of  $X$ . *Examples* are the number of red corpuscles on a microscope slide, the rate of emission of electrons or  $\alpha$ -particles, the number of incoming calls to a telephone exchange.

**Normal Distribution.** Let  $X$  be a continuous random variable with density

$$(14) \quad \phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-(x-\mu)^2/2\sigma^2}.$$

Then  $X$  is said to have a *normal distribution*; its mean and standard deviation are  $\mu$  and  $\sigma$ . One terms  $\phi(x)$  the *normal density function* of mean  $\mu$  and standard deviation  $\sigma$ ; the corresponding distribution

$$(15) \quad \Phi(x) = \int_{-\infty}^x \phi(t) dt$$

is the *normal distribution function*. The function  $\Phi$  is tabulated for  $\mu = 0$  and  $\sigma = 1$ , and any other case is reduced to this by replacing  $X$  by its standardized variable  $X^*$  (Sect. 2). See Table 2, Sect. 9.

For large values of  $n$ , the binomial distribution may be approximated by the normal distribution having  $\mu = np$ ,  $\sigma = \sqrt{npq}$ . More precisely, if  $X$  has a binomial distribution, then as  $n \rightarrow \infty$

$$(16) \quad Pr\left(\frac{X - np}{\sqrt{npq}} \leq t\right) = Pr(X^* \leq t) \rightarrow \Phi(t).$$

**The  $\chi^2$ -Distribution.** Let  $X$  be a continuous random variable with values in the range  $0 \leq x < \infty$ . Then  $X$  is said to have a  $\chi^2$ -distribution with  $n$  degrees of freedom, if  $X$  has density

$$(17) \quad k_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2}, \quad x \geq 0.$$

One finds

$$(18) \quad \mu = n, \quad \sigma = \sqrt{2n} \quad (n = 1, 2, \dots).$$

This type of distribution is of great importance in the theory of sampling of normal populations (Sect. 4). See Table 3, Sect. 9.

**Student  $t$ -Distribution.** Let  $X$  be a continuous random variable with density

$$(19) \quad s_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}.$$

Then  $X$  is said to have a Student  $t$ -distribution with  $n$  degrees of freedom ( $n = 1, 2, \dots$ ). One finds

$$(20) \quad \mu = 0, \quad \sigma = \sqrt{\frac{n}{n-2}}.$$

As  $n \rightarrow \infty$ ,  $s_n(x)$  approaches the normal density function of mean 0 and standard deviation 1. The  $t$ -distribution is of value in sampling theory (Sect. 4). See Table 4, Sect. 9.

#### 4. SAMPLING

In a great variety of practical problems a precise answer is obtainable only by making a very large number of measurements. For the sake of economy, one makes a smaller number of measurements and estimates the true answer from these. The theory of such methods of estimation is called sampling. *Examples.* The average height of 1,000,000 soldiers can be estimated by averaging the heights of a selected 1000 soldiers. The outcome of a presidential election can be estimated by polling a small number of voters.

The successive measurements in an experiment yield a random sequence  $X_1, \dots, X_n$  called a *sample*.

**EXAMPLE.** The measurements of the height of 1000 soldiers yield 1000 numbers. One can regard each soldier as a sample point  $\xi$ , the aggregate of all 1,000,000 soldiers as the sample space  $S$ . If the heights follow some definite pattern, then there will be a definite probability that the height  $X_1$  be less than a fixed value. Hence, there is a distribution function  $F_1(x_1)$  associated with  $X_1$  and  $x_1$  can be regarded as the value of a random variable  $X_1$ . Similar statements apply to the measurements  $X_2, \dots, X_n$ . If the measurements are independent (i.e., each one is made without considering the others), all measurements have the same distribution  $F(x)$  and  $X_1, \dots, X_n$  are random variables with joint distribution

$$(21) \quad G(x_1, \dots, x_n) = F(x_1)F(x_2)\cdots F(x_n).$$

The assumptions of the example considered will be assumed to hold generally. A sample space is assumed given, with associated probabilities.

A measurement  $x$  is a value of a random variable  $X$ ; the probability that  $X \leq a$  is  $F(a)$ , where  $F$  is the distribution of  $X$ . Successive measurements yield random variables  $X_1, \dots, X_n$ . It will be assumed that these are independent, so that eq. (21) gives the joint distribution.

**Sample Moments.** The *sample mean* or *average* is the number

$$(22) \quad \bar{x} = \frac{X_1 + \dots + X_n}{n}.$$

The sample *moments* about the origin and about the mean are defined respectively as

$$(23) \quad m'_k = \frac{1}{n} \sum_{i=1}^n X_i^k, \quad m_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{x})^k,$$

so that  $\bar{x} = m'_1$ . The number  $s^2 = m_2$  is the *sample variance*. One has the formula

$$(24) \quad s^2 = m'_2 - \bar{x}^2.$$

One can regard  $\bar{x}$  and  $s^2$  as *estimates* for the mean  $\mu$  and variance  $\sigma^2$  of  $X$ ;  $\bar{x}$  and  $s^2$ , and indeed all the moments, are random variables, being functions of  $X_1, \dots, X_n$ .

From the fact that all  $X_i$  have a common distribution  $F(x)$ , one can deduce properties of the distribution of the various moments. For *example*,

$$(25) \quad E(\bar{x}) = E\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n} \sum E(X_i) = \frac{1}{n} \sum \mu = \mu$$

Similarly,

$$(26) \quad E(s^2) = \frac{n-1}{n} \sigma^2$$

**Unbiased Estimate.** A sample estimate is termed *unbiased* if its expectation is equal to the parameter being estimated. Equation (25) shows that  $\bar{x}$  is an unbiased estimate of  $\mu$ ; eq. (26) shows that  $s^2$  is not an unbiased estimate of  $\sigma^2$ , although  $[n/(n-1)]s^2$  is such an unbiased estimate. Unbiasedness is a useful property of an estimate, but it is not as important as some other properties. The bias in  $s^2$  need be considered only if  $n$  is sufficiently small (less than 20, for example), so that  $(n-1)/n$  is appreciably different from 1.

### Computational Procedures

**Data Classification.** The computation of sample moments for large samples is simplified by the classification of the data. In this procedure the *sample range*, the interval from the smallest to the largest sample value, is divided into approximately fifteen *class intervals* of equal width (the *class width*). The number of measurements whose  $x_i$  value lies in each class interval, the *frequency* of the class interval, is then recorded, as well as the midpoint of each interval, the *class mark*. In the subsequent computation one then replaces each sample value  $x_i$  by the class mark of the corresponding class interval; usually a negligible error is introduced by this replacement. *Example.* In measuring height of a population to the nearest 0.1 in. one can choose class intervals 1 in. in width; to avoid ambiguity the end points of the class intervals should be 60.05 in., 61.05 in.,  $\dots$ , for instance, rather than 60 in., 61 in.,  $\dots$ .

**Computation.** If there are  $h$  class intervals with frequencies  $f_j$  and class marks  $\bar{x}_j$  ( $j = 1, \dots, h$ ), then the moments are computed as follows:

$$(27) \quad \bar{x} = \frac{1}{n} \sum_{j=1}^h f_j \bar{x}_j,$$

$$(28) \quad m'_2 = \frac{1}{n} \sum_{j=1}^h f_j \bar{x}_j^2,$$

$$(29) \quad s^2 = \frac{1}{n} \sum_{j=1}^h f_j (\bar{x}_j - \bar{x})^2 = m'_2 - \bar{x}^2.$$

The computation can be further simplified by *coding* the data; that is, by introducing new measurements  $y_j$  by a linear change of variables:

$$(30) \quad \bar{x}_j = ay_j + b \quad (a \neq 0),$$

where the coefficients  $a$ ,  $b$  are chosen to simplify the  $y_j$  data. The new mean and variance  $\bar{y}$  and  $s_y^2$  are related to the old,  $\bar{x}$  and  $s_x^2$ , by the equations

$$(31) \quad \bar{x} = a\bar{y} + b, \quad s_x^2 = a^2 s_y^2.$$

If  $a$  is chosen to be the class width and  $b$  is taken to be one of the class marks (usually chosen near the middle of the range), then the  $y_j$  are integers, positive or negative, so that the computation is considerably simplified. After  $\bar{y}$  and  $s_y^2$  are computed,  $\bar{x}$  and  $s_x^2$  are found from eq. (31). The procedure is illustrated in tabular form in Table 1.

TABLE 1. COMPUTATION OF SAMPLE MEAN AND VARIANCE

Class Intervals $a_j - a_{j-1} = a$	Frequency $f_j$	Class Mark $\tilde{x}_j$	Coded Marks		
			$y_i = \frac{\tilde{x}_j - b}{a}$	$f_j y_i$	$f_j y_i^2$
$a_0 - a_1$	//	$\tilde{x}_1 = a_0 + \frac{1}{2}a$	-5	-10	50
$a_1 - a_2$	///	$\tilde{x}_2 = a_1 + \frac{1}{2}a$	-4	-24	96
$a_2 - a_3$			-3		
$a_3 - a_4$			-2		
$a_4 - a_5$			-1		
$a_5 - a_6$		$b = a_5 + \frac{1}{2}a$	0	0	0
$a_6 - a_7$			1		
$a_7 - a_8$			2		
...	...	...	...	...	...
$a_{h-1} - a_h$	/// ...	$\tilde{x}_h$			
Totals	$n$			$n\bar{y}$	$nm'_{2,y}$

$$s_y^2 = m'_{2,y} - \bar{y}^2, \quad \bar{x} = a\bar{y} + b, \quad s_x^2 = a^2 s_y^2.$$

**Distribution of Sample Moments**

If some information is known concerning the distribution  $F(x)$  of the random variable  $X$  being measured, then one can draw conclusions as to the distributions of the sample moments. These conclusions in turn permit one to make statements as to the accuracy of the sample moments as estimates of the true moments. For example, suppose that the variable  $X$  is distributed uniformly over an interval of length 1; that is,  $F'(x) = f(x) = 1$  for  $c \leq x \leq c + 1$ , and  $f(x) = 0$  otherwise. If  $c$  is unknown, each sample will give information as to its value. A single measurement  $X$  then allows one to conclude that  $X - 1 \leq c \leq X$ , the mean  $c + \frac{1}{2}$  would be estimated as  $X$  and one knows that, with probability 1, the mean lies between  $X - \frac{1}{2}$  and  $X + \frac{1}{2}$ .

One now proceeds to list properties of the distribution of sample moments when various assumptions are made concerning the distribution  $F(x)$ . These results are applied below to estimation of accuracy of the estimates.

**Distribution of  $\bar{x}$  When  $\sigma$  Is Known.** If  $X$  is normally distributed, then  $\bar{x}$  is also normally distributed, with mean  $\mu$  and variance  $\sigma^2/n$  (Sect. 3). Equivalently, one can state that

$$(32) \quad x' = \frac{\bar{x} - \mu}{\sigma} \sqrt{n}$$

has a normal distribution of mean 0 and variance 1. The conclusion is

approximately true even if  $X$  does not have a normal distribution, provided  $n$  is large. (See Chap. 12.)

**Distribution of  $\bar{x}$  When  $\sigma$  Is Unknown.** Let  $s = \sqrt{s^2}$ , the *sample standard deviation* and let

$$(33) \quad t = \frac{\bar{x} - \mu}{s} \sqrt{n - 1},$$

so that  $t$  can be considered as a random variable. If  $X$  is normally distributed, then  $t$  has a Student  $t$ -distribution with  $n - 1$  degrees of freedom. Again the conclusion is approximately true even if  $X$  does not have a normal distribution, provided  $n$  is large. Furthermore, the  $t$ -distribution approaches the normal distribution of mean 0 and variance 1 as  $n \rightarrow \infty$ .

**Distribution of  $s$  When  $\sigma$  Is Unknown.** Let

$$(34) \quad u = \frac{ns^2}{\sigma^2}.$$

If  $X$  is normally distributed, then  $u$  has a  $\chi^2$ -distribution with  $n - 1$  degrees of freedom. Again the conclusion is approximately true for large  $n$ , regardless of the form of  $F(x)$ .

### Confidence Intervals and Hypothesis Testing

The results described are now applied to obtain estimates for the accuracy of  $\bar{x}$  and  $s^2$  as estimates of  $\mu$  and  $\sigma^2$ . The accuracy will be described in the terminology of confidence intervals. The statement "the interval  $(a, b)$  is a 95 per cent confidence interval for  $\mu$ " means that  $Pr(A \leq \mu \leq B)$  is 0.95, where  $A, B$  are random variables with observed values  $a, b$ . One can also say "either  $a \leq \mu \leq b$  or an event of probability only 0.05 has occurred in the sampling."

**Confidence Intervals for  $\mu$  When  $\sigma$  Is Known.** The 95 per cent interval is obtained from the fact that  $(\bar{x} - \mu)\sqrt{n}/\sigma$  has a normal distribution of mean 0 and variance 1. By means of tables (Sect. 9) one determines the number  $t_{0.95}$  on the normal density curve such that 95 per cent of the area lies between  $-t_{0.95}$  and  $t_{0.95}$ ; that is,

$$(35) \quad \Phi(t_{0.95}) - \Phi(-t_{0.95}) = 0.95.$$

(See Fig. 1.) One finds  $t_{0.95} = 1.96$ . Hence with probability 0.95

$$-1.96 \leq (\bar{x} - \mu)\sqrt{n}/\sigma \leq 1.96$$

or equivalently,

$$(36) \quad \bar{x} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{1.96\sigma}{\sqrt{n}}.$$

Thus the 95 per cent confidence interval has end points  $\bar{x} \pm 1.96 \sigma/\sqrt{n}$ . In a similar manner one obtains confidence intervals for other percentages.

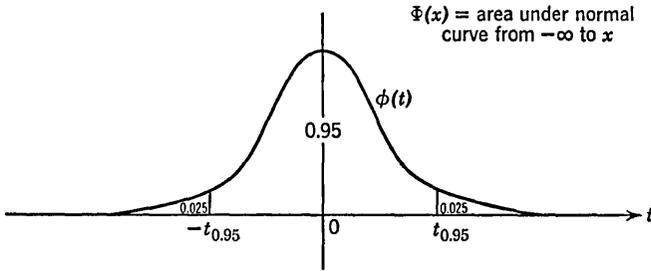


FIG. 1. Normal density curve with 95 per cent limits indicated.

**Confidence Intervals for  $\mu$  When  $\sigma$  Is Unknown.** Now  $s$  must be used instead of  $\sigma$  and  $(\bar{x} - \mu)\sqrt{n-1}/s$  has a  $t$ -distribution with  $n-1$  degrees of freedom. The point  $t_{0.95}$  is obtained from the tables of Sect. 9. Then with probability 0.95

$$-t_{0.95} \leq (\bar{x} - \mu)\sqrt{n-1}/s \leq t_{0.95}$$

or

$$(37) \quad \bar{x} - \frac{st_{0.95}}{\sqrt{n-1}} \leq \mu \leq \bar{x} + \frac{st_{0.95}}{\sqrt{n-1}}$$

Thus  $\bar{x} \pm st_{0.95}/\sqrt{n-1}$  are the end points of the 95 per cent confidence interval for  $\mu$ . A similar procedure is used for other percentages.

**Confidence Intervals for  $\sigma$ .** One uses the fact that  $ns^2/\sigma^2$  has a  $\chi^2$ -distribution with  $n-1$  degrees of freedom. Since the  $\chi^2$ -distribution

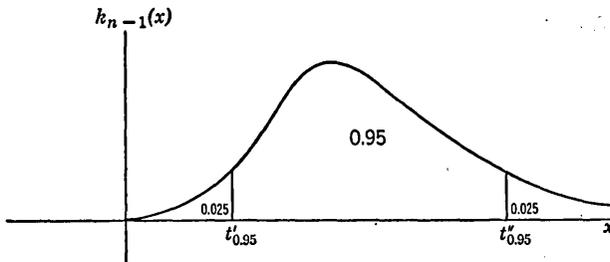


FIG. 2.  $\chi^2$ -density curve with 95 per cent confidence limits indicated.

is unsymmetrical, two points  $t'_{0.95}$  and  $t''_{0.95}$  must be found with the aid of tables (Sect. 9) such that the areas under the  $\chi^2$ -density curve to the left of  $t'_{0.95}$  and to the right of  $t''_{0.95}$  both equal 0.025 (Fig. 2). Then with

probability 0.95

$$t'_{0.95} \leq \frac{s^2 n}{\sigma^2} \leq t''_{0.95}$$

or

$$(38) \quad \frac{ns^2}{t''_{0.95}} \leq \sigma^2 \leq \frac{ns^2}{t'_{0.95}}$$

**Confidence Interval for Binomially Distributed Population.** The preceding confidence intervals were derived under the assumption that the underlying distribution  $F(x)$  was normal, or else that  $n$  was so large that appropriate approximations could be made. Another case of frequent occurrence is that for which the distribution is binomial (Sect. 3) with unknown probability  $p$  of success. The obvious estimate for  $p$  is the sample proportion of successes  $x/n$ . To obtain a confidence interval we use the fact that for  $n$  large the binomial distribution can be approximated by the normal distribution of mean  $np$  and variance  $npq$ . Hence with probability 0.95

$$-t_{0.95} \leq \frac{x - np}{\sqrt{npq}} \leq t_{0.95},$$

where  $t_{0.95} = 1.96$ , as obtained from Table 2 in Sect. 9. The end points of the confidence interval for  $p$  are obtained by setting  $(x - np)/\sqrt{np(1-p)}$  equal to  $\pm t_{0.95}$  and solving the resulting quadratic equation for  $p$ . In this way the end points are found as

$$\frac{c^2 + 2x \pm c\sqrt{c^2 + 4x(1-x/n)}}{2(c^2 + n)}, \quad c = t_{0.95}$$

or, approximately for large  $n$ , as

$$\frac{x}{n} \pm t_{0.95} \frac{\sqrt{x(n-x)}}{n^{3/2}}$$

**Hypothesis Testing.** Frequently, instead of obtaining estimates or confidence intervals for moments, one is merely interested in answering "yes" or "no" to certain hypotheses about the population. For example, one is asked, "On the basis of the observed data, are we justified in rejecting the assumption that  $\mu$  has a specified value?" One method of answering such a question is to construct a confidence interval, say a 95 per cent confidence interval, for  $\mu$ . If the specified value lies outside (inside) this interval, one replies, "Yes, one is (is not) justified in rejecting the assumption at the 95 per cent level of significance." See Refs. 2, 4, 5.

5. BIVARIATE DISTRIBUTIONS

Let  $X$  and  $Y$  be two random variables associated with the same experiment (that is, defined on the same sample space), and not necessarily independent. The following questions arise very frequently: "Is there any functional dependence between  $X$  and  $Y$ ?" "Knowing the value of  $X$ , how would one best predict the value of  $Y$ ?" If there is a strict functional relation  $Y = \phi(X)$ , then all the values  $(x, y)$  fall on the curve  $y = \phi(x)$ . In practice, see Fig. 3, this arises very rarely. Instead, one

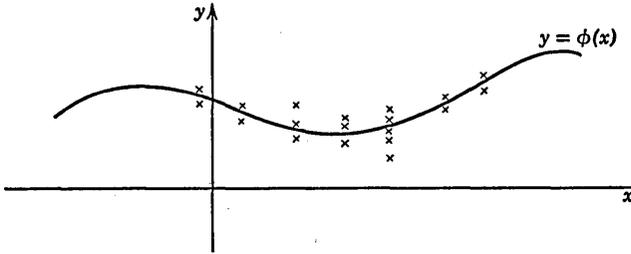


FIG. 3. Regression curve of  $y$  on  $x$ .

may find the values  $(x, y)$  distributed very close to a curve  $y = \phi(x)$ ; that is, the density  $f(x, y)$  may be very high near the curve and negligible far from the curve. For fixed  $x = t$ , the values are distributed along the line  $x = t$  with one-dimensional density, again large near  $y = \phi(x)$  and small when  $y$  differs greatly from  $y = \phi(x)$ . This one-dimensional density is simply the conditional density (Chap. 12)

$$(39) \quad p(y|t) = \frac{f(t, y)}{\int_{-\infty}^{\infty} f(t, y) dy} = \frac{f(t, y)}{f_1(t)},$$

where  $f_1(x), f_2(y)$  are the density functions for  $X$  and  $Y$ . The best estimate for  $\phi(x)$  at  $x = t$  is the mean value of  $Y$  on the line  $x = t$ ; that is, the value of

$$(40) \quad \mu_{y|x} = E(y|x) = \int_{-\infty}^{\infty} yp(y|x) dy.$$

Equation (40) defines  $\mu_{y|x}$  as a function of  $x$ , the true regression function of  $Y$  on  $X$ .

**Computation.** If  $f(x, y)$  is not completely known, measurements in an experiment will lead to various averages from which  $\mu_{y|x}$  can be estimated. If the pairs  $(X_1, Y_1), \dots, (X_n, Y_n)$  are the experimental values

(random vectors) then one computes the quantities

$$(41) \quad \bar{x} = \frac{1}{n} \sum_{k=1}^n X_k, \quad \bar{y} = \frac{1}{n} \sum_{k=1}^n Y_k \quad (\text{sample means})$$

$$s_x^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{x})^2, \quad s_y^2 = \frac{1}{n} \sum_{k=1}^n (Y_k - \bar{y})^2,$$

$$(42) \quad s_{xy} = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{x})(Y_k - \bar{y}) \quad (\text{sample moments of second order}),$$

$$(43) \quad r = \frac{s_{xy}}{s_x s_y} \quad (\text{sample coefficient of correlation}).$$

These can be considered as estimates for the true means, moments about the mean, and coefficient of correlation (Chap. 12)

$$(41') \quad \mu_x = \int_{-\infty}^{\infty} x f_1(x) dx, \quad \mu_y = \int_{-\infty}^{\infty} y f_2(y) dy,$$

$$(42') \quad \sigma_{ij} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)^i (y - \mu_y)^j f(x, y) dx dy \quad (i, j = 1, 2),$$

$$(43') \quad \rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}.$$

Computation of the experimental quantities (41), (42), and (43) can be simplified by coding techniques analogous to those of Sect. 4. (See Refs. 3, 8.)

### Curve-Fitting by Least Squares

From a graphical representation of the data  $(x_k, y_k)$ , one is usually led to some notion of the form of the regression function  $\mu_{y|x}$  (mean of  $Y$  for each  $X$ ). One then chooses an estimate  $\hat{\mu}_{y|x}$  for this function with one or more adjustable parameters. The parameters are then chosen so that

$$\sum_{k=1}^n (y_k - \hat{\mu}_{y|x_k})^2$$

has its minimum value. This is the method of *fitting data by least squares*. The simplest form for  $\hat{\mu}_{y|x}$  is  $Ax + B$ , where  $A, B$  are adjustable. The method of least squares leads to the values

$$(44) \quad A = \frac{r s_y}{s_x}, \quad B = \bar{y} - r \bar{x} \frac{s_y}{s_x}.$$

More generally, one can let

$$(45) \quad \hat{\mu}_{y|x} = A_0 + A_1 x + \cdots + A_m x^m,$$

where the constants  $A_0, \dots, A_m$  are adjustable. The method of least squares leads to the following linear equations for  $A_0, \dots, A_m$ :

$$(46) \quad \sum_{i=0}^m \sum_{k=1}^n A_i x_k^{i+j} = \sum_{k=1}^n x_k^j y_k \quad (j = 0, 1, \dots, m).$$

**Confidence Intervals**

In order to obtain confidence intervals for the estimates, it is necessary to make certain assumptions concerning the form of the underlying density  $f(x, y)$ . Let us assume that the true regression function  $\mu_{y|x}$  is linear, and that the distribution of  $Y$  for fixed  $X$  is normal, with variance independent of  $x$ . These assumptions are often satisfied in practical problems. In practice, the variable  $X$  is usually not determined by random samples, but it is given preassigned values and the corresponding values of  $Y$  are determined experimentally. (*Example.* In a problem of detection of electromagnetic radiation,  $X$  might denote the range and  $Y$  the signal strength; the range  $X$  would be varied at regular intervals and the signal strength  $Y$  recorded once or several times for each value of  $X$ .) Under these assumptions, the true regression function is of form  $ax + b$ . The coefficients  $a$  and  $b$  are estimated by  $A, B$  as given in eq. (44). The two quantities for which confidence intervals are usually required are  $\mu_{y|x}$  itself and the regression coefficient  $a$ .

The required intervals are obtained as follows. One defines

$$(47) \quad S^2 = \frac{1}{n} \sum_{k=1}^n (y_k - Ax_k - B)^2 = s_y^2(1 - r^2).$$

It can then be shown that the quantities

$$(48) \quad \frac{(\mu_{y|x} - Ax - B) \cdot s_x \cdot \sqrt{n - 2}}{\sqrt{(x - \bar{x})^2 + s_x^2} \cdot S}, \quad \frac{(a - A)s_x \sqrt{n - 2}}{S}$$

possess  $t$ -distributions with  $n - 2$  degrees of freedom. Thus 95 per cent confidence intervals for  $\mu_{y|x}$  and  $a$  have end points

$$(49) \quad Ax + B \pm t_{0.95} \frac{\sqrt{(x - \bar{x})^2 + s_x^2} \cdot S}{s_x \sqrt{n - 2}},$$

$$A \pm t_{0.95} \frac{S}{s_x \sqrt{n - 2}},$$

respectively, where  $t_{0.95}$  is determined as in Sect. 4 from the  $t$ -distribution with  $n - 2$  degrees of freedom.

The preceding process can be generalized to the case of more than two

variables. For details on this and on computational procedures, see Refs. 3, 5, 8.

## 6. TESTS FOR GOODNESS OF FIT

A problem of frequent occurrence is to determine whether a set of experimentally determined data is consistent with, or fits, some pre-assigned hypothetical probability distribution. The usual way of studying this is to divide the range of the variable  $X$  into  $k$  subintervals, not necessarily equal in length. The number  $f_i$  of observations falling in the  $i$ th subinterval can then be counted. Let  $p_i$  be the probability assigned to the  $i$ th subinterval by the given hypothetical probability distribution. Let  $n$  be the total number of observations. It can then be shown that the quantity

$$(50) \quad u = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i}$$

has approximately a  $\chi^2$ -distribution with  $n - 1$  degrees of freedom. From eq. (50) it is apparent that large values of  $u$  correspond to large differences between the observed and theoretical distributions. If it is greater than the 0.95 point of the  $\chi^2$ -table, one can say (with 95 per cent certainty) that the sample did not come from the given hypothetical distribution.

The approximation by a  $\chi^2$ -distribution is usually sufficiently accurate if each  $p_i > 5/n$  and  $k > 5$ .

Frequently in such a problem the hypothetical distribution is not completely specified, but contains some adjustable parameters. For example, one might wish to test whether a sample comes from a normal population, in which case the mean and variance of the population must first be estimated from the sample. It can be shown that the  $\chi^2$ -test usually remains valid, provided one further degree of freedom is subtracted for each parameter estimated. More precisely, in order for the test to be valid, the parameters must be estimated by the *method of maximum likelihood*. See Refs. 4, 5.

## 7. SEQUENTIAL ANALYSIS

The usual method of collecting data consists of the determination of a fixed number of observations and their subsequent statistical analysis. Frequently a considerable reduction in the number of observations required can be made by making the observations in sequence and re-analyzing the data after each observation. Such a process is known as a *sequential analysis* and is particularly useful for such problems as production testing.

**EXAMPLE.** Consider a population whose density function  $f(x; \theta)$  depends on some parameter  $\theta$  (mean, variance, etc.) whose value is not known; let

us suppose that  $\theta$  can take only one of two given values  $\theta_0, \theta_1$ . The problem is to decide which value is the correct one. In such a decision problem, errors can be made in two ways: by deciding that  $\theta_1$  is correct when  $\theta_0$  is actually the true value of  $\theta$ , or by deciding that  $\theta_0$  is correct when  $\theta_1$  is actually the true value. Denote the probabilities to be assigned to these two types of errors by  $\alpha$  and  $\beta$  respectively. The values of  $\alpha$  and  $\beta$  can be preassigned by an experimenter, and clearly both should be small if one wants to have great confidence in one's decision; however, the smaller  $\alpha$  and  $\beta$  are taken to be the more observations will be required to come to a decision.

Let  $x_1, x_2, \dots$  be the sequence of observed values, and let  $f(x; \theta_j)$  denote the density function of the population when  $\theta_j$  is the true value of  $\theta, j = 0, 1$ . Define the quantities

$$(51) \quad P_{jn} = \prod_{i=1}^n f(x_i; \theta_j) \quad (j = 0, 1), \quad q_n = \frac{P_{1n}}{P_{0n}}.$$

Each  $P_{jn}$  can be found from the preceding one after each observation by multiplying by the corresponding  $f(x_n; \theta_j)$ . The decision rule is then the following.

If

$$(52) \quad \frac{\beta}{1 - \alpha} < q_n < \frac{1 - \beta}{\alpha},$$

take another observation. If  $q_n \leq \beta/(1 - \alpha)$ , decide that  $\theta_0$  is the correct value of  $\theta$ ; if  $q_n \geq (1 - \beta)/\alpha$ , decide that  $\theta_1$  is the correct value of  $\theta$ . Thus the rule is to continue sampling until  $q_n$  leaves the interval (52), choosing  $\theta_0$  if  $q_n$  first leaves the interval to the left and choosing  $\theta_1$  if  $q_n$  first leaves the interval to the right.

A formula is available which allows one to estimate the mean number of observations required before a decision is reached. In general, this number is substantially less than the number required when the sample size is fixed, although one must expect as many or more observations to be required in a small percentage of cases. (See Ref. 7.)

## 8. MONTE CARLO METHOD

A great variety of mathematical and physical problems of apparently nonstatistical nature can be reformulated in statistical form and solved by sampling techniques. For example, the area under a curve  $y = f(x)$ ,  $0 \leq x \leq 1$ ,  $0 \leq f(x) \leq 1$  can be formulated as the probability that a "random point"  $(x, y)$  in the square  $0 \leq x \leq 1$ ,  $0 \leq y \leq 1$  satisfies the inequality  $y \leq f(x)$ . Thus computation of  $\int_0^1 f(x) dx$  is achieved by choos-

ing many random points  $(x, y)$  and counting the proportion which satisfy the condition  $y \leq f(x)$ .

The general program of such statistical solutions to essentially non-statistical problems is described as the Monte Carlo method. The basic idea is very old, but the method has been the subject of unusual interest in the last decade, especially because of the availability of high-speed digital computing machines.

For information on the subject see Ref. 9, which contains an extensive bibliography.

9. STATISTICAL TABLES

TABLE 2. THE CUMULATIVE NORMAL DISTRIBUTION FUNCTION (Ref. 10)

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{x^2}{2}} dx \text{ FOR } 0.00 \leq u \leq 2.99.$$

u	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.9 <sup>2</sup> 0097	.9 <sup>2</sup> 0358	.9 <sup>2</sup> 0613	.9 <sup>2</sup> 0863	.9 <sup>2</sup> 1106	.9 <sup>2</sup> 1344	.9 <sup>2</sup> 1576
2.4	.9 <sup>2</sup> 1802	.9 <sup>2</sup> 2024	.9 <sup>2</sup> 2240	.9 <sup>2</sup> 2451	.9 <sup>2</sup> 2656	.9 <sup>2</sup> 2857	.9 <sup>2</sup> 3053	.9 <sup>2</sup> 3244	.9 <sup>2</sup> 3431	.9 <sup>2</sup> 3613
2.5	.9 <sup>2</sup> 3790	.9 <sup>2</sup> 3963	.9 <sup>2</sup> 4132	.9 <sup>2</sup> 4297	.9 <sup>2</sup> 4457	.9 <sup>2</sup> 4614	.9 <sup>2</sup> 4766	.9 <sup>2</sup> 4915	.9 <sup>2</sup> 5060	.9 <sup>2</sup> 5201
2.6	.9 <sup>2</sup> 5339	.9 <sup>2</sup> 5473	.9 <sup>2</sup> 5604	.9 <sup>2</sup> 5731	.9 <sup>2</sup> 5855	.9 <sup>2</sup> 5975	.9 <sup>2</sup> 6093	.9 <sup>2</sup> 6207	.9 <sup>2</sup> 6319	.9 <sup>2</sup> 6427
2.7	.9 <sup>2</sup> 6533	.9 <sup>2</sup> 6636	.9 <sup>2</sup> 6736	.9 <sup>2</sup> 6833	.9 <sup>2</sup> 6928	.9 <sup>2</sup> 7020	.9 <sup>2</sup> 7110	.9 <sup>2</sup> 7197	.9 <sup>2</sup> 7282	.9 <sup>2</sup> 7365
2.8	.9 <sup>2</sup> 7445	.9 <sup>2</sup> 7523	.9 <sup>2</sup> 7599	.9 <sup>2</sup> 7673	.9 <sup>2</sup> 7744	.9 <sup>2</sup> 7814	.9 <sup>2</sup> 7882	.9 <sup>2</sup> 7948	.9 <sup>2</sup> 8012	.9 <sup>2</sup> 8074
2.9	.9 <sup>2</sup> 8134	.9 <sup>2</sup> 8193	.9 <sup>2</sup> 8250	.9 <sup>2</sup> 8305	.9 <sup>2</sup> 8359	.9 <sup>2</sup> 8411	.9 <sup>2</sup> 8462	.9 <sup>2</sup> 8511	.9 <sup>2</sup> 8559	.9 <sup>2</sup> 8605

Example:  $\Phi(2.57) = .9<sup>2</sup>4915 = .994915.$

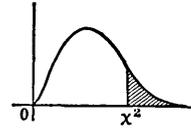


TABLE 3. THE  $\chi^2$  DISTRIBUTION

$P$  = the probability of a  $\chi^2$  deviation greater than the tabulated value

Degrees of freedom	$P = 0.99$	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0.000157	0.000628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.787	5.989	7.779	9.488	11.668	13.277
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.360	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.273
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

For degrees of freedom greater than 30, the expression  $\sqrt{2\chi^2} - \sqrt{2n'} - 1$  may be used as a normal deviate with unit variance, where  $n'$  is the number of degrees of freedom.

Reproduced from *Statistical Methods for Research Workers*, 6th ed., with the permission of the author, R. A. Fisher, and his publisher, Oliver and Boyd, Edinburgh.

TABLE 4. STUDENT'S  $t$  DISTRIBUTION \*

Degrees of freedom $n$	Probability of a deviation greater than $t$					
	.005	.01	.025	.05	.1	.15
1	63.657	31.821	12.706	6.814	3.078	1.963
2	9.925	6.965	4.303	2.920	1.886	1.386
3	5.841	4.541	3.182	2.353	1.638	1.250
4	4.604	3.747	2.776	2.132	1.533	1.190
5	4.032	3.365	2.571	2.015	1.476	1.156
6	3.707	3.143	2.447	1.943	1.440	1.134
7	3.499	2.998	2.365	1.895	1.415	1.119
8	3.355	2.896	2.306	1.860	1.397	1.108
9	3.250	2.821	2.262	1.833	1.383	1.100
10	3.169	2.764	2.228	1.812	1.372	1.093
11	3.106	2.718	2.201	1.796	1.363	1.088
12	3.055	2.681	2.179	1.782	1.356	1.083
13	3.012	2.650	2.160	1.771	1.350	1.079
14	2.977	2.624	2.145	1.761	1.345	1.076
15	2.947	2.602	2.131	1.753	1.341	1.074
16	2.921	2.583	2.120	1.746	1.337	1.071
17	2.898	2.567	2.110	1.740	1.333	1.069
18	2.878	2.552	2.101	1.734	1.330	1.067
19	2.861	2.539	2.093	1.729	1.328	1.066
20	2.845	2.528	2.086	1.725	1.325	1.064
21	2.831	2.518	2.080	1.721	1.323	1.063
22	2.819	2.508	2.074	1.717	1.321	1.061
23	2.807	2.500	2.069	1.714	1.319	1.060
24	2.797	2.492	2.064	1.711	1.318	1.059
25	2.787	2.485	2.060	1.708	1.316	1.058
26	2.779	2.479	2.056	1.706	1.315	1.058
27	2.771	2.473	2.052	1.703	1.314	1.057
28	2.763	2.467	2.048	1.701	1.313	1.056
29	2.756	2.462	2.045	1.699	1.311	1.055
30	2.750	2.457	2.042	1.697	1.310	1.055
$\infty$	2.576	2.326	1.960	1.645	1.282	1.036

The probability of a deviation numerically greater than  $t$  is twice the probability given at the head of the table.

\* This table is reproduced from *Statistical Methods for Research Workers*, with the generous permission of the author, Professor R. A. Fisher, and the publishers, Messrs. Oliver and Boyd.

REFERENCES

1. I. W. Burr, *Engineering Statistics and Quality Control*, McGraw-Hill, New York, 1953.
2. H. Cramer, *The Elements of Probability Theory*, Wiley, New York, 1954.
3. W. J. Dixon and F. J. Massey, *Introduction to Statistical Analysis*, McGraw-Hill, New York, 1951.
4. P. G. Hoel, *Introduction to Mathematical Statistics*, Wiley, New York, 1947.
5. A. M. Mood, *Introduction to the Theory of Statistics*, McGraw-Hill, New York, 1950.
6. J. Neyman, *First Course in Probability and Statistics*, Henry Holt, New York, 1950.
7. A. Wald, *Sequential Analysis*, Wiley, New York, 1947.
8. G. U. Yule and M. G. Kendall, *Introduction to the Theory of Statistics*, Giffin and Co., London, 1937.
9. *Symposium on Monte Carlo Methods*, H. A. Meyer, Editor, Wiley, New York, 1956.
10. A. Hald, *Statistical Tables and Formulas*, Wiley, New York, 1952.



# NUMERICAL ANALYSIS

## B. NUMERICAL ANALYSIS

Richard F. Clippinger  
and Joseph H. Levin, Editors

14. Numerical Analysis, by *Bernard Dimsdale*  
*Murray Mannos*  
*J. M. Cameron*  
*R. F. Clippinger*  
*J. B. Diaz*  
*Bernard Friedman*  
*Eugene Isaacson*  
*Robert Richtmyer*



## Numerical Analysis

Richard F. Clippinger  
and Joseph H. Levin, Editors

1. Interpolation, Curve Fitting, Differentiation, and Integration, by Bernard Dimsdale	14-01
2. Matrix Inversion and Simultaneous Linear Equations, by Murray Mannos	14-13
3. Eigenvalues and Eigenvectors, by Murray Mannos	14-28
4. Digital Techniques in Statistical Analysis of Experiments, by Joseph M. Cameron	14-48
5. Ordinary Differential Equations, by Richard F. Clippinger	14-55
6. Partial Differential Equations, by J. B. Diaz, Richard F. Clippinger, Bernard Friedman, Eugene Isaacson, and Robert Richtmyer	14-64
References	14-88

### 1. INTERPOLATION, CURVE FITTING, DIFFERENTIATION, AND INTEGRATION

Bernard Dimsdale

**Definitions.** Suppose  $f(x)$  is a function about which the following is known: at each of  $n + 1$  points  $x_0, x_1, \dots, x_n$ , called the *basic set* of points, the numerical value of  $f$  or of one of its derivatives is known. It is to be noted that  $x$  may represent one or more independent variables. Suppose  $g(x; a_0, a_1, \dots, a_n)$  is given analytically and the  $a$ 's are determined so that  $g$  has the same numerical property as  $f$  at each point of the basic set. Then  $g$  is called an *interpolating function* for  $f$ , and  $R = f - g$  is called the *remainder*.

In the event that  $g$  is linear with respect to the  $a$ 's, that is  $g(x; a) = a_0g_0(x) + a_1g_1(x) + \cdots + a_ng_n(x)$  the interpolating function is called *linear*, and the functions  $g_0, g_1, \cdots, g_n$  are called *basic interpolating functions*. In the further event that  $x$  is a single variable and  $g_i(x) = x^i$  the function  $g$  is called an *interpolating polynomial*.

If a function  $\bar{g}(x; a_0, \cdots, a_m)$  is given analytically for  $m \leq n$ , any requirement whatsoever on  $f - g$  over the basic set establishes  $\bar{g}$  as a *curve-fitting function*. If that requirement is that

$$\sum_{i=0}^n [f(x_i) - g(x_i; a)]^2 w(x_i)$$

be minimal then  $g$  is a *least square fit* to  $f$ , relative to the *weight function*  $w$ , which is presumed to be positive. Again  $g$  may be nonlinear, linear, or polynomial.

### Interpolation

**General Solution of Interpolating Problem.** For nonlinear  $g$  the definitions imply that the  $a$ 's can be determined by solving  $n + 1$  simultaneous nonlinear algebraic equations. For linear  $g$  the equations for  $a$  are linear and the problem is solved when an  $(n + 1)$ st order matrix is inverted, which of course presupposes that it is not singular. No element of this matrix depends on the values of  $f$  or its derivatives, so that the inverted matrix can be used for all those functions  $f$  for which the conditions of interpolation, the basic interpolating functions, and the basic set of points are the same.

**Interpolating Polynomials for Arbitrary Basic Point Sets.** If the derivatives of  $f$  are not involved in the interpolation, then

$$g(x) = \sum_{i=0}^n \frac{h_i(x)}{h_i(x_i)} f(x_i),$$

$$R(x) = \frac{f^{(n+1)}(\xi)h(x)}{(n+1)!},$$

where  $h(x)$  is the product of all  $x - x_\nu$ ,  $\nu = 0, 1, \cdots, n$ ;  $h_i(x)$  is the same except that the factor  $x - x_i$  is deleted,  $f^{(n+1)}(\xi)$  is the  $(n + 1)$ st derivative of  $f(x)$ ,  $\xi$  is an unknown function of  $x$ , but is some number between the least and the greatest of the basic set of points. This is *Lagrange's formula*, and has been put in practicable computing form by Aitken. Form the table

$x_0$	$f_{0,0}$			
$x_1$	$f_{1,0}$	$f_{1,1}$		
$x_2$	$f_{2,0}$	$f_{2,1}$	$f_{2,2}$	
$x_3$	$f_{3,0}$	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$
...	...	...	...	...

where  $f_{j,0} = f(x_j)$ ,

$$f_{k,j+1} = f_{j,j} + \frac{(x_j - x)f_{j,j} - f_{k,j}}{x_k - x_j}, \quad j > 0.$$

If sufficient information about some derivative, say the  $p$ th, is available to show that  $R = f^{(p)}(\xi)h(x)/p!$  is sufficiently small for interpolation purposes,  $p$  values of  $x$  and  $f$  will give the interpolated value of  $f$  as the rightmost number in the table, within an error bounded by  $R$ . If not, then an iterative process can be undertaken, starting with two points and adding one point at a time together with entries in the table until two consecutive values in the next to the last column are sufficiently close.

EXAMPLE.  $x = 0.4$ .

$k$	$x_k$	$f_{k,0}$	$f_{k,1}$	$f_{k,2}$	$f_{k,3}$	$f_{k,4}$
0	0	0				
1	1	1	0.4			
2	2	8	1.6	-0.32		
3	0.5	0.125	0.1	0.04	0.064	
4	-1	-1	0.4	0.4	0.064	0.064

Hence  $f(0.4)$  is taken to be 0.064. Here  $f(x) = x^3$ ,  $f^{(4)}(x) = 0$ . Thus four points would have been sufficient.

In the event that derivatives are also given, *Neville's procedure* applies (see Ref. 1).

**Interpolating Polynomials for Uniformly Spaced Points.** In the event that the basic set of points has the property that  $x_{p+1} - x_p = h$ , where  $h$  does not change with  $p$ , the procedure to be followed, if derivatives do not enter, involves a difference table as follows:

$x_p$	$f_p$		$\Delta^2 f_{p-1}$	
		$\Delta f_p$		$\Delta^3 f_{p-1} \dots$
$x_{p+1}$	$f_{p+1}$		$\Delta^2 f_p$	
		$\Delta f_{p+1}$		$\Delta^3 f_p \dots$
$x_{p+2}$	$f_{p+2}$		$\Delta^2 f_{p+1}$	
		$\Delta f_{p+2}$		$\Delta^3 f_{p+1} \dots$
$x_{p+3}$	$f_{p+3}$		$\Delta^2 f_{p+2}$	

where  $\Delta^k f_q = \Delta^{k-1} f_{q+1} - \Delta^{k-1} f_q$  and  $f_j = f(x_j)$ , that is, any element with a  $\Delta$  is the difference of its two adjacent left neighbors, and is obtained by

subtracting the upper one from the lower one, and the subscripts on  $f$  are constant along a line running diagonally downward to the right.

Let

$$u = \frac{x - x_0}{h},$$

$$(u)_r = \frac{u(u-1)\cdots(u-r+1)}{r!},$$

then

$$g(x) = f_p + (u-p)_1 \Delta f_p + (u-p)_2 \Delta^2 f_p + \cdots + (u-p)_k \Delta^k f_p,$$

and

$$R(x) = f^{(k+1)}(\xi) h^{k+1} (u-p)_{k+1},$$

which is *Newton's forward formula* over the basic set of points

$$x_p, x_{p+1}, \cdots, x_{p+k}.$$

There is just one interpolation polynomial through these points, but it has many representations in terms of differences. In fact, there is a representation involving the set of differences obtained by starting with the highest order difference, adding to the set one of its two left neighbors, and repeating until the  $f$  column is reached. The precise representation can be written down by use of a Lozenge diagram (see Ref. 2 or 3). Since all these are representations of the same interpolation polynomial, the remainder term is the same for all of them, and depends only on the particular highest order difference used.

**The Newton Interpolating Formulas.** With the notation  $[x] =$  largest integer in  $x$ , so that, for *example*,  $[2] = [2.2] = [5/2] = 2$ , the following four formulas serve three purposes:

(a) *Newton forward* (basic set  $x_0, x_1, \cdots, x_n$ ):

$$f(x) = f_0 + \sum_{p=1}^n (u)_p \Delta^p f_0 + f^{(n+1)}(\xi) h^{n+1} (u)_{n+1}.$$

(b) *Newton backward* (basic set  $x_{-n}, \cdots, x_{-1}, x_0$ ):

$$f(x) = f_0 + \sum_{p=1}^n (u+p-1)_p \Delta^p f_{-p} + f^{(n+1)}(\xi) h^{n+1} (u+n)_{n+1}.$$

(c) *Newton-Gauss 1* (basic set  $x_{-\nu}, x_{1-\nu}, \cdots, x_{n-\nu}$ ;  $\nu = [n/2]$ ):

$$f(x) = f_0 + \sum_{p=1}^n (u + [p/2 - \frac{1}{2}])_p \Delta^p f_{-[p/2]} + f^{(n+1)}(\xi) h^{n+1} (u + [n/2])_{n+1}.$$

(d) *Newton-Gauss 2* (basic set  $x_{-\nu}, x_{1-\nu}, \dots, x_{n-\nu}; \nu = [n/2 + \frac{1}{2}]$ ):

$$f(x) = f_0 + \sum_{p=1}^n (u + [p/2])_p \Delta^p f_{-[p/2+\frac{1}{2}]} + f^{(n+1)}(\xi) h^{n+1} (u + [n/2 + \frac{1}{2}])_{n+1}.$$

All the  $\xi$ 's, of course, are different and have values somewhere in the interval defined by the basic set of points.

EXAMPLE. Find  $f(3.4256)$ .

$x$	$f(x)$	$f$	$\Delta^2 f$	$\Delta^3 f$
3.0	0.4771213			
		0.0280287		
3.2	0.5051500		-0.0016998	
		0.0263289		0.0001945
3.4	0.5314789		-0.0015053	
		0.0248236		0.0001628
3.6	0.5563025		-0.0013425	
		0.0234811		0.0001378
3.8	0.5797836		-0.0012047	
		0.0222764		
4.0	0.6020600			

by using

(a) Newton forward,  $x_0 = 3.4$ .

$$\begin{aligned} (u)_1 = u = 0.128, \quad (u)_2 = -0.055808, \quad (u)_3 = +0.05224, \\ f(x) = 0.5314789 + 0.0248236(u)_1 - 0.0013425(u)_2 + 0.0001378(u)_3 \\ = 0.5347384. \end{aligned}$$

(b) Newton backward,  $x_0 = 3.6$ .

$$\begin{aligned} (u)_1 = u = -0.872, \quad (u+1)_2 = -0.055808, \quad (u+2)_3 = -0.02098, \\ f(x) = 0.5563025 + 0.0248236(u)_1 - 0.0015053(u+1)_2 \\ + 0.0001945(u+2)_3 \\ = 0.5347362. \end{aligned}$$

(c) Newton-Gauss 1,  $x_0 = 3.4$ .

$$\begin{aligned} (u)_1 = u = 0.128, \quad (u)_2 = -0.055808, \quad (u+1)_3 = -0.02098, \\ f(x) = 0.5314789 + 0.0248236(u)_1 - 0.0015053(u)_2 + 0.0001628(u+1)_3 \\ = 0.5347369. \end{aligned}$$

Here  $f(x)$  is  $\log_{10} x$  and  $f(x)$  is in fact 0.5347366.

In interpolating for  $f(x)$  it is reasonable to select a formula such that  $x_0, x_1$  (or  $x_0, x_{-1}$ ) contains  $x$ , for then the contribution of the differences to the computed  $f(x)$  will in general be smallest, a point of considerable interest if remainders cannot be computed. If this condition is accepted, then forward differences must be used at the beginning of a table or immediately following a discontinuity in  $f$  or its pertinent derivatives. Likewise backward differences must be used at the end of a table or preceding a discontinuity. Centered differences may be used in the body of the table. In this connection a further comment should be made. The function  $(u)_{n+1}$  takes on much smaller values on the average near the center of its basic set of points than near the edge. Using centered differences will therefore normally achieve a given accuracy with less differences.

### Curve Fitting

**Linear Least Square Curve Fitting.** Upon differentiating

$$\sum_{s=0}^n [a_0 g_0(x_s) + a_1 g_1(x_s) + \cdots + a_m g_m(x_s) - f(x_s)]^2 w(x_s),$$

with respect to each of the  $a$ 's, and setting these derivatives equal to zero, the equations

$$G_{k0}a_0 + G_{k1}a_1 + \cdots + G_{km}a_m = F_k, \quad k = 0, 1, \cdots, m$$

result, where

$$G_{ij} = \sum_{s=0}^n g_i(x_s)g_j(x_s)w(x_s),$$

$$F_i = \sum_{s=0}^n g_i(x_s)f(x_s)w(x_s),$$

thus leading again to a linear system of equations to be solved. Note that the matrix  $G$  is the product of a rectangular matrix whose elements are  $g_i(x_j)$  by its transpose and is therefore a non-negative definite symmetric matrix. Symmetry implies that the amount of labor in solving the linear system is half the normal amount, and non-negative definiteness implies that *any* indication of singularity obtained means that the system is singular, which is not necessarily true in the more general case. Note that nothing up to this point prevents  $x$  from representing more than one independent variable, so that surfaces in many variables are amenable to this kind of fit.

The task of forming the  $G$  matrix will greatly exceed the task of inverting if  $n$  is much greater than  $m$ , since the number of multiplications (and additions) is measured by  $\frac{1}{2}mn^2$ , whereas the corresponding number for the

inversion is  $\frac{1}{2}m^3$ . If, however,  $g_k(x)$  is  $x^k$ ,  $e^{kx}$ , or  $e^{kx\sqrt{-1}}$ , then  $G_{ij} = G_{rs}$  if  $i + j = r + s$ . Thus only  $2n + 1$  of these numbers need be computed, at a cost of about  $mn$  multiplications and additions. That is, linear least square fitting with polynomials in  $x$ , in  $e^x$ , or with trigonometric sums involving terms such as  $\sin kx$ ,  $\cos kx$  is much less laborious than the more general case.

**Machine Solution for Linear Fitting.** In any but the simplest cases, for very limited values of  $m$  and  $n$ , this kind of calculation is fit material for an automatic computer. If subroutines for evaluation of  $g_k(x)$ , for the product of two rectangular matrices and for inversion of positive definite symmetric matrices are available, it is a relatively simple matter to program the solution of the problem, including evaluation of residuals, that is, of

$$\sum_{s=0}^n a_s g_s(x_k) - f(x_k),$$

for all  $k$ .

The technique for polynomial least square fitting has been reworked by von Holdt (Ref. 4) so that any given set of data (one independent variable) can be fitted with polynomials of every order from 1 to  $m$ , and all residuals computed for each polynomial, with a total computing labor no greater than that required for direct fitting, and calculation of residuals, of an  $m$ th order polynomial. With such a procedure it is possible, for example, to continue the computation until a polynomial is reached for which the largest absolute residual is sufficiently small. There is no doubt that this is possible, barring singularities, since for  $m = n$  every residual is zero.

The method of von Holdt loses its value for large values of  $n$ , because there is a rapid accumulation of roundoff error. This difficulty is minimized by a method which makes use of orthogonal polynomials (Ref. 4a).

**Nonlinear Least Square Curve Fitting.** The problem of selecting  $a = (a_0, a_1, \dots, a_m)$  so as to minimize

$$T(a) = \sum_{s=0}^n [g(x_s; a) - f(x_s)]^2 w(x_s)$$

may be approached by Newton's method as follows. Let  $a^0 = a_0^0, a_1^0, \dots, a_m^0$  be arbitrarily assigned and compute

$$G_{ij} = \sum_{s=0}^n \frac{\partial g(x_s; a^0)}{\partial a_i} \frac{\partial g(x_s; a^0)}{\partial a_j},$$

$$F_i = \sum_{s=0}^n [f(x_s) - g(x_s; a^0)] \frac{\partial g(x_s; a^0)}{\partial a_i},$$

the notation  $\partial g(x_s; a^0)/\partial a_i$  meaning that  $g$  is differentiated partially with respect to  $a_i$  and  $a$  is then set equal to  $a^0$ , which is a set of known numbers. Solve the system of equations

$$\sum_{j=0}^m G_{ij} \Delta a_j = F_i, \quad i = 0, 1, \dots, m,$$

for  $\Delta a_j$  and let

$$a_j^1 = a_j^0 + \Delta a_j, \quad j = 0, 1, \dots, m.$$

With  $a_j^1$  as a new  $a_j^0$ , iterate the process, which may converge to a solution of the problem or, more precisely, to a relative minimum of  $T$ . The  $G$  matrix here is again non-negative definite.

**Levenberg's Method.** Unfortunately this process may not converge. Levenberg (Ref. 5) has provided a modification which guarantees that each step of the process reduces the value of  $T$  unless  $T$  is already at a relative minimum. The iterations go as follows:

1. Compute the  $G$  matrix as before. If  $G$  has no zeros on the diagonal, call it the current  $G$  matrix. If it has, replace them by arbitrary positive numbers, for which neighboring nonzero diagonal elements will suffice, and call this the current  $G$  matrix. No damage is suffered if small but nonzero elements are replaced in this way.

2. If the current  $G$  matrix is singular, or effectively singular for practical purposes, go to (4). If not, compute  $\Delta a$  and hence  $a^1$ , by using the current  $G$  matrix.

3. Compute  $T(a^1)$ . If  $T(a^1) < T(a^0)$ , let  $a^1$  become the new  $a^0$  and go to (1). If  $T(a^1) \geq T(a^0)$ , go to (4).

4. Double the diagonal elements of the current  $G$  to form a new current  $G$  and go to (2).

The description is complete, except for a choice of criterion on which to terminate the process. The criterion may involve specifying an acceptable value of  $T$ , an acceptable value of the maximum residual, or a maximum number of doublings of the diagonal, for example.

In practice, it develops that for most iterations no replacing of diagonal elements and no doubling are necessary. Such iterative steps are identical with Newton's method. Only when Newton's method fails does the additional procedure become operative. Therefore, speaking loosely, Levenberg's method resembles Newton's method as closely as possible without being subject to the possibility of divergence.

## Differentiation

**Forward and Backward Formulas for Numerical Differentiation.** The most commonly used formulas are those obtained by differentiating

the Newton formulas (see subsection, The Newton Interpolating Formulas) for equally spaced arguments. For *Newton's forward difference formula* the result is

$$h^m f^{(m)}(x_0) = \sum_{s=m}^n b_{m,s} \Delta^s f_0 + h^{n+1} b_{m,n+1} f^{(n+1)}(\xi),$$

where  $f^{(m)}(x_0)$  is the  $m$ th derivative of  $f(x)$  at  $x = x_0$  and Table 1 gives values of  $b_{m,s}$  adequate for finding derivatives up to and including order 6 by interpolating over as many as eight points.

For *Newton's backward formula* the result is

$$h^m f^{(m)}(x_0) = \sum_{s=m}^n (-1)^{s+m} b_{m,s} \Delta^s f_{-s} + (-1)^{n+m+1} h^{n+1} b_{m,n+1} f^{(n+1)}(\xi).$$

Again, Table 1 gives the values of  $b$ .

TABLE 1. COEFFICIENTS FOR NEWTON'S DIFFERENCE FORMULAS

$s$	$b_{1,s}$	$b_{2,s}$	$b_{3,s}$	$b_{4,s}$	$b_{5,s}$	$b_{6,s}$
1	1					
2	$-\frac{1}{2}$	1				
3	$\frac{1}{3}$	-1	1			
4	$-\frac{1}{4}$	$\frac{1}{12}$	$-\frac{3}{2}$	1		
5	$\frac{1}{5}$	$-\frac{5}{6}$	$\frac{7}{4}$	-2	1	
6	$-\frac{1}{6}$	$\frac{137}{180}$	$-\frac{15}{8}$	$\frac{17}{6}$	$-\frac{5}{2}$	1
7	$\frac{1}{7}$	$-\frac{7}{10}$	$\frac{29}{15}$	$-\frac{7}{2}$	$\frac{25}{6}$	-3
8	$-\frac{1}{8}$	$\frac{363}{560}$	$-\frac{469}{240}$	$\frac{967}{240}$	$-\frac{35}{6}$	$\frac{23}{4}$

**Centered Formulas for Numerical Differentiation.** Upon differentiating the two Newton-Gauss formulas given in the preceding section it develops that for derivatives of even order every other term vanishes, and that for odd derivatives a similar result can be achieved by adding the two formulas and dividing by 2. This is *Stirling's formula* which, for a basic set of  $2N + 1$  points, is

$$h^{2m-1} f^{(2m-1)}(x_0) = \sum_{s=m}^N c_{m,s} \frac{1}{2} [\Delta^{2s-1} f_{-s} + \Delta^{2s-1} f_{1-s}] + R_N,$$

$$R_N = \sum_{s=0}^{m-1} \frac{(2N + 1)!(2m - 1)! c_{s+1, N+1}}{(2s + 1)!(2N + 2m - 2s - 1)!} h^{2N+2m-2s-1} f^{2N+2m-2s-1}(\xi_s).$$

For even derivatives over  $2N + 1$  points the result is

$$h^{2m}f^{(2m)}(x_0) = \sum_{s=m}^N d_{m,s} \Delta^{2s}f_{-s} + \bar{R}_N,$$

$$\bar{R}_N = \sum_{s=0}^{m-1} \frac{(2N+1)!(2m)!c_{s+1,N+1}}{(2s+1)!(2N+2m-2s)!} h^{(2N+2m-2s)} f^{(2N+2m-2s)}(\xi_s).$$

Table 2 gives values of the  $c$ 's and  $d$ 's for derivatives of orders one through six for interpolations over as many as eleven points.

TABLE 2. COEFFICIENTS FOR STIRLING'S FORMULA

$j$	$c_{1j}$	$c_{2j}$	$c_{3j}$	$d_{1j}$	$d_{2j}$	$d_{3j}$
1	1			1		
2	$-\frac{1}{6}$	1		$-\frac{1}{12}$	1	
3	$\frac{1}{30}$	$-\frac{1}{4}$	1	$\frac{1}{90}$	$-\frac{1}{6}$	1
4	$-\frac{1}{140}$	$\frac{7}{20}$	$-\frac{1}{3}$	$-\frac{1}{560}$	$\frac{7}{40}$	$-\frac{1}{4}$
5	$\frac{1}{630}$	$-\frac{41}{3024}$	$\frac{13}{144}$	$\frac{1}{3150}$	$-\frac{41}{7560}$	$\frac{13}{240}$
6	$-\frac{1}{27720}$	$\frac{479}{151200}$	$-\frac{139}{6048}$			

EXAMPLE. Find  $f'(3.4)$ ,  $f''(3.4)$  by Stirling's formulas by using the data for the example in the subsection on the Newton Interpolating Formulas. Here

$$x_0 = 3.4,$$

$$hf'(x_0) = c_{1,1} \frac{1}{2} (\Delta f_{-1} + \Delta f_0) + c_{1,2} \frac{1}{2} (\Delta^3 f_{-2} + \Delta^3 f_{-1}),$$

$$0.2f'(x_0) = \frac{1}{2}(0.0263289 + 0.0248236) - \frac{1}{12}(0.0001945 + 0.0001628)$$

$$= 0.1277282,$$

$$h^2f''(x_0) = d_{1,1} \Delta^2 f_{-1} + d_{1,2} \Delta^4 f_{-2}$$

$$= -0.0015053 - \frac{1}{12}(-0.0000317)$$

$$= -0.03757.$$

Since  $f'(x) = \ln e/x$ ,  $f''(x) = \ln e/x^2$ , the true values are 0.1277337 and 0.0375687, respectively.

**Remarks on Numerical Differentiation.** The formulas given above apply only to calculation of derivatives at tabular points. For derivatives at other points and for derivatives of functions tabulated at unequal intervals, it is possible, but not practical, to differentiate Lagrange's formula. The most obvious thing to do in this case is to form a new table at equal intervals which includes the point at which a derivative is required. In the first case, the coefficients to be used in the interpolation formulas do not change in moving from one point to the next, which simplifies matters considerably. It is to be observed that the situation

then becomes complex as regards the remainder term. On the other hand, it is also to be observed that derivatives cannot be computed with an accuracy greater than the number of significant figures in the highest order difference used. If, therefore, an interpolation can be performed with sufficient accuracy to guarantee the last significant figure written down, differentiation on the interpolated table will give as much accuracy as is available from the original table. Similar remarks apply to the differentiation of functions having more than one variable.

**Integration**

**General Remarks on Numerical Integration.** The most useful characteristic of the integral, from the point of view of numerical integration is the fact that

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx,$$

which means that the problem can be reduced to finding integrals over small ranges.

There are many possibilities in developing formulas with regard to choice of basic sets of points for the interpolation polynomials to be used. The two sets commonly used will be discussed here: (1) equally spaced points starting with  $a$  and ending with  $b$ , which leads to a set of formulas called *Cotes's formulas*; (2) a spacing developed by *Gauss*.

**Cotes's Formulas for Numerical Integration.** Let

$$x_i = a + ih, \quad i = 0, 1, \dots, n,$$

with  $x_n = b$ , that is, define  $h = (b - a)/n$ . Let  $f_i = f(x_i)$ . Then the formulas are as shown in Table 3. The formulas for  $n = 1$  and  $n = 2$  are known respectively as the *trapezoidal rule* and *Simpson's rule*. Except

TABLE 3. COTES'S FORMULAS

$$\begin{aligned} \int_a^b f(x) dx &= \frac{1}{2}(f_0 + f_1) - \frac{h^3}{12}f^{(2)}(\xi), & n = 1 \\ &= \frac{h}{2}(f_0 + 4f_1 + f_2) - \frac{h^5 f^{(4)}(\xi)}{90}, & n = 2 \\ &= \frac{2h}{45} [7(f_0 + f_4) + 32(f_1 + f_3) + 12f_2] - \frac{8h^7 f^{(6)}(\xi)}{945}, & n = 4 \\ &= \frac{h}{140} [41(f_0 + f_6) + 216(f_1 + f_5) + 27(f_2 + f_4) \\ &\quad + 272f_3] - \frac{9h^9}{1400} f^{(8)}(\xi), & n = 6 \end{aligned}$$

for  $n = 1$ , formulas for odd  $n$  are not given, since their remainder term has the same order of magnitude in  $h$  as the preceding even  $n$ .

The choice of  $n$  is a matter of judgment. Generally, the larger  $n$  the less evaluation of integrand is required for given accuracy. If the integrand requires much computing, this is important.

EXAMPLE. If  $f(x) = e^{-x}/x$ ,  $a = 1$ ,  $b > a$  then  $|f^{(n)}(x)| \leq 1.36n!$  If it is required that the remainder term shall not exceed  $10^{-10}$ , then for the above  $n$ 's the  $h$ 's are 0.0008, 0.013, 0.03, 0.06, and the number of evaluations of integrand per unit  $b - a$  is 1250, 77, 33, 16 respectively.

**Gauss's Formula.** For any  $n$  let

$$\begin{aligned} x_i &= a + (b - a)\xi_i, & i &= 0, 1, \dots, [n/2], \\ &= b - (b - a)\xi_i, & i &= [n/2] + 1, \dots, n. \end{aligned}$$

Then, for  $n = 2N$

$$\int_a^b f(x) dx = (b - a) \left[ A_N f_N + \sum_{i=0}^{N-1} A_i (f_i + f_{2N-i}) \right];$$

for  $n = 2N + 1$

$$\int_a^b f(x) dx = (b - a) \sum_{i=0}^N A_i (f_i + f_{2N-i}),$$

where the  $A$ 's and the  $\xi$ 's are given in Table 4.

TABLE 4. VALUES OF  $A_i$  AND  $x_i$  IN GAUSS'S FORMULA

$n = 1$	$\xi_0 = 0.21132$	48654	$A_0 = 0.5$	
$n = 2$	$\xi_0 = 0.11270$	16654	$A_0 = \frac{5}{18}$	
	$\xi_1 = 0.5$		$A_1 = \frac{3}{8}$	
$n = 3$	$\xi_0 = 0.06943$	18442	$A_0 = 0.17392$	74226
	$\xi_1 = 0.33000$	94782	$A_1 = 0.32607$	25774
$n = 4$	$\xi_0 = 0.04691$	00770	$A_0 = 0.11846$	34425
	$\xi_1 = 0.23076$	53449	$A_1 = 0.23931$	43352
	$\xi_2 = 0.5$		$A_2 = 0.28444$	44444
$n = 5$	$\xi_0 = 0.03376$	52429	$A_0 = 0.08566$	22462
	$\xi_1 = 0.16939$	53068	$A_1 = 0.18038$	07865
	$\xi_2 = 0.38069$	04070	$A_2 = 0.23395$	69672
$n = 6$	$\xi_0 = 0.02544$	60438	$A_0 = 0.06474$	24831
	$\xi_1 = 0.12923$	44072	$A_1 = 0.13985$	26957
	$\xi_2 = 0.29707$	74243	$A_2 = 0.19091$	50253
	$\xi_3 = 0.5$		$A_3 = 0.20897$	95918

The remainder term is of order  $h^{2n+1}$ . For further development, see Hobson (Ref. 6).

**Other Integration Methods.** Tchebysheff has developed a method in which the numerical integral has the form

$$k(f_0 + f_1 + \cdots + f_n)$$

which is useful if  $f$  represents data subject to uniform errors, since no error is weighted more than another.

For multiple integration the methods given here may be applied repeatedly. If the number of repeated integrations is quite large, the Monte Carlo method is useful.

For integrals over an infinite range and for infinite integrands, transformations of the variable of integration can frequently be found which remove the difficulty.

## 2. MATRIX INVERSION AND SIMULTANEOUS LINEAR EQUATIONS

Murray Mannos

**General Remarks.** The development of large scale electronic digital computers has made it numerically possible to invert many large size matrices and to solve large systems of linear equations heretofore considered impractical because of their large size. Problems being attacked by matrix inversion include:

(a) The numerical solution of a differential equation, a partial differential equation, or an integral equation satisfying boundary conditions is often achieved by resolving the problem into a large approximating set of algebraic equations.

(b) A nonlinear problem is frequently replaced by a sequence of linear systems yielding successively improved approximations to the original problem.

(c) Large systems of linear equations, at least in part, are serving as preliminary models for economic and business type problems. The object in linear programming (see Chap. 15), for *example*, is to maximize (minimize) a linear objective function such as profit (cost) subject to the restraints imposed by a system of linear equations (or inequalities). If the inverse of the matrix of coefficients of a linear system of equations is already known, the solution to the system is obtained by merely multiplying the inverse by the column vector whose components consist of constants on the right-hand side of the equalities. In the revised simplex technique (Ref. 7) designed for solving linear programming problems it is the inverse of certain basic column vectors that is calculated at each iteration or stage of the algorithm.

Practical ways of solving systems of *linear equations* are divided into two categories: the direct and indirect methods.

(a) The *direct* method yields an exact solution in a finite number of steps provided no roundoff errors are permitted.

(b) The *indirect* method usually involves an infinite number of iterations to get an exact solution. In practice one accepts the fact that one cannot get a precise answer but must be satisfied with a result sufficiently close to the exact result. At this point in the indirect method the calculation is broken off. To be really sure that the answer is sufficiently close either some estimate of roundoff errors must be made or the closeness must be determined perhaps by some physical considerations. Severity of roundoff errors may easily render useless results.

The discussion will be confined to matrices whose elements are *real* and to linear systems whose coefficients are real. Many of the methods and results described apply equally well to the complex elements and coefficients simply by making appropriate word changes. Furthermore, *any matrix of order  $n$  with complex coefficients may be represented by a real matrix of order  $2n$ .*

No "best method" for either inverting matrices or solving linear systems of equations can be recommended. For a given technique, a matrix or a linear system of equations can always be constructed which will not work too well but which may work better with some other technique. In some cases it is a combination of methods, perhaps a direct followed by an indirect method, that works well for a system of linear equations. Ill-conditioned matrices, of which the favorite seems to be the Hilbert matrix, impose an extremely stringent test upon the accuracy of any given matrix inversion technique. A measure of the ill-conditioning of a matrix may be looked upon as the relative smallness of its determinant compared with that of its individual elements. This will suffice here although more sophisticated measures could be used to interpret the notion of ill-conditioned matrices. The Hilbert matrix is denoted by  $H = (h_{ij})$  where  $h_{ij} = 1/i + j + 1$  ( $i, j = 1, 2, \dots, n$ ).

Having obtained by a given technique a not entirely satisfactory approximation for the inverse of a matrix or for a solution to a system of linear equations, one may consider using techniques for improving the inverse of the matrix or the solution to the linear system of equations as the case may be.

*To facilitate the evaluation of procedures for matrix inversion or solution of linear systems for use on digital computers, a summary table of approximate storage requirements and number of operations is presented at the end of the section.*

**Matrix Inversion**

Each nonsingular square matrix  $A$  of order  $n$  has an inverse  $A^{-1}$  such that

$$(1) \quad AA^{-1} = A^{-1}A = I.$$

If for  $A = (a_{ij})$  the elements  $a_{ij}$  ( $i, j = 1, \dots, n$ ) are real, then the elements  $b_{ij}$  of  $A^{-1} = (b_{ij})$  ( $i, j = 1, \dots, n$ ) are also real. If the  $a_{ij}$  of the matrix  $A$  are specified, the problem is to find the numbers  $b_{ij}$  of  $A^{-1}$ . For certain types of matrices this is relatively simple.

(a) If  $D = (d_{ij})$  is a diagonal matrix, that is,  $d_{ij} = 0, i \neq j$  and  $d_{ii} \neq 0$  ( $i = 1, 2, \dots, n$ ), then the elements of its inverse  $D^{-1} = (b_{ij})$  are  $b_{ij} = 0, i \neq j$ , and  $b_{ii} = 1/d_{ii}$  ( $i = 1, 2, \dots, n$ ).

(b) If  $T = (a_{ij})$  is a nonsingular lower triangular matrix, that is,  $a_{ij} = 0, i < j$ , and  $a_{ii} \neq 0$  ( $i = 1, 2, \dots, n$ ), the elements of its inverse  $T^{-1} = B = (b_{ij})$  can be obtained essentially by solving a series of linear equations in one unknown. Multiplying each of the columns of  $B$  by the first row of  $T$  yields

$$a_{11}b_{11} = 1; \quad a_{11}b_{1j} = 0 \quad (j = 2, \dots, n).$$

This yields  $b_{11} = 1/a_{11}$  and  $b_{1j} = 0$  ( $j = 2, \dots, n$ ). Similarly, multiplying  $B$  by the second row of  $T$  gives

$$a_{21}b_{12} + a_{22}b_{22} = 1; \quad a_{21}b_{1j} + a_{22}b_{2j} = 0 \quad (j = 1, \dots, n: j \neq 2).$$

Substituting the known  $b_{1j}$  ( $j = 1, \dots, n$ ) into the latter equations yields new values  $b_{2j}$  ( $j = 2, \dots, n$ ) from the resulting  $n$  linear equations in each of these unknowns. By continuing in this way, multiplication of each of the columns of  $B$  by the  $n$ th row of  $T$  gives

$$a_{n1}b_{1n} + a_{n2}b_{2n} + \dots + a_{nn}b_{nn} = 1; \\ a_{n1}b_{1j} + a_{n2}b_{2j} + \dots + a_{nn}b_{nj} = 0 \quad (j = 1, \dots, n - 1).$$

Substituting the known  $b_{ij}$  ( $i = 1, \dots, n - 1; j = 1, \dots, n$ ) yields the values  $b_{nj}$  ( $j = 1, \dots, n$ ) of the last row of  $B$ .

(c) An old standard method for inverting matrices is given by  $A^{-1} = (1/\det A)(\dots)$ , where the expression in parenthesis is the transpose of the matrix of cofactors of the elements  $a_{ij}$  of the given matrix  $A$ . This method is not to be recommended as practical for  $n$  greater than 3 or 4.

(d) If one has already computed the characteristic polynomial or better still the minimum polynomial of a matrix

$$m(x) = x^m + a_1x^{m-1} + \dots + a_{m-1}x + a_m; \quad a_m \neq 0,$$

then  $A^{-1} = (-1/a_m)(A^{m-1} + a_1A^{m-2} + \dots + a_{m-1}I)$  since  $A$  satisfies its

minimum equation. In general it may be as much trouble calculating the characteristic or minimum polynomial as it is to invert the matrix itself.

(e) Let  $A_i$  denote the  $i$ th row of the nonsingular matrix  $A$  and  $I_i$  the  $i$ th row of the *identity* matrix. Then

$$A_i = \sum_{j=1}^n a_{ij} I_j \quad (j = 1, \dots, n).$$

If one has solved for the  $I_j$ 's in terms of the  $A_i$ 's, then

$$I_j = \sum_{k=1}^n b_{jk} A_k,$$

and the matrix of coefficients of the latter equation is the desired inverse, i.e.,  $A^{-1} = (b_{ij})$ . In general, this method is more cumbersome than a number of the methods described below.

**Jordan-Gauss Method.** Write the matrix  $A$  with the identity matrix beside it as shown

$$(2) \quad \left[ \begin{array}{cccc|cccc} a_{11} & a_{12} & \cdots & a_{1n} & 1 & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & a_{2n} & 0 & 1 & \cdots & 0 \\ \cdot & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} & 0 & 0 & \cdots & 1 \end{array} \right]$$

A series of elementary row operations will be applied to  $A$  and these will also be applied in the same order to  $I$ . When  $A$  has been reduced to  $I$  by a series of elementary row transformations, then  $I$  will in turn be transformed into  $A^{-1}$  by the same transformations, and the process will be finished. If  $A$  is nonsingular, then for some  $i = 1, \dots, n$  it follows that  $a_{i1} \neq 0$ . One can by an exchange of rows guarantee that the element in the first row of the first column is different from zero.

In case the matrix  $(a_{ij})$  has been altered by an exchange of rows one now denotes the left-hand matrix of (2) by  $(b_{ij})$ . Then adding to the  $i$ th row  $-b_{i1}/b_{11}$  times the first row ( $i = 2, \dots, n$ ) the new left-hand matrix of (2) takes the form

$$(3) \quad \left[ \begin{array}{cccc} b_{11} & c_{12} & \cdots & c_{1n} \\ 0 & c_{22} & \cdots & c_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & c_{n2} & \cdots & c_{nn} \end{array} \right]$$

The minor of order  $n - 1$  in the lower right-hand corner of the matrix (3) has rank  $n - 1$  so that at least one of the elements  $c_{2j} \neq 0$  ( $j = 2, \dots,$

$n$ ). Applying the same argument as before to this minor, all elements below the diagonal element of column 2 of the left-hand matrix in (2) may be reduced to zero. Similarly the element in the first row, second column may be reduced to zero. The first column remains unchanged while the second one has been altered to the desired form.

By continuing in this way the left-hand side of (2) may be reduced to the diagonal form

$$\begin{bmatrix} b_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & z_{nn} \end{bmatrix}$$

with each diagonal element being different from 0. By dividing the first row of (2) by  $b_{11}$ , the second row by  $d_{22}$ , etc., the left-hand matrix of (2) is finally reduced to the identity and the right-hand matrix is now  $A^{-1}$ .

The diagonal elements  $b_{11}, d_{22}, \dots$  of the first, second,  $\dots$  columns which are used to reduce the remaining elements of their respective columns to zero are referred to as pivots. Care should be exercised whenever possible not to select a pivot which is too small or too large; otherwise, loss of significance among other difficulties may arise. Numerous variations of the use of elementary row operations for inverting matrices exist in the literature (Ref. 8).

**Partition Method.** Let the nonsingular  $n \times n$  matrix  $A$  be partitioned as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where  $A_{11}$  is an  $m \times m$  minor ( $m < n$ ) which is likewise nonsingular. Then the inverse  $A^{-1}$  of  $A$  is given by the matrix

$$A^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

where  $B_{11} = A_{11}^{-1} + X\Delta^{-1}Y,$   
 $B_{12} = -X\Delta^{-1},$   
 $B_{21} = -\Delta^{-1}Y,$   
 $B_{22} = \Delta^{-1},$

and

$$X = A_{11}^{-1}A_{12}, \quad Y = A_{21}A_{11}^{-1},$$

$$\Delta = A_{22} - YA_{12} = A_{22} - A_{21}X.$$

Inverting a matrix of order  $n$  has been reduced to inverting a matrix of  $m$ , and another of order  $n - m$ . However, one has to pay the price of performing a number of matrix multiplications afterwards.

**Morris Escalator Method.** By starting with the inverse of the  $2 \times 2$  principal minor  $M_{22}$  in the upper left-hand corner of the nonsingular matrix  $A$  one may by the partition method obtain the inverse of the  $3 \times 3$  principal minor  $M_{33}$  in the upper left-hand corner of  $A$ . Then  $M_{33}$  is used to compute the inverse  $M_{44}$  of the  $4 \times 4$  principal minor in the first four rows and columns. Step by step, one dimension at a time, the partition procedure is carried out until  $A^{-1}$  is obtained. The process is uninterrupted until the inverse of one of the  $M_{ii}$  fails to exist, a fact which is established by noting that the corresponding  $\Delta_i = 0$ . This situation is remedied by interchanging the  $i$ th row with an appropriate row, say the  $j$ th, of the remaining  $n - i$  rows of  $A$ , computing the inverse of the new  $i \times i$  principal minor in the left-hand corner, and then continuing as before.

In order to obtain  $A^{-1}$  one must interchange the  $i$ th and  $j$ th columns of the resulting inverse so obtained. If several of the inverses of principal minors encountered fail to exist, a similar procedure applies in each instance.

**Gram Schmidt Orthogonalization Method.** Premultiplication of the nonsingular matrix  $A$  by an appropriate matrix  $P$  transforms  $A$  into an *orthogonal matrix*, i.e.,

$$(4) \quad PA = O.$$

Since the inverse of an orthogonal matrix is its own transpose, it follows from eq. (4) that

$$A^{-1} = A'P'P,$$

where  $P = DN$

$$N = P_{n-1} \cdots P_2 P_1; P_{i-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ & & \ddots & & & & \ddots & \\ & & & \cdot & & & & \cdot \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ c_{i1} & c_{i2} & \cdots & c_{i,i-1} & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ & & \ddots & & \cdot & & \cdot & \\ & & & \cdot & \cdot & & \cdot & \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$D = \begin{bmatrix} 1/|Q_1| & 0 & \cdots & 0 \\ 0 & 1/|Q_2| & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 1/|Q_n| \end{bmatrix}$$

and in turn

$$c_{ij} = -\frac{A_i Q'_j}{Q_j Q'_j}, \quad j = 1, 2, \dots, i - 1,$$

where  $A_i$  denotes the  $i$ th row of  $A$ ,  $Q_j$  denotes the  $j$ th row of  $Q = NA$ , and  $|Q_i|$  denotes the length of the  $i$ th row of  $Q$  considered as a row vector.

**Inversion of Modified Matrices.** If the inverse of a matrix  $A$  is known, the inverse of a matrix differing from  $A$  in only an element, a row, or a column can be found as a result. If the matrix differs from  $A$  by several elements, rows, or columns, its inverse may be realized by repeated application of this method. The method is based on the matrix identity

$$(5) \quad (A + xy')^{-1} = A^{-1} - \frac{(A^{-1}x)(y'A^{-1})}{(1 + y'A^{-1}x)},$$

where  $x$  and  $y$  are arbitrary column vectors. The matrix  $xy'$  can be made to consist of all zeros except the element in the  $i$ th row and  $j$ th column where it is to contain a fixed value  $c$ . This is easily achieved by taking  $x = ce_i$  and  $y = e_j$  where  $e_i$  is the unit column vector containing a 1 in the  $i$ th position and 0 elsewhere. By taking  $y = e_i$  the matrix  $xy'$  has  $x$  for its  $i$ th column and all other columns consist of zeros. Hence, if the vector  $x$  stands for the vector difference of the  $i$ th column of the matrix whose inverse is desired and the  $i$ th column of  $A$ , the required inverse is obtained from eq. (5). A similar argument applies if the matrices differ only in one row.

**Improving a Computed Inverse** (Hotelling and Bodewig, see Refs. 9 and 10). Suppose that the matrix  $C_0$  is considered a sufficiently good approximation to the inverse of the matrix  $A$  so that  $B = I - AC_0$  has very small elements. If necessary, for some specific purpose, the computed inverse can be improved by forming the sequence

$$C_k = C_{k-1}(I + B^{2^{k-1}}), \quad k = 1, 2, \dots$$

Actually, the sequence converges to  $A^{-1}$  and so  $A^{-1}$  is expressible in the following form of an infinite product

$$(6) \quad A^{-1} = C_0 \prod_{k=1}^{\infty} (I + B^{2^{k-1}}).$$

Very frequently the improvement found by computing  $C_0(I + B)$  or perhaps  $C_0(I + B)(I + B^2)$  is sufficiently satisfactory. Although there are a number of variations other than eq. (6) for expressing  $A^{-1}$ , the present scheme has some merit when using an electronic digital computing machine, since it is only necessary to keep successively squared powers of  $B$ , adding this to the identity matrix  $I$ , and premultiplying by the last computed approximation to  $A^{-1}$ .

### Systems of Linear Equations. Direct Methods

Direct methods arrive at an exact solution in a finite sequence of arithmetical operations.

**Elimination.** Given a set of  $m \leq n$  linear equations in  $n$  unknowns

$$(7) \quad \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \dots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

or more briefly in matrix notation

$$Ax = b,$$

the augmented matrix  $(A|b)$  is operated on by a sequence of elementary row operations which reduce the matrix of coefficients  $A$  to echelon form (see Chap. 3). If a row of the reduced form of  $(A|b)$  is of the form  $(0, 0, \dots, 0, c)$ , where  $c \neq 0$ , the system (7) is inconsistent; otherwise, it is consistent. Arbitrary values are assigned to those  $x$ 's which do not correspond to a leading coefficient of 1 in some line; while the remaining  $x$ 's may be solved for in terms of these parameters one at a time as a linear equation in one unknown whose coefficient is 1.

*Note.* In the remainder of this section only the case with  $m = n$  and the matrix  $A$  nonsingular will be considered.

*Use of Cramer's Rule.* Let  $A(k)$  denote the matrix constructed from  $A$  in (7) by replacing column  $k$  by the column  $b$  of right-hand coefficients. Then the unique solution to (7) is given by Cramer's rule in the following form as a ratio of determinants

$$x_k = \frac{\det A(k)}{\det A} \quad (k = 1, \dots, n).$$

For  $n > 3$  or 4 this method is not to be recommended as efficient.

*Known Inverse.* If the inverse  $A^{-1}$  of  $A$  has already been calculated by any of the previously described or perhaps other methods, the solution in matrix form is given by

$$x = A^{-1}b.$$

However, if  $A^{-1}$  must be computed for the sole purpose of getting  $x$ , the method is not always efficient for large values of  $n$ .

*Conjugate Gradient Method.* Most of the iterative schemes involve an infinite number of iterations and so are classified as indirect methods. However, an outstanding iterative scheme called the conjugate gradient method involves but a finite number of iterations and so is classified as a direct method. Because of the way in which the algorithm for this scheme is built up, it seems more appropriate to discuss it after the gradient method, an indirect method. The elegant finite algorithm for the conjugate gradient method seems to have been independently discovered by Stiefel, Hestenes, and Lanczos (Ref. 11). For a linear system  $Ax = b$ ,  $\det A \neq 0$ , of  $n$  equations the algorithm starts with an initial guess  $x_0$  building up successive approximations  $x_1, \dots, x_n$  and finally terminates after at most  $n$  of these steps or iterations. The corresponding residual vectors

$$r_i = Ax_i - b \quad (i = 0, 1, \dots, n)$$

so formed are mutually orthogonal to the preceding ones. If  $r_i \neq 0$  ( $i = 0, 1, \dots, n - 1$ ) then  $r_n$  orthogonal to each  $r_i$  means  $r_n$  must be the null vector  $0$ ; since  $n + 1$  linearly independent vectors of dimension  $n$  cannot exist.

### Systems of Linear Equations: Indirect Methods

By and large this discussion includes most iterative methods since it takes an infinite number of steps to carry through the whole process. An iteration for solving a system of linear equations is a set of rules for operating on an approximate solution  $(x_1^{(k)}, \dots, x_n^{(k)})$  to obtain an improved or more precise solution  $(x_1^{(k+1)}, \dots, x_n^{(k+1)})$ . The sequence of approximate solutions so defined must converge to the actual solution of the given system of equations. In some cases it is a pronounced advantage to start out with a rather good initial approximation  $(x_1^{(0)}, \dots, x_n^{(0)})$ , whereas in others this is not necessarily true. It is frequently advantageous to improve the solution obtained by a direct method by a few iterations, since the direct solution usually is afflicted with roundoff errors.

**Seidel Method.** One starts off with a guess  $(x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})$  as the initial solution to the linear system (7). Substituting in the first equation of (7) the values  $x_2^{(0)}$  for  $x_2$ ,  $x_3^{(0)}$  for  $x_3$ ,  $\dots$ , and finally  $x_n^{(0)}$  for  $x_n$ , and then solving for  $x$  yields a new value  $x_1^{(1)}$  as the first component

of the next approximate solution. Next by substituting in the second equation of (7) the newly gained value  $x_1^{(1)}$  for  $x_1$  and  $x_3^{(0)}$  for  $x_3, \dots, x_n^{(0)}$  for  $x_n$ , and then solving for  $x_2$ , one obtains a new value  $x_2^{(1)}$  as the second component of the next approximation. Continuing in this fashion and finally substituting in the  $n$ th or last equation of (7) the values  $x_1^{(1)}$  for  $x_1, x_2^{(1)}$  for  $x_2, \dots, x_{n-1}^{(1)}$  for  $x_{n-1}$  and solving for  $x_n$  yields the final component  $x_n^{(1)}$  of the new iteration  $(x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)})$ . The approximation  $(x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)})$  is used primarily in the next iteration to obtain the improved approximation  $(x_1^{(2)}, x_2^{(2)}, \dots, x_n^{(2)})$ . One continues in this way.

The process is very well adapted to machine usage. Convergence is assured when either the matrix of coefficients  $A$  is positive definite or when the diagonal element of the  $i$ th row dominates the rest of the row for each  $i$ , that is, when

$$|a_{ii}| > \sum_{i \neq j} |a_{ij}| \quad (i = 1, 2, \dots, n).$$

Convergence is also guaranteed for additional types of matrices, and there are a number of variations of this procedure. In particular, the "back and forth" Seidel method due to Aitken and Rosser was especially designed to handle those cases in which convergence of the regular Seidel method was erratic.

**Relaxation Method.** First write the system (7) in the form

$$(8) \quad \begin{aligned} b_1 - a_{11}x_1 - a_{12}x_2 - \dots - a_{1n}x_n &= 0 \\ b_2 - a_{21}x_1 - a_{22}x_2 - \dots - a_{2n}x_n &= 0 \\ \dots & \\ b_n - a_{n1}x_1 - a_{n2}x_2 - \dots - a_{nn}x_n &= 0 \end{aligned}$$

and assume that none of the diagonal elements  $a_{ii}$  ( $i = 1, \dots, n$ ) is equal to zero. Then take  $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \dots, x_i^{(0)}, \dots, x_n^{(0)})$  as an initial guess to the solution. If it should accidentally happen that  $x^{(0)}$  satisfies (8), one is finished. If not, define the *residual vector* by  $r^{(0)} = (r_1^{(0)}, r_2^{(0)}, \dots, r_n^{(0)})$ , where  $r_i^{(0)}$  ( $i = 1, 2, \dots, n$ ) is the value or residual obtained by substituting  $x^{(0)}$  in the left-hand side of the  $i$ th equation of (8). Suppose that  $r_i^{(0)}$  is a component of largest magnitude in  $r^{(0)}$ . The object then is to reduce the residual  $r_i^{(0)}$  to 0 by altering the value of the  $i$ th component  $x_i^{(0)}$  of  $x^{(0)}$  while keeping the remaining components of  $x^{(0)}$  fixed. The next trial solution  $x^{(1)}$  is constructed as follows:

$$\begin{aligned} x_k^{(1)} &= x_k^{(0)} & (k = 1, \dots, n; k \neq i) \\ x_i^{(1)} &= x_i^{(0)} + \frac{r_i^{(0)}}{a_{ii}}. \end{aligned}$$

This effects a new set of residuals  $r^{(1)}$  with  $i$ th residual equal to 0. Select

the residual of maximum magnitude and similarly apply the same scheme as above to obtain  $x^{(2)}$ . This process is repeated again and again so as ultimately to reduce all residuals to as close to 0 as possible.

It is sometimes possible to speed up convergence by picking residuals not necessarily of maximum magnitude. In fact, by varying several of the variables at one time it may be possible to speed up convergence considerably. However, it would be very difficult to write a code including many such variations and tricks.

*Note.* In the following sections it is often convenient to introduce a measure or *metric* different from the usual one in order to cut down on the amount of computation required.

**Approximations.** Let  $A$  be a symmetric positive definite matrix, then the *length of a vector  $x$  with respect to the metric  $A$*  is defined as  $|x|_A = (x'Ax)^{1/2}$ , and any two vectors  $x$  and  $y$  are *conjugate* or  *$A$ -orthogonal* if  $x'Ay = 0$ . These are extensions of the usual definitions of length and orthogonality. The latter may be obtained from the new definitions by taking  $A = I$ .

With respect to the usual metric,  $|Ax - b|^2 = 0$  if and only if  $Ax - b = 0$ . This means that solving  $Ax = b$  is equivalent to finding an  $x$  such that  $|Ax - b|^2$  is minimized, since it is known that 0 is its minimum value. Likewise with respect to the generalized metric  $B$ ,  $|Ax - b|_B^2 = 0$  if and only if  $Ax - b = 0$  since  $B$  must be positive definite.

Now let

$$(9) \quad f(x) = |Ax - b|_B^2$$

and consider the family of hyperellipsoids

$$(10) \quad f(x) = k,$$

where  $k$  may take on any constant value. Then the solution of the system  $Ax = b$  is the common center of the family of ellipsoids eq. (10). The game then is to construct a set of approximations  $x^{(0)}, x^{(1)}, \dots$  which get us to or close to this center. The more rapidly this happens the less computation is involved.

**Gradient Method.** Start with a guess  $x^{(0)}$  as an initial approximation to the solution of  $Ax = b$ . The ellipsoid of the family (10) obtained by setting  $k = f(x^{(0)})$  passes through the point  $x^{(0)}$  in  $n$ -dimensional space. Then proceed in the direction of the gradient of  $-f(x)$  at  $x^{(0)}$  that is, along the inner normal to the ellipsoid  $f(x) = f(x^{(0)})$ . It is known that  $f(x)$  decreases most rapidly along the latter direction and so it is natural to proceed in this direction until one arrives at the minimum of  $f(x)$  along this inner normal. This happens at that point  $x^{(1)}$  where the inner normal becomes a tangent to one of the family of ellipsoids in eq. (10). Similarly, proceed along the inner normal of the ellipsoid  $f(x) = f(x^{(1)})$  until the

minimum of  $f(x)$  in this direction is reached. Continue in this way and work in closer and closer to the common center of the family of ellipsoids in (10).

The algebraic procedure for solving  $Ax = b$  with respect to the metric  $B$  according to the geometric scheme described above is as follows:

(a) Compute  $C = A'BA$

$$c = A'Bb.$$

(b) Make an initial guess  $x^{(0)}$ .

(c) Use the following algorithm to obtain the approximation  $x^{(i+1)}$  from that of  $x^{(i)}$ . (i) Calculate the vector  $z^{(i)}$  in the direction of the gradient of  $f(x)$  at  $x^{(i)}$ ; i.e.,  $z^{(i)} = Cx^{(i)} - c$ . (ii) Calculate

$$a_i = \frac{(z^{(i)})'z^{(i)}}{(z^{(i)})'Cz^{(i)}}.$$

(iii) Obtain  $x^{(i+1)} = x^{(i)} - a_i z^{(i)}$ , where the coefficient  $a_i$  determines the minimum value of  $f(x)$  in eq. (9) along the inner normal to  $f(x) = f(x^{(i)})$  at  $x^{(i)}$ .

If  $A$  is a symmetric positive definite matrix, it is most convenient to choose the metric  $B = A^{-1}$ , for then  $A$  replaces  $B$  and  $b$  replaces  $c$  throughout the above algorithm with a resulting simplification.

A considerable advantage of the gradient method is that there need not be an accumulation of roundoff error since the vector  $z^{(i)}$  along the gradient can be recalculated for each iteration. The function  $f(x)$  in eq. (9) may be regarded as a measure of the closeness of an approximation  $x$  to the true solution  $A^{-1}b$ . For the gradient method it is true that  $f(x^{(i+1)}) < f(x^{(i)})$  for each  $i$  and that  $f(x^{(i)})$  approaches 0 in the limit; that is,  $x^{(i)}$  converges steadily toward the true solution  $A^{-1}b$ . However, it is still true that the convergence may be slow or, in other words, it may take many iterations to get close to the center of the ellipsoids. A number of variations of the gradient method have been devised to try to speed up the convergence.

**Conjugate Gradient Method.** First consider the case where  $A$  is symmetric and positive definite. The object in the conjugate gradient method as in the gradient method is to get to the common center of the family of ellipsoids eq. (10). However, the route taken in the conjugate gradient method is different from that of the gradient method and is so modified as to get to the center of the family eq. (10) in but a finite number of steps, namely, at most  $n$  iterations.

*The procedure in three dimensions* will be described. The discussion in higher dimensions follows along similar lines.

As before, make an initial guess  $x^{(0)}$  and proceed from  $x^{(0)}$  along the negative gradient of  $f(x)$  or what is the same along the inner normal of

the three-dimensional ellipsoid  $f(x) = f(x^{(0)})$ . Take as the next approximation the point  $x^{(1)}$ , which is the midpoint of the resulting chord of the ellipsoid  $f(x) = f(x^{(0)})$ . Consider the diametral plane through  $x^{(1)}$  containing the locus of midpoints of the chords of  $f(x) = f(x^{(0)})$ , which are parallel to the direction of the inner normal. The diametral plane so formed cuts out a two-dimensional elliptic cross section from the ellipsoid  $f(x) = f(x^{(0)})$ . The common center of the ellipsoids (10) of interest lie in this two-dimensional elliptic cross section, and the method is designed so that all subsequent approximating points shall remain trapped in this cross section. The diametral plane of  $f(x) = f(x^{(0)})$  is likewise a diametral plane of the interior ellipsoid  $f(x) = f(x^{(1)})$  of the family (10) and cuts it in a two-dimensional elliptic cross section lying within the previous one cut from  $f(x) = f(x^{(0)})$ . Next proceed from  $x^{(1)}$  along the gradient of  $f(x)$  within the last elliptic cross section formed and take for  $x^{(2)}$  the midpoint of the chord so formed in the ellipse. In other words, instead of proceeding from  $x^{(1)}$  along the inner normal of the ellipsoid  $f(x) = f(x^{(1)})$  as in the gradient method, proceed along the inner normal of its cross section made by the diametral plane through  $x^{(1)}$ .

Again the locus of centers of chords parallel to the chord through  $x^{(1)}$  and  $x^{(2)}$  forms a diameter of the elliptic cross section of  $f(x) = f(x^{(1)})$ , which contains not only  $x^{(2)}$  but also the center of the family (10). Next proceed from  $x^{(2)}$  along this diameter, choosing its center as the new and final approximation  $x^{(3)}$ . By barring roundoff error,  $x^{(3)}$  yields the exact solution to a linear system of three equations in three unknowns. If either of the chords mentioned above passing through  $x^{(0)}$ ,  $x^{(1)}$  happens to pass also through the center of the family (10), the process will end in only one or two iterations, respectively, instead of three. This will be indicated by the residual  $r^{(1)} = 0$  or  $r^{(2)} = 0$ , respectively.

*Algorithms for a symmetric positive definite matrix* of order  $n$  and for the general  $n$ -dimensional case, respectively, will be given below, where  $p_i$  denotes a vector in the direction of  $x^{(i)}$  to  $x^{(i+1)}$ .

(a) Pick  $x^{(0)}$ ; then let  $p^{(0)} = r^{(0)} = b - Ax^{(0)}$ ,

$$(b) \quad a_i = \frac{|r^{(i)}|^2}{(p^{(i)})'Ap^{(i)}},$$

$$(11) \quad (c) \quad x^{(i+1)} = x^{(i)} + a_i p^{(i)},$$

$$(d) \quad r^{(i+1)} = r^{(i)} - a_i A p^{(i)},$$

$$(e) \quad b_i = \frac{|r^{(i+1)}|^2}{|r^{(i)}|^2},$$

$$(f) \quad p^{(i+1)} = r^{(i+1)} + b_i p^{(i)},$$

where the coefficient  $a_i$  is selected to make  $x^{(i+1)}$  the appropriate distance from  $x^{(i)}$  and  $b_i$  to keep  $p^{(i+1)}$  in the appropriate direction as described above.

The algorithm eqs. (11) may be applied to a matrix which is symmetric and positive semidefinite as well as to a symmetric positive definite matrix.

*In the case where  $A$  is a general matrix,* the system  $Ax = b$  is replaced by the equivalent system

$$(12) \quad A'Ax = A'b,$$

where  $A'A$  is a symmetric and positive semidefinite. The algorithm (11) could thus be applied to eq. (12), but in order to avoid the roundoff errors due to computing  $A'A$ , it is better to use the following algorithm which leads to theoretically equivalent results.

$$(a) \text{ Pick } x^{(0)}, \text{ then let } r^{(0)} = b - Ax^{(0)}, p^{(0)} = A'r^{(0)},$$

$$(b) a_i = \frac{|A'r^{(i)}|^2}{|Ap^{(i)}|^2},$$

$$(c) x^{(i+1)} = x^{(i)} + a_i p^{(i)},$$

$$(d) r^{(i+1)} = r^{(i)} - a_i A p^{(i)},$$

$$(e) b_i = \frac{|A'r^{(i+1)}|^2}{|A'r^{(i)}|^2},$$

$$(f) p^{(i+1)} = A'r^{(i+1)} + b_i p^{(i)}.$$

The conjugate gradient method has numerous advantages in addition to those already mentioned. One may start all over again with the last approximation obtained as the initial approximation in order to nullify the effects of accumulated roundoff errors. Also, each successive approximation is better than its predecessor. It is very important to note that the given matrix is unchanged during the procedure so that the original data are used again and again. This permits use of special properties of the given matrix such as its particular form or sparseness. A number of variations of this technique have been devised.

A great many of the most important works in the field are to be found in the extensive bibliographies of works by Forsythe and Householder (Refs. 8, 9, 12, and 13).

**Computer Storage Requirements and Number of Operations.** Storage requirements for a given problem will vary in general with the machine, with the programmer, and with the layout of the program. Hence, in Table 5 the number of storage locations required for the program of a given technique of matrix inversion or solution of a linear system shall simply be denoted by the symbol  $w$ .

A multiplication or a division will be identified simply as a multiplication. Likewise an addition or a subtraction will be identified as an addition. Since a multiplication requires from about 2 to 10 times as much time as an addition on most computers, greater weight should be accordingly apportioned to the number of multiplications. If the number of multiplications required for a given technique turns out to be, for example  $2n^3 + 3n + 1$ ; then  $3n + 1$  is negligible compared with  $2n^3$  when  $n$  is sufficiently large. One says the number of multiplications required in this case is of the order  $2n^3$ , and this is simply indicated by  $2n^3$ .

In the case of the indirect procedures such as the Seidel, relaxation, and gradient methods the number of iterations necessary for a satisfactory solution varies from problem to problem. In fact, the number of iterations required depends upon the original system of equations, the choice of the initial solution, and the accuracy stipulated beforehand. In these cases storage requirements and the number of operations are given for one iteration. For the conjugate gradient method these will be given totally for all  $n$  iterations.

TABLE 5. COMPUTER STORAGE REQUIREMENTS AND NUMBER OF OPERATIONS FOR MATRIX INVERSION AND LINEAR SYSTEMS OF EQUATIONS

$n$  = the order of matrix involved  
 $w$  = the number of storage locations required for the computer program of a given technique.

Method	Storage Requirements	Multiplications	Additions
<i>Matrix Inversion</i>			
Jordan-Gauss	$2n^2 + w$	$n^3$	$n^3$
Morris escalator	$n^2 + w$	$\frac{2}{3}n^3$	$\frac{4}{3}n^3$
Gram-Schmidt	$\frac{5}{2}n^2 + w$	$\frac{1}{6}n^3$	$\frac{1}{6}n^3$
Modified matrix	$n^2 + 2n + w$	(a) one element $n^2$	(a) $n^2$
		(b) one row or column $2n^2$	(b) $2n^2$
		(c) whole matrix $2n^3$	(c) $2n^3$
<i>Linear Systems of Equations</i>			
Elimination	$n^2 + n + w$	$n^3/3$	$n^3/3$
Seidel (one iteration)	$n^2 + n + w$	$n^2$	$n^2$
Relaxation (one iteration)	$n^2 + n + w$	$n^2$	$n^2$
Gradient (one iteration)	$n^2 + 5n + 1 + w$	$2n^2$	$2n^2$
Conjugate gradient (one iteration)	Symmetric positive definite		
	$2n^2 + 6n + 2 + w$	$n^2$	$n^2$
	General case		
	$4n^2 + 5n + 2 + w$	$3n^2$	$3n^2$

### 3. EIGENVALUES AND EIGENVECTORS

Murray Mannos

**General Remarks.** The characteristic equation of a matrix together with the corresponding eigenvalues (characteristic values) and eigenvectors (characteristic vectors) plays a fundamental role in the theory of mechanical or electrical vibrations. *Examples:* the flutter vibrations of an airplane wing, the elastic vibrations of a skyscraper or bridge, the buckling of an elastic structure, the transient oscillations of an electric network, and mechanical wave vibrations of molecules and atoms. Similar remarks concerning direct and indirect methods, roundoff errors, etc., apply to the finding of the eigenvalues and eigenvectors as to the inverting of a matrix and the solution of a linear system of equations (see Refs. 8, 9, and 12-14).

In practice, it usually happens that all the eigenvalues of a matrix are distinct. This gives rise to a matrix  $A$  which can be diagonalized by a similarity transformation. Under a similarity transformation the eigenvalues of  $A$  remain invariant. A symmetric matrix can be diagonalized by an orthogonal transformation and similarly a Hermitian matrix can be diagonalized by a unitary transformation. Hence, these types of matrices are frequently singled out for special treatment by somewhat less general methods than apply to the most general type of matrix. Matrices which cannot be diagonalized by means of a similarity transformation or whose eigenvalues are multiple or very closely spaced cause the procedures to become more complex. Results concerning the bounds of eigenvalues are sometimes useful in helping to isolate them. In numerous cases it suffices to find either the dominant or the least eigenvalue.

The elements of the matrix  $A$  will usually be complex elements but they may be confined to be real numbers in some instances. The matrix  $A$  itself will always be of finite order.

Approximations for *digital computer storage requirements* and *number of operations* for finding the eigenvalues and eigenvectors of a matrix cannot be given as readily as in the cases of matrix inversion and the solution of systems of linear equations. This is because the solution of an eigenvalue problem often consists of a number of major segments, such as an iteration, the reduction of a matrix to a direct sum of triple diagonal matrices whose sizes depend on the original matrix, the solution of complex equations, or the evaluation of transcendental functions at specific places, or the consideration of a sequence of Sturm functions. In the case of the triple diagonal method, consideration of the computer aspects has been

broken down in terms of the more important segments. Similarly, computer information for one step of the reduction process for finding eigenvalues of a symmetric matrix by the Jacobi method is also given in Table 6 at the end of this section.

**Characteristic Polynomial.** The characteristic polynomial  $f(x)$  of a matrix  $A$  of order  $n$  over the complex number system may be defined as

$$\begin{aligned}
 (13) \quad \det(\lambda I - A) &= \det \begin{bmatrix} \lambda - a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & \lambda - a_{22} & \cdots & -a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ -a_{n1} & -a_{n2} & \cdots & \lambda - a_{nn} \end{bmatrix} \\
 &= \lambda^n + c_1\lambda^{n-1} + \cdots + c_n \\
 &= f(\lambda).
 \end{aligned}$$

The matrix  $\lambda I - A$  has elements which are polynomials in  $\lambda$  with complex coefficients. The characteristic polynomial may be found by the following methods:

1. The theory of determinants for such matrices is developed along the same line as for those matrices whose elements are real or complex numbers. Hence, the  $\det(\lambda I - A)$  can be expanded along any row or column to obtain its characteristic polynomial. This method is not to be recommended for  $n > 3$ .

2. The coefficients  $c_1, c_2, \dots, c_n$  of the characteristic polynomial in eq. (13) may be obtained from *subdeterminants* of the matrix  $A$  itself:  $c_1 = -(a_{11} + a_{22} + \cdots + a_{nn})$  is the negative of the sum of the diagonal elements of  $A$  or simply the negative of the trace of  $A$ ;  $c_2$  is the sum of the determinants of the  $2 \times 2$  principal minors of  $A$  (i.e., the totality of minors having two of their elements on the diagonal of  $A$ );  $c_3$  is the negative of the sum of the determinants of the  $3 \times 3$  principal minors of  $A$ ,  $\dots$ ,  $c_n = (-1)^n \det A$ . Likewise, this method is not to be recommended for  $n > 3$ .

3. A *finite iterative scheme* based on repeated premultiplication by the matrix  $A$  yields the coefficients  $c_1, c_2, \dots, c_n$  of (13) also. This is the so-called *Souriau-Frame algorithm*.

$$\begin{aligned}
 A_1 &= A, & c_1 &= -\text{trace } A_1, & B_1 &= A_1 + c_1 I \\
 A_k &= AB_{k-1}, & c_k &= -\text{trace } \frac{A_k}{k}, & B_k &= A_k + c_k I \\
 & & & & & (k = 2, 3, \dots, n)
 \end{aligned}$$

4. Another way of finding the characteristic polynomial of a matrix  $A$  is

to build it up one degree at a time by finding the characteristic polynomial of the *upper left-hand minors* of  $A$  in increasing size.

Let  $M_i$  denote the upper left-hand minor of  $A$  of order  $i$ ,  $I_i$  the unit matrix of order  $i$ , and  $f_i(\lambda)$  the characteristic polynomial of  $M_i$ .

Since

$$(\lambda I_i - M_i) \operatorname{adj} (\lambda I_i - M_i) = f_i(\lambda) I_i,$$

it follows from a consideration of the last column that

$$(14) \quad \lambda I_i \begin{bmatrix} b_{1i}(\lambda) \\ b_{2i}(\lambda) \\ \vdots \\ \vdots \\ b_{i-1,i}(\lambda) \\ f_{i-1}(\lambda) \end{bmatrix} = M_i \begin{bmatrix} b_{1i}(\lambda) \\ b_{2i}(\lambda) \\ \vdots \\ \vdots \\ b_{i-1,i}(\lambda) \\ f_{i-1}(\lambda) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ f_i(\lambda) \end{bmatrix},$$

where

$$\begin{bmatrix} b_{1i}(\lambda) \\ b_{2i}(\lambda) \\ \vdots \\ \vdots \\ b_{i-1,i}(\lambda) \\ f_{i-1}(\lambda) \end{bmatrix}$$

is the  $i$ th or last column of  $\operatorname{adj} (\lambda I_i - M_i)$ .

From the first  $i - 1$  rows of the expressions in eq. (14) the coefficients of the polynomials  $b_{ki}(\lambda)$  ( $k = 1, 2, \dots, i - 1$ ) are determined by comparing the various powers of  $\lambda$ . The leading coefficient of each of the  $b_{ki}(\lambda)$  ( $k = 1, 2, \dots, i - 1$ ) is determined by comparing coefficients of  $\lambda^{i-1}$ . Then by using these known coefficients and by comparing coefficients of  $\lambda^{i-2}$ , the second coefficients of each of the polynomials  $b_{ki}(\lambda)$  ( $k = 1, 2, \dots, i - 1$ ) are obtained. By continuing in this way the  $b_{ki}$  ( $k = 1, 2, \dots, i - 1$ ) are completely determined. If the known  $b_{ki}$  ( $k = 1, 2, \dots, i - 1$ ) are now substituted in the resulting equation formed by setting the  $i$ th or last rows of eq. (14) equal, the polynomial  $f_i(\lambda)$  is determined.

One first forms  $f_1(\lambda) = \lambda - a_{11}$  and uses the above technique to find  $f_2(\lambda)$  from  $f_1(\lambda)$ , etc., until finally  $f(\lambda) = f_n(\lambda)$  is obtained from  $f_{n-1}(\lambda)$ .

5. The *method of finite iterations* may be used to obtain a polynomial equation from which some of the eigenvalues of a matrix  $A$  may be obtained. Let  $x \neq 0$  be an arbitrary vector and form  $Ax$ . If  $x$  and  $Ax$  are linearly



of them becomes zero and the process terminates. From  $x_0$  and  $y_0$  one gets  $a_0$  from the left-hand equation of (18) by setting  $i = 1$ . This determines the polynomial  $P_1(\lambda)$  and in turn one gets the vectors  $x_1$  and  $y_1$  from eq. (17) by setting  $i = 2$ . From  $x_1$  and  $y_1$  one gets the coefficients  $a_1$  and  $b_0$  from (18) by setting  $i = 2$ . This in turn determines  $P_2(\lambda)$  from which one determines the vectors  $x_2$  and  $y_2$  by the use of eq. (17) with  $i = 3$ . Continuing in this manner ultimately shows that either the vector  $x_k = 0$  or  $y_k = 0$  for some  $k$ . When this occurs the polynomial  $P_k(\lambda)$ , whose coefficients are now determined, is singled out. The polynomial  $P_k(\lambda)$  as before is a factor of the minimum polynomial  $m(\lambda)$  and coincides with  $m(\lambda)$  if  $k = m$ , and with the characteristic polynomial  $f(\lambda)$  if  $k = n$ .

**Determination of Eigenvalues and Eigenvectors.**  $f(\lambda) = 0$  is called the characteristic equation of the matrix  $A$  and the  $n$  roots of this equation are called the *eigenvalues* of the matrix  $A$ . From eq. (13) it follows that if  $\lambda$  is an eigenvalue of  $A$ , then  $\det(\lambda I - A) = f(\lambda) = 0$  so that the system of linear equations

$$Ax = \lambda x$$

has a nontrivial solution  $x \neq 0$ , and any such solution  $x \neq 0$  is called an *eigenvector* of the matrix  $A$ .

Once the coefficients of the characteristic polynomial have been determined, the characteristic equation can be solved by Graeffe's, Bernoulli's, or any other known method for solving a polynomial equation to obtain the eigenvalues of  $A$ . If  $n$  is fairly high, a large amount of precision in the calculations must be exercised or roundoff error may easily invalidate the results.

Apart from multiplicity it is possible to find the eigenvalues of  $A$  by considering a polynomial equation of lower degree than  $n$ . In this connection the minimum polynomial of the matrix  $A$  will be defined below. By the well-known Cayley-Hamilton theorem it follows that  $f(A) = 0$ . In general, however,  $A$  satisfies polynomial equations of lower degree than  $n = \deg f(\lambda)$ . One denotes by  $m(\lambda)$  that polynomial of lowest degree with leading coefficient 1 such that  $m(A) = 0$ . This polynomial is unique. Furthermore, the minimum polynomial  $m(\lambda)$  divides the characteristic polynomial  $f(\lambda)$ , and each of the eigenvalues of  $A$  is a root of  $m(\lambda) = 0$ . The multiplicity of such a root  $\lambda$  of  $m(\lambda) = 0$  is less than or equal to the multiplicity of  $\lambda$  as a root of  $f(\lambda) = 0$ . Hence, if a certain procedure leads to the construction of the minimum polynomial of  $A$ , it may be sufficient to obtain the necessary information concerning the eigenvalues of  $A$  from  $m(\lambda)$ , which may be of considerably lower degree than the characteristic polynomial of  $A$  and so easier to work with. If one denotes by  $g(\lambda)$  the greatest common divisor of the polynomial elements of  $\text{adj}(\lambda I - A)$  it may be shown that

$$m(\lambda) = \frac{f(\lambda)}{g(\lambda)}.$$

**Direct Methods**

Apart from roundoff errors, the procedures described under the heading of direct methods terminate in a finite number of steps with exact results.

**The Escalator Method.** If the eigenvalues of a symmetric matrix  $A_i$  of order  $i$  are known and distinct and the eigenvectors are also known, the symmetric matrix

$$A_{i+1} = \begin{bmatrix} A_i & a_{i+1} \\ a'_{i+1} & a_{i+1,i+1} \end{bmatrix}$$

obtained by bordering  $A_i$  with an additional row and column also has eigenvalues and eigenvectors which can be found in terms of the eigenvalues and eigenvectors of  $A_i$ . Furthermore, the eigenvalues of  $A_{i+1}$  are distinct and interlace with those of  $A_i$ .

Let  $\lambda_k$  ( $k = 1, 2, \dots, i$ ) denote the eigenvalues of  $A_i$ , and  $u_k$  denote the eigenvectors of  $A_i$ . Then the eigenvalues of  $A_{i+1}$  are obtained by solving the equation

$$\mu - a_{i+1,i+1} = \sum_{k=1}^i \frac{(a'_{i+1}u_k)^2}{\mu - \lambda_k}$$

for the  $i + 1$  values of  $\mu$  which satisfy this equation.

The eigenvector  $v_k$  ( $k = 1, 2, \dots, i + 1$ ) of  $A_{i+1}$  corresponding to  $\mu_k$  ( $k = 1, 2, \dots, i + 1$ ) may be given by

$$v_k = (U(\mu_k I - \Lambda)^{-1}U'a_{i+1}, 1), \quad k = 1, 2, \dots, i + 1,$$

where  $U = (u_1, u_2, \dots, u_i)$ ,

$$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_i \end{bmatrix},$$

and  $U(\mu_k I - \Lambda)^{-1}U'a_{i+1}$  give the first  $i$  components of  $v_k$ .

Starting with the matrix  $(a_{11})$ , which has an eigenvalue of  $a_{11}$  and an eigenvector 1, yields the eigenvalues and eigenvectors of the matrix

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

and continuing step by step finally yields the eigenvalues and eigenvectors of the matrix  $A$  itself. It should be observed that it is necessary to calculate the eigenvalues and eigenvectors of each of the submatrices  $A_i$  ( $i = 2, 3, \dots, n - 1$ ) as well as of the matrix  $A$  itself.

**Triple Diagonal Method.** Let  $A$  be a real symmetric matrix. The method consists of first reducing the matrix  $A$  to a triple diagonal form by means of a specially formed orthogonal transformation to be described below. Then the eigenvalues of the resulting matrix  $S$ , which are the same as those of  $A$ , are obtained with the aid of a Sturm sequence of functions consisting of the determinants of the first principal minors or upper left-hand corner minors of the matrix  $\lambda I - S$ . Then also the eigenvectors of  $S$  associated with an eigenvalue  $\lambda$  are obtained directly from the solution of the homogeneous equations  $(\lambda I - S)x = 0$  because of their exceedingly simple form. From the eigenvectors of  $S$ , one then constructs the eigenvectors of  $A$  itself.

1. In the triple diagonal form of a matrix each element not on the main diagonal, the diagonal just above it, or the diagonal just below it is 0. To obtain this form one attempts by appropriate orthogonal transformations to reduce to 0 all elements of the first row beyond the second column and likewise all elements of the first column beyond the second row. If all these elements are already 0, no manipulation is required. If not, one next looks at the element  $a_{12}$ . If  $a_{12} = 0$  and  $a_{1j}$  is the first nonzero element of the first row following  $a_{12}$ , interchange the second and  $j$ th columns and do likewise with the second and  $j$ th rows. Thus the new element in the first row second column is nonzero. If  $a_{12} \neq 0$  to begin with, look at the element  $a_{13}$ . If  $a_{13} = 0$ , make an exchange similar to the one above so as to bring a nonzero element into its position. If  $a_{13} \neq 0$ , postmultiply  $A$  by the orthogonal matrix  $R_{23}$  and premultiply  $A$  by  $R_{23}^{-1} = R'_{23}$ , where

$$(19) \quad R_{23} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & c & -s & 0 & 0 & \cdots & 0 \\ 0 & s & c & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} U_{23} & 0 \\ 0 & I_{n-3} \end{bmatrix}$$

$$\text{and} \quad c = \left[ 1 + \left( \frac{a_{13}}{a_{12}} \right)^2 \right]^{-1/2}; \quad s = \left( \frac{a_{13}}{a_{12}} \right) c.$$

This amounts to a rotation in the  $x_2x_3$ -plane, and  $c, s$  are the cosine and

sine, respectively, of appropriate angles for making the element  $a'_{13}$  of the matrix  $R_{23}^{-1}AR_{23} = (a'_{ij})$  equal to 0 and hence also making  $a'_{31} = 0$ . Also  $a'_{12}$  has larger magnitude than  $a_{12}$ , and so  $a'_{12} \neq 0$  also. Furthermore,  $a'_{1j} = a_{1j}$  and  $a'_{i1} = a_{i1}$  ( $i, j = 4, \dots, n$ ). If  $a'_{14} \neq 0$ , one may interchange the third and fourth columns and the third and fourth rows of  $(a'_{ij})$  and apply the same type of transformation as before, and give rise to an additional 0 in the first row and column of the newly formed matrix. If  $a'_{14} = 0$ , one looks at  $a'_{15}$ , etc. By continuing in this fashion one forms a new matrix whose first row and first column, except possibly for the first two elements in each case, consist of zeros.

2. The same scheme can next be applied to the resulting submatrix of order  $n - 1$  in the lower right-hand corner. Here instead of  $R_{23}$  one uses  $R_{34}$  where

$$(20) \quad R_{34} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & U_{34} & 0 \\ 0 & 0 & I_{n-4} \end{bmatrix}$$

to reduce all elements in first row and column of the  $(n-1)$ -st order submatrix to zero except possibly for the first two elements. No elements in the first row or column of the  $n$ th order matrix are affected by this.

Continue in this way and, if necessary, finally use

$$R_{n-1,n} = \begin{bmatrix} I_{n-3} & 0 \\ 0 & U_{n-1,n} \end{bmatrix}$$

to effect the final reduction to the following triple diagonal form.

$$S = \begin{bmatrix} a_1 & b_1 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ b_1 & a_2 & b_2 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & b_2 & a_3 & b_3 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & b_3 & a_4 & b_4 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & b_{n-2} & a_{n-1} & b_{n-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{n-1} & a_n \end{bmatrix}.$$

It follows that

$$(21) \quad S = T'AT,$$

where  $T$  consists of a finite product of orthogonal matrices of the type eqs. (19), (20), etc., and also of the type obtained from interchanging two columns of the identity matrix.





pendent eigenvector is associated with  $\lambda_i$ , any nonzero column of  $F(\lambda_i)$ , which has rank 1, is the desired eigenvector associated with  $\lambda_i$ . If there are two linearly independent eigenvectors associated with  $\lambda_i$ ,  $F(\lambda_i) = 0$  and  $F'(\lambda_i)$  is of rank 2, and any two linearly independent columns yield the desired eigenvectors. Lastly, if there are three linearly independent eigenvectors,  $F(\lambda_i) = F'(\lambda_i) = 0$ . But  $F''(\lambda_i)$  is of rank 3, and any three linearly independent columns of  $F''(\lambda_i)$  are the desired eigenvectors associated with  $\lambda_i$ . This procedure can be extended all the way to a root of  $f(\lambda) = 0$  having multiplicity  $n$ .

### Indirect Methods

Here the number of arithmetic operations necessary to arrive at exact answers is infinite. The procedures are iterative and the eigenvalues and eigenvectors of a matrix  $A$  are found without explicitly calculating the characteristic polynomial of  $A$ .

**Iterative Procedures for Hermitian Matrices.** It is easier to handle the case of a Hermitian matrix since it has real eigenvalues; eigenvectors associated with distinct eigenvalues are mutually orthogonal; and, because it can be diagonalized, the multiplicity of each eigenvalue  $\lambda$  equals the number of linearly independent eigenvectors associated with  $\lambda$ .

Assume, for the time being, that the eigenvalues of a given Hermitian matrix  $A$  are distinct. Also all the eigenvalues of  $A + \mu I$ , which is also Hermitian, can be made positive by picking  $\mu$  sufficiently large so that there is no restriction in assuming that the matrix  $A$  has a single dominant eigenvalue, i.e., an eigenvalue whose absolute value is greater than that of any other eigenvalue of  $A$ . One first concentrates attention upon a method of finding the dominant eigenvalue and its associated eigenvector. Several methods are then available for finding the remaining eigenvalues of  $A$ .

The procedure starts with an initial vector  $x_0$  and by repeated premultiplication of  $A$  builds up the sequence of vectors

$$(24) \quad x_p = Ax_{p-1} = A^p x_0 \quad (p = 1, 2, \dots).$$

In the nonexceptional case for  $p$  sufficiently large, the direction of the vector  $x_p$  will approach the direction of the eigenvector  $u_1$  associated with the dominant eigenvalue  $\lambda_1$ . In the exceptional cases,  $x_p$  will approach either some  $u_i$  associated with the eigenvalue  $\lambda_i$  ( $i = 2, \dots, n$ ) or else 0. The latter rarely happens, but at any rate, an  $x_0$  can be easily picked so that the former case will apply. Again if  $p$  is sufficiently large, the ratio of the  $i$ th component ( $i = 1, 2, \dots, n$ ) of  $x_{p+1}$  to that of  $x_p$  can be made arbitrarily close to the dominant eigenvalue. The closeness with which these ratios agree may be regarded as a measure of the accuracy of the

approximation to  $\lambda_1$ . An error made during the course of the computation of  $x_p$  will not lead to an erroneous result since subsequent multiplication by  $A$  will pull the computation back into line.

One may alternatively calculate  $\lambda_1$  by means of a ratio of numbers as defined below. Let

$$a_p = \bar{x}'_0 x_p \quad (p = 1, 2, \dots);$$

then

$$\lambda_1 = \lim_{p \rightarrow \infty} \frac{a_{p+1}}{a_p}.$$

If next one desires to find the minimum eigenvalue  $\lambda_n$  of  $A$  together with its associated eigenvector, one may consider the matrix  $cI - A$  where  $c > \lambda_1$ . The matrix  $cI - A$  is Hermitian, and the same techniques may be applied to it to find its maximum eigenvalue and its associated eigenvector. To get the minimum eigenvalue  $\lambda_n$  of  $A$ , one simply changes the sign of the maximum eigenvalue of  $cI - A$  and adds  $c$ . The eigenvector associated with the maximum eigenvalue of  $cI - A$  is also the eigenvector associated with the minimum eigenvalue  $\lambda_n$  of  $A$ .

After  $\lambda_1$  and  $u_1$  have been calculated, the determination of the remaining  $n - 1$  eigenvalues and their associated eigenvectors of the  $n$ th order matrix  $A$  may be done in terms of a matrix whose order is  $n - 1$  instead of  $n$ . If the normalized form of  $u_1$  is denoted by  $u_1^*$ , i.e.,  $u_1^* = u_1/|u_1|$ , from the vector  $u_1^*$  a unitary matrix  $U$  may be constructed so that

$$\bar{U}'AU = \begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix},$$

where  $A_1$  is a Hermitian matrix of order  $n - 1$  whose eigenvalues  $\lambda_2, \lambda_3, \dots, \lambda_n$  are the  $n - 1$  remaining unknown eigenvalues of  $A$ . The dominant eigenvalue  $\lambda_2$  of  $A_1$  and its associated eigenvector  $v_2$  can be found as previously. The eigenvector  $u_2^*$  associated with the eigenvalue  $\lambda_2$  of  $A$  is the vector  $U \begin{pmatrix} 0 \\ v_2 \end{pmatrix}$ . The following construction of the unitary matrix  $U$  is due to Feller and Forsythe. One writes  $u_1^*$  as follows

$$u_1^* = \begin{pmatrix} a \\ z \end{pmatrix},$$

where  $a$  is a complex number and  $z$  is an  $n - 1$  dimensional vector with complex components. Then

$$U = \begin{bmatrix} a & -\bar{z}' \\ z & I_{n-1} - kzz\bar{z}' \end{bmatrix},$$

where  $k = (1 - \bar{a})/(1 - \bar{a}a)$ .

Next one replaces  $A_1$  by the matrix

$$\begin{bmatrix} \lambda_2 & 0 \\ 0 & A_2 \end{bmatrix}$$

where  $A_2$  is a Hermitian matrix of order  $n - 2$  having eigenvalues  $\lambda_3, \dots, \lambda_n$ , and then one repeats the previous step. This is continued until all eigenvalues and eigenvectors of  $A$  are obtained.

Another way to find the eigenvalues  $\lambda_2, \dots, \lambda_n$  and their associated eigenvectors, once  $\lambda_1$  and  $u_1^*$  are known, is to form the new Hermitian matrix

$$(25) \quad A_1 = A - \lambda_1 u_1^* \overline{u_1}'$$

of order  $n$  also. The eigenvalues of  $A_1$  are  $0, \lambda_2, \dots, \lambda_n$ . The known eigenvector  $u_1^*$  is associated with the eigenvalue  $0$  of  $A_1$ ; while the unknown eigenvector  $u_i^*$  is associated with the eigenvalue  $\lambda_i$  ( $i = 2, \dots, n$ ) of  $A_1$  as well as of  $A$ . Thus the dominant eigenvalue  $\lambda_2$  of  $A_1$  and its associated eigenvector  $u_2^*$  can be found as before by forming powers of  $A_1$  instead of powers of  $A$  as in eq. (24).

Next one forms the Hermitian matrix

$$A_2 = A_1 - \lambda_2 u_2^* \overline{u_2}'$$

of order  $n$  which has eigenvalues  $0, 0, \lambda_3, \dots, \lambda_n$ , and the unknown  $u_i^*$  is associated with the  $\lambda_i$  ( $i = 3, \dots, n$ ) of  $A_2$  as well as of  $A$ . Thus one obtains  $\lambda_3$  and its associated eigenvector  $u_3^*$ . Again one continues in this fashion until all eigenvalues of  $A$  and their associated eigenvectors are found.

**Multiple Roots.** So far the possibility of multiple roots has not been considered. Suppose, as before, one starts with  $x_0$  and builds up sequence (24), one obtains as before an eigenvector associated with  $\lambda_1$ . A distinct starting vector  $y_0$  may be selected to build up a new sequence which will again lead to the eigenvalue  $\lambda_1$ . But it may happen that  $y_0$  leads to an eigenvector which is linearly independent of the one to which  $x_0$  leads. In this case  $\lambda_1$  is a multiple eigenvalue. If  $y_0$ , as in the case of distinct eigenvalues, leads only to an eigenvector which is linearly dependent or simply a multiple of the eigenvector to which  $x_0$  leads, then  $\lambda_1$  is a simple eigenvalue. If  $x_0$  and  $y_0$  lead to linearly independent eigenvectors, and a third arbitrary vector  $z_0$  leads to an eigenvector which is linearly dependent upon the first two eigenvectors,  $\lambda_1$  is an eigenvalue of multiplicity 2; whereas, if  $z_0$  leads to an eigenvector linearly independent of the first two calculated eigenvectors,  $\lambda_1$  is at least of multiplicity 3. One can continue this process for eigenvalues of higher multiplicity also.

Let  $\lambda_1$  be a root of multiplicity 2. Then in the two-dimensional vector space generated by the two linearly independent eigenvectors obtained one may select  $u_1^*$ ,  $u_2^*$ , which are orthogonal and of unit length, and which are eigenvectors associated with  $\lambda_1$ . By starting with  $u_1^*$  and  $u_2^*$  one may similarly, as before, build up a unitary matrix  $U$  such that

$$U^{-1}AU = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & A_2 \end{bmatrix},$$

where  $A_2$  is a Hermitian matrix of order  $n - 2$  containing the remaining eigenvalues  $\lambda_3, \dots, \lambda_n$  of  $A$ . Similarly, one proceeds in the case of multiple eigenvalues, as outlined before.

A number of additional variations for obtaining the eigenvalues and eigenvectors of  $A$  is possible.

**Iterative Process for General Type Matrices.** If the matrix  $A$  can be diagonalized, the method of successively premultiplying by  $A$  applies, with some small appropriate modifications, to this case as well as to the case of the Hermitian matrix. No longer are the eigenvalues of  $A$  necessarily real. There may be several distinct dominant eigenvalues. The eigenvectors of  $A$  can no longer be assumed mutually orthogonal. In order to get around this situation, one introduces the concept of row eigenvectors as well as column eigenvectors. Associated with each eigenvalue  $\lambda_i$  of  $A$  is the row eigenvector  $u^{(i)}$ , where  $u^{(i)}A = \lambda_i u^{(i)}$ , and the column eigenvector  $u_i$ , where  $Au_i = \lambda_i u_i$  ( $i = 1, 2, \dots, n$ ). In this case

$$u^{(i)}u_j = \delta_{ij},$$

where

$$\delta_{ij} = 0, \quad i \neq j,$$

$$\delta_{ij} = 1, \quad i = j.$$

Here  $u^{(i)}$  and  $u_i$  ( $i = 1, 2, \dots, n$ ) need not be unit vectors but only  $u^{(i)}u_i = 1$  ( $i = 1, 2, \dots, n$ ).

One again starts with an arbitrary initial vector  $x_0$  and forms the sequence  $x_p$  of eq. 24. A unique dominant eigenvalue and its associated eigenvector are found exactly as before. Whereas, in the case of a Hermitian matrix one forms the matrix  $A_1$  as in eq. (25) in order to study the remaining eigenvalues and their associated eigenvectors, one now forms the matrix

$$A_1 = A - \lambda_1 u^{(1)}u_1.$$

Finding the eigenvalues in the case where several eigenvalues are dominant is more complicated, as these are not computed as a simple ratio but

rather as described below. Suppose that  $x_{1p}, x_{2p}, \dots, x_{kp}$  are  $k$  linearly independent vectors obtained from the sequence (24). For  $p$  sufficiently large, these are arbitrarily close to the actual eigenvectors. It is desired to find the eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_k$  associated with these. Take  $z$  as an arbitrary vector and form

$$a_{ip} = z'x_{ip} \quad (i = 1, 2, \dots, k),$$

then

$$\det \begin{bmatrix} 1 & a_{1p} & a_{2p} & \cdots & a_{kp} \\ \lambda & a_{1,p+1} & a_{2,p+1} & \cdots & a_{k,p+1} \\ \lambda^2 & a_{1,p+2} & a_{2,p+2} & \cdots & a_{k,p+2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \lambda^k & a_{1,p+k} & a_{2,p+k} & \cdots & a_{k,p+k} \end{bmatrix} = 0$$

has  $k$  roots which are close approximations to the eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_k$ .

The great majority of matrices appearing in applications have distinct eigenvalues and so can be diagonalized. Therefore, the method of iterating by premultiplication of a given matrix is applicable. In the rare case in which  $A$  has a root of multiplicity  $r$  and whose associated eigenvectors number less than the full complement of  $r$ , it is not possible to diagonalize  $A$ . Nevertheless, even in this case, it is still possible to use this iteration scheme to find the dominant eigenvalue and the associated eigenvector of matrix  $A$  having but a single dominant eigenvalue. One must, however, consider the linear dependence of a finite number of successive  $x_p$ 's in sequence (24) for  $p$  sufficiently large to obtain the dominant eigenvalue  $\lambda_1$ . The associated eigenvector may be obtained as a linear combination of a finite number of the  $x_p$ 's whose components contain powers of  $\lambda_1$ .

**Jacobi Method.** The technique applies to Hermitian and so to real symmetric matrices too. The method hinges on the fact that a  $2 \times 2$  Hermitian matrix

$$H = \begin{bmatrix} a_{11} & ae^{i\psi} \\ ae^{-i\psi} & a_{22} \end{bmatrix}, \quad a > 0$$

can be reduced to diagonal form  $D$  by a unitary transformation  $U^{-1}HU = D$ , where

$$(26) \quad U = \begin{bmatrix} e^{i\psi/2} \cos \theta & -e^{i\psi/2} \sin \theta \\ e^{-i\psi/2} \sin \theta & e^{-i\psi/2} \cos \theta \end{bmatrix}$$

and  $\theta$  is an angle in the first quadrant which satisfies  $\tan 2\theta = 2a/(a_{11} - a_{22})$ . If  $\psi = 0$ , i.e.,  $H$  is real symmetric, the matrix (26) reduces to the familiar form

$$U = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix},$$

which corresponds to a rotation in the plane.

If the  $n$ th order Hermitian matrix  $A = (a_{ij})$  is written in the form

$$A = \begin{bmatrix} H & A_1 \\ \bar{A}'_1 & A_{22} \end{bmatrix},$$

then the unitary matrix

$$(27) \quad U_1 = \begin{bmatrix} U & 0 \\ 0 & I \end{bmatrix},$$

where  $U$  is the matrix (26) and transforms  $A$  into a matrix  $B = (b_{ij})$ , where  $b_{12} = b_{21} = 0$ ; furthermore, the sum of the squares of the diagonal elements of  $B$  exceeds the corresponding sum of  $A$  by the positive quantity  $2a^2$ . If it is desired to transform  $A$  into a matrix  $B$  such that  $b_{ij} = b_{ji} = 0$ , the elements of the matrix  $U$  in eq. (26) must be positioned in the  $i$ th and  $j$ th rows as well as in the  $i$ th and  $j$ th columns of  $U_1$  in eq. (27).

One might hope that after applying the product of a finite number of the above unitary transformations one might reduce the matrix  $A$  to diagonal form, in which case the sum of the squares of the diagonal elements will have the maximum possible value. Unfortunately, this is not true, as some of the elements, which have previously been reduced to zero, will not remain so while some additional elements are likewise being reduced to zero.

The procedure is to reduce to zero a pair of off-diagonal elements of greatest modulus. It is the infinite product of all these transformations which will reduce  $A$  to diagonal form and whose diagonal contains the eigenvalues of  $A$ . The infinite product of unitary matrices of the type (27) converges to a matrix whose columns are the eigenvectors of the matrix  $A$ .

### Eigenvalues of Special Matrices

The types of eigenvalues to which certain important classes of matrices give rise are worth noting.

<i>Matrix A</i>	<i>Every Eigenvalue <math>\lambda_i</math> of A</i>
(a) Real and symmetric	(a) Real
(b) Real, symmetric, and positive definite	(b) Real and positive
(c) Real, symmetric and positive semidefinite	(c) Real and non-negative
(d) Orthogonal	(d) $ \lambda_i  = 1$ for every $i$

In (a), (b), and (c), if a real symmetric matrix is replaced by a Hermitian matrix, the conclusions still remain valid. In (d) if  $A$  is unitary, the conclusion drawn there still holds.

Some additional properties concerning dominant roots of important classes are listed below:

(i) If  $A$  is real and symmetric, the maximum eigenvalue  $\lambda_{\max}$  is given by

$$\lambda_{\max} = \max_{x \neq 0} \frac{x'Ax}{x'x},$$

and the minimum eigenvalue  $\lambda_{\min}$  is given by

$$\lambda_{\min} = \min_{x \neq 0} \frac{x'Ax}{x'x}.$$

(ii) If  $A$  is a real positive matrix (i.e.,  $A$  has positive elements),  $\lambda_{\max}$  is a real number.

### Bounds on Eigenvalues

It is often a helpful guide to establish bounds for the eigenvalues of a matrix at the outset, as this may influence the procedure. It is extremely advantageous when this leads to the isolation of some of the eigenvalues of a given matrix. Some of the criteria for determining bounds are easily applied. A number of such results will be stated and in some cases additional information will be given concerning the associated eigenvectors. First, the case of matrices with complex elements will be treated, and subsequently this will be specialized to matrices with positive and also matrices with non-negative elements. However, when results on bounds apply to a large class of matrices, the bounds cannot be expected to be as sharp as those applying to a smaller more specialized class of matrices. The following cases are of interest.

1. Let  $A$  be an *arbitrary matrix* of order  $n$  with complex elements. Then

$$|\lambda| \leq n^M,$$

where  $\lambda$  is any eigenvalue of  $A$ , and  $M$  is the maximum of the moduli of the elements  $a_{ij}$  ( $i, j = 1, \dots, n$ ) of  $A$ . This result is due to Hirsch in 1902.

2. Let

$$R_i = \sum_{j=1}^n |a_{ij}| \quad \text{and} \quad T_j = \sum_{i=1}^n |a_{ij}|.$$

Also let  $R = \max R_i$  ( $i = 1, \dots, n$ ) and  $T = \max T_j$  ( $j = 1, \dots, n$ ). Then

$$|\lambda| \leq \min (R, T).$$

A number of variations on these two bounds (1) and (2) exists. Some of these variations are a bit sharper, but these are simple to apply and will be sufficient for the purposes at hand.

3. Let  $P_i$  denote the sum of the moduli of the off-diagonal elements of the  $i$ th row of the matrix  $A$  and  $Q_j$  the sum of the moduli of the off-diagonal elements of the  $j$ th column of  $A$ . That is,

$$P_i = \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \quad \text{and} \quad Q_j = \sum_{\substack{i=1 \\ i \neq j}}^n |a_{ij}|.$$

Then a result due to Levy and Hadamard states that each eigenvalue of  $A$  lies in at least one of the circles

$$(28) \quad |z - a_{ii}| \leq P_i \quad (i = 1, \dots, n)$$

and in at least one of the circles

$$|z - a_{jj}| \leq Q_j \quad (j = 1, \dots, n).$$

In other words, if one takes the diagonal element  $a_{ii}$  and draws a circle with  $a_{ii}$  as the center and  $P_i$  ( $i = 1, \dots, n$ ) as radius, all the eigenvalues of  $A$  will be trapped in these  $n$  circles. A similar remark applies to the  $n$  circles with the  $Q_j$  ( $j = 1, \dots, n$ ) as radii. It is to be noted that an eigenvalue of  $A$  may be in several of the  $n$  circles.

4. An interesting offshoot of this result is the following: If one of the  $n$  circles is isolated from the remaining  $n - 1$  circles, that is, has no point in common with the remaining  $n - 1$  circles, exactly one eigenvalue of  $A$  will be found in the isolated circle. More generally Geršgorin showed that when  $m$  circles intersect in a connected region isolated from the remaining  $n - m$  circles, the connected region thus formed contains exactly  $m$  eigenvalues of  $A$ .

5. The following results concerning the *number of associated eigenvectors* is noteworthy. If an eigenvalue  $\lambda$  of the matrix  $A$  lies in only one of  $n$  circles (28),  $\lambda$  has only one linearly independent eigenvector associated with it. This result is due to Taussky (Ref. 15). Stein has shown that if an eigenvalue  $\lambda$  has associated with it  $m \leq n$  linearly independent eigenvectors,  $\lambda$  lies in at least  $m$  of the circles (28).

6. Before passing to the case of positive and non-negative matrices, it is worth noting a result of Frobenius which gives a connection between the eigenvalues of a matrix with complex elements and a dominating matrix with non-negative elements. Let  $B = (b_{ij})$  be a matrix with complex elements and  $A = (a_{ij})$  be a matrix with non-negative elements such that  $|b_{ij}| \leq a_{ij}$  ( $i, j = 1, \dots, n$ ). Then the characteristic circle of  $A$  contains the characteristic circle of  $B$ . (The *characteristic circle* of a matrix is the

smallest circle about the origin containing the eigenvalues of the given matrix.)

7. Turning attention next to *real matrices with non-negative elements*, one can draw some additional and sharper conclusions. If  $A$  is a matrix whose elements  $a_{ij} \geq 0$  ( $i, j = 1, \dots, n$ ) then (a)  $A$  has a real eigenvalue  $\lambda_d \geq 0$  which is dominant (there may be other dominant eigenvalues), (b)  $\lambda_d$  has an associated eigenvector  $x \geq 0$ , i.e., all components of  $x$  are non-negative, and (c)  $\lambda_d$  does not decrease when an element of  $A$  increases.

The above results are due to Herstein and Debreu, who paralleled for the case of non-negative matrices the results of Frobenius given below.

8. These results grow sharper when one further restricts the matrix  $A$  to be indecomposable. A non-negative matrix  $A$  is called *indecomposable* if  $A$  cannot be transformed to a matrix of the form

$$\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$$

by the same permutations of rows and columns where  $A_{11}$  and  $A_{22}$  are square submatrices of  $A$ .

If  $A$  is a non-negative indecomposable matrix, then (a)  $A$  has a real simple eigenvalue  $\lambda_d > 0$  which is dominant; (b)  $\lambda_d$  has an associated eigenvector  $x > 0$ , i.e., all the components of  $x$  are strictly positive; and (c)  $\lambda_d$  increases when an element of  $A$  increases. These important results were first demonstrated by Frobenius nearly a half century ago.

9. If the matrix  $A$  is still further restricted so that all its *elements are positive*, that is,  $a_{ij} > 0$  ( $i, j = 1, \dots, n$ ), then the statement (a) above can be strengthened to include the fact that  $\lambda_d$  is the only dominant eigenvalue of  $A$ .

10. Again, suppose  $A$  is a positive matrix and let

$$R_i = \sum_{j=1}^n a_{ij} \quad (i = 1, \dots, n), \quad R = \max \{R_1, R_2, \dots, R_n\},$$

$$\text{and } r = \min \{R_1, R_2, \dots, R_n\}.$$

Frobenius first noted that

$$r \leq \lambda_d \leq R.$$

Also  $\lambda_d = r = R$  if and only if all  $R_i$  are equal; otherwise, the inequality

$$r < \lambda_d < R$$

holds. Suppose that not all the  $R_i$  ( $i = 1, \dots, n$ ) are equal and let

$$\kappa = \min a_{ij},$$

$$\delta = \max_{R_i < R_j} \{R_i/R_j\},$$

and

$$\sigma = \sqrt{(r - \kappa)/(R - \kappa)}$$

so that  $\kappa > 0$ ,  $\delta < 1$ , and  $\sigma < 1$ . Ledermann improved the bounds on Frobenius' result as follows

$$r + \kappa \left( \frac{1}{\sqrt{\delta}} - 1 \right) < \lambda_d < R - \kappa(1 - \sqrt{\delta}),$$

and Ostrowski further sharpened the bounds with the inequalities

$$r + \kappa \left( \frac{1}{\sigma} - 1 \right) \leq \lambda_d \leq R - \kappa(1 - \sigma).$$

In fact, the right-hand side of Ostrowski's inequalities applies to matrices with complex elements when in the definitions of  $R$  and  $\kappa$  one uses the modulus of the elements. More recently, Brauer announced further improvement of the above bounds, stating that the best possible bounds have been attained. That is, in order to get sharper bounds one would have to restrict further the class of positive matrices.

11. Some specialized examples of non-negative matrices are the *stochastic matrices* and the *oscillation matrices*. The eigenvalues of the former play an important role in the theory of stochastic processes while the latter type matrices are applicable in the theory of small oscillations of mechanical systems.

The matrix  $A = (a_{ij})$  is called *stochastic* if  $a_{ij} \geq 0$  ( $i, j = 1, \dots, n$ ) and if

$$\sum_{j=1}^n a_{ij} = 1 \quad (i = 1, \dots, n).$$

If  $a_{ij} > 0$  ( $i, j = 1, \dots, n$ ), the matrix  $A$  is called a positive stochastic matrix. All the eigenvalues of a stochastic matrix lie within or on the boundary of the unit circle. Also  $\lambda = 1$  is a dominant eigenvalue of any stochastic matrix. Previous results on non-negative and positive matrices may be directly applied to stochastic matrices.

The matrix  $A$  of order  $n$  is said to be *completely non-negative* (*completely positive*) if all minors of all orders from 1 to  $n$  of  $A$  are *non-negative* (*positive*).

If  $A$  is completely non-negative and there exists a positive integer  $k$  such that  $A^k$  is completely positive, then  $A$  is said to be an *oscillation matrix*. A non-negative matrix  $A$  will specialize to an oscillation matrix if and only if  $\det A \neq 0$ ,  $a_{i,i+1} > 0$  and  $a_{i+1,i} > 0$  ( $i = 1, \dots, n - 1$ ). The eigenvalues of an oscillation matrix have the interesting property that they are all strictly positive and simple.

For an extensive bibliography on the bounds of eigenvalues, see Ref. 15.

TABLE 6. COMPUTER STORAGE REQUIREMENTS AND NUMBER OF OPERATIONS FOR FINDING EIGENVALUES AND EIGENVECTORS

$n$  = order of matrix involved  
 $w, w', w''$  = program storage requirements

Method	Storage Requirements	Multipli- cations	Addi- tions
<i>Triple Diagonal</i>			
Triple diagonal form $S$	$n^2/2 + n/2 + w$	$\frac{4}{3}n^3$	$\frac{2}{3}n^3$
Eigenvalues of $A$ <sup>a</sup>			
Eigenvectors of $S$	$2n - 1 + w'$	$3n^2$	$2n^2$
Eigenvectors of $A$	$2n^2 + w''$	$2n^3$	$n^3$
Total for eigenvalues and eigenvectors <sup>b</sup>	$\frac{5}{2}n^2 + \frac{5}{2}n - 1 + w'''$	$\frac{10}{3}n^3$	$\frac{5}{3}n^3$
<i>Jacobi (Symmetric matrix)</i>			
Eigenvalues (one step of reduc- tion)	$n^2 + 4 + w$	$4n$	$2n$

<sup>a</sup> In finding the eigenvalues of  $A$ , the number of operations depends on the stipulated requirements for accuracy. If  $n$  is sufficiently large, the number of operations required to find the eigenvalues, once the matrix is in triple diagonal form, is negligible compared with the number of operations required to reduce the original matrix to triple diagonal form.

<sup>b</sup>  $w'''$  is the sum of  $w, w', w''$ , and the number of cell locations used in finding the eigenvectors of  $A$ .

#### 4. DIGITAL TECHNIQUES IN STATISTICAL ANALYSIS OF EXPERIMENTS

Joseph M. Cameron

**Introduction.** In scientific experiments a variable is measured under several different conditions with a view to assessing the effect of these conditions on the variable under study. There may be factors present in the measurement process which, if not balanced out or their effect reduced by randomization or replication, may invalidate the estimates of the effects the experiment seeks to measure. The branch of statistics called the *design of experiments* is concerned with the construction of experimental arrangements that permit the balancing out of such extraneous factors and at the same time minimizing (for a given number of observations) the uncertainties in the estimates of the effects under study.

In most applications the analysis required is the usual *least squares* analysis for estimating the parameters postulated to represent the data. In a designed experiment the normal equations that arise in the estimation of the parameters take on a particularly simple form and the calculations have been systematized and given the name *analysis of variance*. *Example.* Consider a set of measurements  $x_1, x_2, \dots, x_n$  all postulated to be estimates of a single quantity. The least squares estimate for that quantity is, of course, the average  $\bar{x} = \Sigma x_i/n$ . One can also compute from the data a measure of the dispersion of the results about this average. Perhaps the most common such measure is the *standard deviation*,  $\sqrt{\Sigma(x_i - \bar{x})^2/(n - 1)}$ . In the analysis of variance one deals not with the standard deviation but rather with its square, which is a quadratic form in the deviations divided by the number of independent deviations, called the *number of degrees of freedom*.

The analysis of variance in its general form is a technique for (a) computing estimates of the parameters involved in the problem and (b) computing the value of quadratic forms, called sums of squares, assignable to certain groupings of the parameters, each sum of squares carrying with it a certain number of degrees of freedom (the rank of the quadratic form). Thus in the case of  $k$  averages each based on  $n$  measurements, the parameters to be estimated are the grand average and the  $(k - 1)$  independent deviations of the individual averages about this grand average. Three sums of squares are to be calculated: one for the grand average (with one degree of freedom), one for the deviation of the individual averages about the grand average [with  $(k - 1)$  degrees of freedom], and one for the deviations of the observations about their own group averages [with  $k(n - 1)$  degrees of freedom].

Several examples of the analysis of variance are presented to illustrate the different techniques of computation that are available. The advantage of one over another probably depends on the nature of the computing device used.

The availability of modern high-speed digital computers makes it feasible to analyze experimental data involving a much greater number of factors, each factor occurring at more levels than would otherwise be the case. The types of calculations described above and in the succeeding pages, because of their systematic nature, lend themselves particularly well to treatment on automatic digital computers.

**Analysis of Factorial Designs Using Hartley Method.** An experiment in which the effects of several factors on a variable are studied by making measurements at all possible combinations of the several states or levels for each of the factors is called a factorial experiment. *Example.* Four temperatures of heat treating can be combined with three time periods

to give rise to twelve conditioning treatments for some alloy. This would be a factorial design with two factors, one at four levels and the other at three levels.

The most general method for the analysis of factorials was developed by Hartley (Ref. 22). His method depends on three operators which he has labeled  $\Sigma$ ,  $D$ , and  $( )^2$ , defined as follows:

- $\Sigma_t$  Sum over all levels  $t = 1, 2, \dots, T$  for each combination of the other subscripts.
- $D_t$  Difference between  $T$  times the original values and the total in the set  $\Sigma_t$  to which the original value contributed.
- $( )^2$  Sum of squares of items indicated in the parentheses.

**Procedure.** The use of this technique will be *illustrated* for a two-factor factorial having one factor at  $k$  levels and the other at  $n$  levels. Denote by  $x_{ij}$  the observation at the  $i$ th level of the first factor and  $j$ th level of the second factor. Let  $x_{.j}$  denote the set of sums  $\sum_i x_{ij} = x_{.j}$ , there being  $n$  such sums. In Table 7 the plan of the calculations is shown. Table 8 shows the analysis of variance table derived from the results of Table 7.

TABLE 7. PLAN OF CALCULATIONS USING HARTLEY TECHNIQUE, TWO-FACTOR FACTORIAL EXPERIMENT

		$A_1$	$A_2$	$\dots$	$A_k$		
Level of fac- tor B	$B_1$	$x_{11}$	$x_{21}$	$\dots$	$x_{k1}$		
	$B_2$	$x_{12}$	$x_{22}$	$\dots$	$x_{k2}$		
	$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$		
	$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$		
	$B_n$	$x_{1n}$	$x_{2n}$	$\dots$	$x_{kn}$		
	$\Sigma_j$	$x_{.1}$	$x_{.2}$	$\dots$	$x_{.k}$	$\Sigma_i \Sigma_j$	$x_{..}$
	$D_j$	$nx_{11} - x_{.1}$	$nx_{21} - x_{.2}$	$\dots$	$nx_{k1} - x_{.k}$	$\Sigma_i D_j$	$nx_{.1} - x_{..}$
		$nx_{12} - x_{.1}$	$nx_{22} - x_{.2}$	$\dots$	$nx_{k2} - x_{.k}$		$nx_{.2} - x_{..}$
		$\dots$	$\dots$	$\dots$	$\dots$		$\dots$
		$nx_{1n} - x_{.1}$	$nx_{2n} - x_{.2}$	$\dots$	$nx_{kn} - x_{.k}$		$nx_{.n} - x_{..}$
	$D_i \Sigma_j$	$kx_{.1} - x_{..}$			$kx_{.k} - x_{..}$		
	$D_i D_j$	$kx_{11} - kx_{.1} - nx_{.1} + x_{..}$			$kx_{k1} - kx_{.k} - nx_{.1} + x_{..}$		
		$\dots$			$\dots$		
		$kx_{1n} - kx_{.1} - nx_{.n} + x_{..}$			$kx_{kn} - kx_{.k} - nx_{.n} + x_{..}$		

The estimates of the parameters are obtained by dividing the entries in the sets  $\Sigma_i \Sigma_j$ ,  $\Sigma_i D_j$ ,  $D_i \Sigma_j$ , and  $D_i D_j$  by  $nk$  giving in that order the grand

TABLE 8. ANALYSIS OF VARIANCE FOR TWO-FACTOR FACTORIAL EXPERIMENT

	No. of Items	Sum of Squares	Degrees of Freedom	Sum of Squares Is Associated with:
$(\Sigma_i \Sigma_j)^2$	1	$(\Sigma_i \Sigma_j)^2 / nk$	1	Grand average
$(\Sigma_i D_j)^2$	$n$	$(\Sigma_i D_j)^2 / n(nk)$	$n - 1$	Effect of different levels of factor $B$
$(D_i \Sigma_j)^2$	$k$	$(D_i \Sigma_j)^2 / k(nk)$	$k - 1$	Effect of different levels of factor $A$
$(D_i D_j)^2$	$nk$	$(D_i D_j)^2 / (nk)^2$	$(n - 1)(k - 1)$	Interaction: lack of constancy between levels of $A$ as level of $B$ is varied
		$\Sigma \Sigma x_{ij}^2$	$nk$	Total (for check)

average, differences among levels of factor  $B$ , differences among levels of factor  $A$ , and the differences due to lack of constancy of the different levels of factor  $A$  as the level of factor  $B$  is changed. This technique can be extended to cover the case of three or more factors by using the basic operations of  $\Sigma$ ,  $D$ , or  $( )^2$  and is adaptable to other designs as well (see Ref. 22).

An alternate procedure necessary when the experiment is run in blocks containing only a fraction of the total number of observations or when a fractional replication design is used is based on the technique described in Ref. 23, and is discussed below. Still another procedure is given in Ref. 16 based on the computation of individual degrees of freedom with orthogonal polynomials tabled in Refs. 20 and 21.

**Balanced Incomplete Blocks.** When there are more objects or treatments than can be compared under the same conditions, i.e., on a given batch of material, in a given time period, or other factor which limits the uniformity of conditions to a few tests, it is necessary to schedule the measurements so that all comparisons of interest may be estimated from the data. The class of designs constructed for such a case is called incomplete block designs, the block being the group of tests within which the environmental or other factor is assumed not to change. The analysis of these block designs will be illustrated for the case of the balanced incomplete block design (see Refs. 16-23).

Observations have index  $x_{bkt}$  referring to  $B$  blocks with  $K$  units per block and  $T$  treatments with  $R$  repetitions of each. The data are entered so that the observations from the first block come first, followed by those from the second block, etc.

*Step I.* Compute total sum of squares of original values,  $\Sigma x_{bk}^2$ .

Step II. Consider only indices  $b$  and  $k$ .

Operation	Number of Items	Result	( ) <sup>2</sup> Applied to Result Gives BK Times
$\Sigma_k$	$B$	$x_b.$	
$D_k$	$BK$	$Kx_{bk} - x_b.$	
$\Sigma_b \Sigma_k$	$1$	$x_{..}$	Correction factor
$D_b \Sigma_k$	$B$	$Bx_{b.} - x_{..}$	Unadjusted blocks sums of squares

Step III. Now consider only indices  $t$  and  $r$ . The values of  $D_k$  are now rearranged into  $T$  groups with  $R$  values each so that the  $R$  values corresponding to the first treatment come in a group followed by a similar grouping for each of the remaining treatments. Call these values  $d_{tr}$  and denote operations after rearrangement with asterisk.  $\Sigma_r^* D_k$  results in

$$d_{t.} = Kx_{t.} - B_t$$

$$(\Sigma_r^* D_k)^2 = \frac{K(K-1)TR}{(T-1)} \times \text{sum of squares for treatments (adjusted),}$$

where  $B_t =$  sum of block totals for blocks containing treatment  $t$ .

Analysis of Variance	Sum of Squares	Degrees of Freedom
Total	$\Sigma x_{bk}^2 - x_{..}^2 / BK$	$BK - 1$
Blocks (unadjusted)	$(D_b \Sigma_k)^2 / BK$	$B - 1$
Treatments (adjusted)	$\frac{(T-1)(\Sigma_r^* D_k)^2}{K(K-1)TR}$	$T - 1$
Error	By subtraction	$BK - B - T + 1$

**Analysis of Factorials by Using Relations among the Indices Associated with the Treatments.** To illustrate the method assume there are three factors  $A, B,$  and  $C$  having levels  $n + 1, n + 1,$  and  $n + 1$  respectively. Each observation is tagged with an index  $x_1 x_2 x_3$ , where  $x_1 = 0, 1, \dots, n, x_2 = 0, 1, \dots, n,$  and  $x_3 = 0, 1, \dots, n,$  where  $n$  is a prime.

For the main effect of  $A$  form the  $(n + 1)$  sums of values whose indices satisfy

$$\begin{aligned} x_1 &= 0 \pmod{n + 1} \\ x_1 &= 1 \pmod{n + 1} \\ &\vdots \\ x_1 &= n \pmod{n + 1} \end{aligned}$$

Denote these sums by  $A_1, A_2, \dots, A_{n+1}$ . The sum of squares for the main effect of  $A$  is given by

$$\frac{\Sigma A^2}{(n+1)^2} - \frac{(\Sigma A)^2}{(n+1)^3} \text{ (degrees of freedom = } n\text{)}.$$

Similar computations give the sum of squares for the main effects of  $B$  and  $C$ .

For the two-factor interactions the sums of values whose indices satisfy the equations below are computed.

$$\begin{cases} x_1 + x_2 = 0 \text{ mod } (n+1) \\ \cdot \\ \cdot \\ x_1 + x_2 = n \text{ mod } (n+1) \\ \cdot \\ \cdot \\ x_1 + nx_2 = 0 \text{ mod } (n+1) \\ \cdot \\ \cdot \\ x_1 + nx_2 = n \text{ mod } (n+1) \end{cases}$$

From the  $(n+1)$  sums corresponding to  $x_1 + \alpha x_2 = 0, 1, \dots, n \text{ mod } (n+1)$  are computed the sum of squares associated with the  $n$  degrees of freedom for  $AB^\alpha$  and the  $AB$  interaction is given by the total of such sums over all values of  $\alpha$ . For the three-factor interaction one computes the  $(n+1)n^2$  sums of values for which the indices satisfy

$$x_1 + \alpha x_2 + \beta x_3 = 0, 1, \dots, n \text{ mod } (n+1),$$

where  $\alpha = 1, 2, \dots, n$ , and  $\beta = 1, 2, \dots, n$ . Each group of  $(n+1)$  sums give the sum of squares associated with the  $n$  degrees of freedom for the effect  $AB^\alpha C^\beta$ . For each group one computes:

$$\frac{\Sigma (\text{Sums})^2}{\text{Number of items in each sum}} - \frac{(\text{Grand total})^2}{\text{Total number of items}}$$

The extension to higher order interactions is straightforward.

This technique is ideally adapted to analysis of variance of factorials where block confounding occurs or to the analysis of fractional replication of factorials. *Example.* A  $3^4$  design in blocks of 9 with  $ABD, ACD^2, AB^2C^2$ , and  $BC^2D^2$  confounded with blocks is computed in the manner

described to get the usual analysis except for the combination of the sums of squares for the three-factor interactions which involve confounding with blocks. For *example* the *ABD* interaction is given by the sum of squares associated with  $AB^2D$ ,  $ABD^2$ , and  $AB^2D^2$  each of which has two degrees of freedom. The sum of squares associated with *ABD* is assigned to blocks.

For fractional factorials (with or without block confounding) the analysis is carried out as if it were a complete design with fewer factors by suppressing one or more of the indices. The individual components,  $A, \dots, B, \dots, AB, AB^2, \dots, ABC, ABC^2, \dots$ , are computed, and an identification is made according to the identity relationships (and block confounding, if any). (For further details see Ref. 23.)

**Analysis of Variance for  $2^n$  Factorials.** An *example* for a  $2^2$  experiment will illustrate this procedure. Enter observations in the order designated.

<i>Observed Values</i>	<i>First Sums and Differences, <math>D_1</math></i>	<i>Second Sums and Differences, <math>D_2</math></i>	<i><math>D_2^2/2^n</math> Will Give</i>
(1) = $x_{00}$	(1) + $a$	(1) + $a + b + ab$	Correc. for mean
$a = x_{01}$	$b + ab$	$a - (1) + ab - b$	$A$
.....	.....	.....	.....
$b = x_{10}$	$a - (1)$	$b + ab - (1) - a$	$B$
$ab = x_{11}$	$ab - b$	$ab - b - a + (1)$	$AB$

In general:

(a) Form a column of sums of the  $2^{n-1}$  pairs followed by  $2^{n-1}$  differences between the first and second element of a pair.

(b) Repeat this operation on the column so formed until the  $n$ th such column is formed.

(c) Then square the entries in the  $n$ th column and divide by  $2^n$  to get analysis of variance table in the order  $A, B, AB, C, AC, BC, ABC, \dots$ . The observations are entered so that their subscripts form an increasing sequence when regarded as binary numbers; e.g., for  $n = 3$  the observations are in the order  $x_{000} x_{001} x_{010} x_{011} x_{100} x_{101} x_{110} x_{111}$ .

**Analysis of Fractional Replication of  $2^n$  Factorials.** Arrange the  $(1/2^k)2^n = 2^s$  observations in the proper order for a  $2^s$  factorial (suppressing the other indices) and carry out the analysis as above. Identify the results of the analysis by using the identity relationships and the block confounding in the manner shown in the following example.

EXAMPLE.  $\frac{1}{4}$  replication of  $2^6$  in blocks of 8.  
 Fundamental identity:  $I = ABEF = ACDF = BCDE$ .  
 Block confounding:  $CD$ .



With *vector notation*, this system takes the form:

$$(29) \quad y'(x) = f(x, y(x)),$$

where  $y$  is a vector whose components are  $y_i(x)$ ,  $i = 1, 2, \dots, n$ , and  $f$  is a vector whose components are  $f_j(x, y_1(x), \dots, y_n(x))$ ,  $j = 1, 2, \dots, n$ .

Vector notation will be used throughout this section covering systems of equations which can be put in this form. The reader who is not familiar with vectors can take the case where  $y(x)$  is a single function of  $x$  and use this section as a guide to the solution of one first order equation.

At the end of this section is a summary table of some useful numerical methods for solving differential equations on a digital computer (see Table 11). Some important characteristics of each of these methods are listed. The prospective user may employ this table as a quick guide in selecting the most suitable method for the problem at hand.

**Requirements for Solution.** A solution of eq. (29) is a vector  $y(x)$  which satisfies eq. (29). It necessarily possesses a first derivative.

The differential equations used by engineers nearly always possess solutions which have continuous derivatives of many or all orders or indeed are analytic (i.e., the Taylor series converges) except at isolated points. They are said to be piecewise continuous and have piecewise continuous derivatives. The isolated discontinuities are of practical importance since engineer's derivatives are such quantities as current, voltage, velocity, and acceleration which he must limit to avoid damage to his equipment. Methods of solving differential equations that are awkward at discontinuities are of restricted value to him.

**Numerical Solution.** The Taylor series for  $y(x)$  in the neighborhood of some point  $x_0$ :

$$y(x) = y(x_0) + y'(x_0)(x - x_0) + \dots + y^{(m)}(x_0)(x - x_0)^m/m! + \dots,$$

enables one to approximate  $y$  by an  $m$ th degree polynomial in  $x - x_0$ . *Most numerical methods of solving differential equations depend directly or indirectly on this fact.*

Consider a set of points

$$x_{i+j} = x_j + ih, \quad i = 0, \pm 1, \pm 2, \dots$$

These points are equally spaced along the  $x$ -axis and the distance between neighboring points is  $h$ , called the *grid size*.

Write the Taylor series of  $y$ ,  $hy'$ ,  $h^2y''$ , etc., at each of these points:

$$(30a) \quad y_{i+j} = y(x_{i+j}) = y_j + ihy'_j + \dots + i^m h^m y_j^{(m)}/m! + R_{m+1},$$

$$(30b) \quad hy'_{i+j} = hy'_j + ih^2y''_j + \dots + i^{m-1} h^m y_j^{(m)}/(m-1)! + R_{m+1},$$

$$(30c) \quad h^2y''_{i+j} = h^2y''_j + \dots + i^{m-2} h^m y_j^{(m)}/(m-2)! + R_{m+1},$$

where  $R_{m+1}$  is a generic notation for a remainder which contains  $h^{m+1}$  as a factor. Equations (30) can be used in an endless variety of ways to obtain procedures for the numerical solutions of eq. (29).

**Solutions for Known  $y_i$  and  $y'_i$ .**  $y_i$  and  $y'_i$  are known at several past points (i.e.,  $i = 0, -1, -2, \dots, -I$ ) and  $y_{j+1}$  is desired. Solve eqs. (30a) and (30b) at  $i = -1, \dots, -I$  for  $2I$  of the quantities:

$$h^2 y_j / 2!, h^3 y_j^{(3)} / 3!, \dots, h^{2I+1} y_j^{(2I+1)} / (2I + 1)!,$$

and substitute into eq. (30a) for  $y_{j+1}$  and obtain a formula accurate to terms of degree  $2I + 2$  in  $h$ . Thus we have Table 9.

TABLE 9. EXTRAPOLATION FORMULAS

Formula	$I$	$y_{j+1}$	$y_j$	$hy'_j$	$y_{j-1}$	$hy'_{j-1}$	$y_{j-2}$	$hy'_{j-2}$	$y_{j-3}$	$hy'_{j-3}$	Error
1	0	1	1	1							$y^{(2)}h^2/2$
2	1	1	-4	4	5	2					$y^{(4)}h^4/8$
3	2	1	-18	9	9	18	10	3			$h^6 y^{(6)}/20$
4	3	1	$-\frac{128}{3}$	16	-36	72	64	48	$\frac{47}{3}$	4	$h^8 y^{(8)}/70$

*First Order Method.* Formula 1 of Table 9 is the simplest and best known of all solution methods and is due to *Euler*. The value of  $y_{j+1}$  is

$$(31) \quad y_{j+1} = y_j + hy'_j,$$

then the value of  $y'_{j+1}$  is obtained from eq. (29). It can be shown that the approximate solution obtained in this fashion converges to the exact solution as the grid size approaches zero, the error at a given point being proportional to  $h$ . This is called a first order method. The principal attraction of this method is its simplicity. Its principal disadvantage whether for hand or electronic digital computation is that it requires a small grid size to obtain a given accuracy.

**Studying the Stability of the Method.** The most illuminating test of any method of solving differential equations (ordinary or partial) is to perturb the solution and study the local properties of the perturbed solution. To illustrate, consider Euler's method for solving a single eq. (29). Suppose that a small error  $\epsilon$  is made at  $x_0$  and that  $z_j$  is the Euler solution of eq. (29) with this error at  $x_0$ .

Let

$$\eta_j = z_j - y_j.$$

Then, since  $y_j$  and  $z_j$  each satisfy eq. (31), one finds, with the mean value theorem, that

$$\eta_{j+1} = \eta_j \left( 1 + h \frac{\partial f}{\partial y} (x_j, y_j + \theta \eta_j) \right), \quad 0 < \theta < 1.$$

Consider now a small enough neighborhood of  $x_0$  so that second order effects may be neglected, i.e., that  $\partial f/\partial y$  may be taken to be a constant  $d$ . Then

$$\eta_{j+1} = \eta_j(1 + hd) = \eta_0(1 + hd)^{j+1} = \epsilon[(1 + hd)^{1/hd}]hd(j + 1).$$

If attention is focused on a fixed point,

$$x = x_0 + (j + 1)h,$$

and if  $h$  is allowed to approach zero,  $\eta_{j+1}$  approaches  $\epsilon \exp(x - x_0)d$ . The error at  $x$ , due to the error  $\epsilon$  at  $x_0$ , thus remains finite as the grid size goes to zero, and the method is said to be locally stable. The error grows with  $x$  if  $\partial f/\partial y$  is positive; otherwise it decreases.

The same method shows that the other extrapolation formulas 2, 3, and 4 of Table 9 cannot by themselves be used to solve differential equations because they are locally unstable, i.e., the error at  $x$  due to a given error at  $x_0$  becomes infinite as the grid size goes to zero.

**Solutions for Known  $y_{j+1}$ .** If  $y_{j+1}$  is obtained in some fashion,  $y'_{j+1}$  can be found from eq. (29). Using eq. (30) for  $hy'_{j+1}$  in addition to the equations used to obtain Table 9 results in Table 10.

TABLE 10. EXTRAPOLATION FORMULAS

Formula	$I$	$y_{j+1}$	$hy'_{j+1}$	$y_j$	$hy'_j$	$y_{j-1}$	$hy'_{j-1}$	$y_{j-2}$	$hy'_{j-2}$	$y_{j-3}$	$hy'_{j-3}$	Error
5	0	1	$\frac{1}{2}$	1	$\frac{1}{3}$				(Trapezoidal formula)			$-h^3y^{(3)}/4$
6	1	1	$\frac{1}{3}$	0	$\frac{4}{3}$	1	$\frac{1}{3}$		(Simpson's rule)			$-h^5y^{(5)}/90$
7	2	1	$\frac{1}{12}$	$-\frac{2}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	1		$\frac{1}{12}$		$-\frac{8}{15}h^7y^{(7)}$

*Heun's second order method* has its basis in formula 5, Table 10, the trapezoidal formula. It is a considerable improvement on Euler's method since a much larger grid size may be used. It is just as stable as Euler's method, requires no past history, and calls for substitution in eq. (29) only once per point.

One uses Euler's formula for a first value of  $y_{j+1}$ , eq. (29) to find  $y'_{j+1}$  and then Heun's formula for a better value of  $y_{j+1}$ . It is not necessary to recompute  $y'_{j+1}$ . The process may be iterated if desired.

*The procedure which Milne* (Ref. 24) *recommends* most highly for solving ordinary differential equations uses

$$y_{j+1} = y_{j-3} + 4hy'_{j-1} + 8h/3(y'_{j-1} - 2y'_{j-2} + y'_{j-3}) + \frac{28}{9}h^5y^{(5)}$$

to extrapolate and formula 6, Table 10, which is *Simpson's rule*,

$$y_{j+1} = y_{j-1} + h/3(y'_{j-1} + 4y'_j + y'_{j+1}) - h^5y^{(5)}/90$$

to recalculate.

*Solution by formulas 2 and 6.* A procedure which requires less past history and therefore is better for starting and at discontinuities uses formula 2, Table 9, to find a third order approximation to  $y_{j+1}$  and Simpson's rule to recalculate. Either procedure calculates derivatives only once per point.

Formula 7, Table 10, is unstable and therefore useful for extrapolation but not for recalculation.

**Method of Adams and Bashforth** (Ref. 25). This approach is best expressed in terms of differences:

$$\begin{aligned} \nabla y'_j &= y'_j - y'_{j-1}, \\ \nabla^2 y'_j &= \nabla(y'_j - y'_{j-1}) = y'_j - 2y'_{j-1} + y'_{j-2}, \\ &\vdots \\ \nabla^n y'_j &= \nabla(\nabla^{n-1} y'_j - \nabla^{n-1} y'_{j-1}), \\ y_{j+1} &= y_j + h(y'_j + \nabla y'_j/2 + 5\nabla^2 y'_j/12 + 3\nabla^3 y'_j/8 + 251\nabla^4 y'_j/720 + \dots). \end{aligned}$$

For solutions whose derivatives of some order are everywhere continuous, this method has the advantage of yielding arbitrarily high order of approximation with only one evaluation of derivatives per point. For automatic computer use, it has several disadvantages which lead to its rare use. A special starting process is required; it is awkward to change grid size; at each isolated discontinuity, the special starting process must be used again.

**The Runge-Kutta Method.** Like Euler's and Heun's methods, this method avoids these difficulties (Refs. 26 and 27). It has several forms. One of the best known, which has a truncation error proportional to  $h^5$ , is:

$$\begin{aligned} y_{j+1} &= y_j + (k_1 + 2k_2 + 2k_3 + k_4)/6 \\ k_1 &= hf(x_j, y_j), \\ k_2 &= hf(x_j + h/2, y_j + k_1/2), \\ k_3 &= hf(x_j + h/2, y_j + k_2/2), \\ k_4 &= hf(x_j + h, y_j + k_3). \end{aligned}$$

The Runge-Kutta method was recently adapted to automatic computers by Gill in a form which concentrates on saving memory and reducing round-off error (Ref. 28).

All forms of the Runge-Kutta method have the disadvantage that the derivatives must be evaluated several times, four in these two cases.

**Fourth Order Method.** This method has been used extensively on automatic computers since 1946 and has been carefully studied by Dims-

dale and Clippinger (Ref. 29). It consists in extrapolating for  $y_{j+2}$  by the third order formula using one past point (see Table 11):

$$(32a) \quad y_{j+2} = y_{j-2} + 4(y_{j-2} - y_j) + 4h(2y'_j + y'_{j-2}) + 2h^4y_j^{(4)}.$$

The derivative  $y'_{j+2}$  is then found and also  $y_{j+1}$  by

$$(32b) \quad y_{j+1} = (y_j + y_{j+2})/2 + (h/4)(y'_j - y'_{j+2}) - h^4y_j^{(4)}/24.$$

The derivative  $y'_{j+1}$  is then found, and  $y_{j+2}$  is redetermined by Simpson's rule:

$$(32c) \quad y_{j+2} = y_j + h/3(y'_j + 4y'_{j+1} + y'_{j+2}) + h^5y_j^{(5)}/90.$$

Isolated discontinuities are made to fall at odd-numbered grid points by adjusting  $h$ . To start, or at points where the grid size is altered, eqs. (32c) and (32b) are iterated, and eq. (32a) is not used. Thus, like Runge-Kutta's or Gill's methods, this method requires no past history, and is well suited to starting, discontinuities, and change of grid size. By the addition of a single point from past history, it achieves the efficiency of Adam's, Milne's, and other methods requiring only one evaluation of derivatives per point.

**Higher Derivatives.** Sometimes eq. (29) can be easily differentiated. In this case a fourth order, stable procedure requiring no past history is obtained by eliminating  $h^3y^{(3)}$ , and  $h^4y^{(4)}$  from eqs. (30a), (30b), and (30c) at  $i = 0, 1$ :

$$y_{j+1} = y_j + h(y'_j + y'_{j+1})/2 + h^2(y''_j - y''_{j+1})/12 + h^5y_j^{(5)}/720.$$

By adding a single point from past history, one obtains the *predictor*,

$$y_{j+1} = 32y_j - 31y_{j-1} - 2h(8y'_j + 7y'_{j-1}) \\ + h^2(9y''_j - 4y''_{j-1})/2 + h^6y_j^{(6)}/720,$$

and the *seventh order corrector*,

$$y_{j+1} = y_{j-1} + 2y_j + 3h(y'_{j+1} - y'_{j-1})/8 \\ + h^2(8y''_j - y''_{j-1} - y''_{j+1})/24 + h^8y_j^{(8)}/60450,$$

which can be used except at the start, at discontinuities, and at grid change points.

*Method of Brock and Murray.* A method which takes advantage of the fact that differential equations are locally linear with constant coefficients and therefore have solutions which are locally linear combinations of exponentials has been developed by Brock and Murray (Ref. 48).

**Extrapolation to Zero Grid Size.** If a method has a local error proportional to  $h^{n+1}$ , it has an error at a given  $x$  proportional to  $h^n$ , since the number of local errors made going from  $x_0$  to  $x$  is  $(x - x_0)/h$ . By call-

ing  $\epsilon$  the error at  $x$ , and the exact answer,  $\bar{y}$ , then,

$$(33) \quad \epsilon = y - \bar{y} = ah^n + bh^{n+1} + \eta,$$

where the remainder,  $\eta$ , goes to zero as  $h^{n+2}$ . If eq. (29) is solved numerically at two grid sizes,  $h_1$  and  $h_2$ , one may write eq. (33) at both grids and solve for  $\bar{y}$ :

$$(34) \quad \bar{y} = y_1 + (y_1 - y_2)r^n/(1 - r^n) + bh_2^{n+1}r^n(1 - r)/(1 - r^n) - (\eta_1 - r^n\eta_2)/(1 - r^n),$$

where  $r = h_1/h_2$ . Richardson (Ref. 30), who invented this procedure, called it "extrapolation to zero grid size." Looking at the next to last term, one sees that it would be more apt to call it "increasing the order of accuracy from  $n$  to  $n + 1$ ." Equation (34) is useful in many ways. For *example*: (a) One can solve (29) at two grid sizes and use eq. (34) to get a better answer at common points. (b) One can solve (29) at one grid size and occasionally take a step at two grid sizes by using the second term to estimate the error and adjust the grid size. (With this procedure it is important to use methods which depend on little past history.) (c) One can take every step at two grids, use eq. (34) to improve the accuracy before proceeding, and also use the second term to adjust the grid size.

**Boundary Value Problems or Distributed Conditions.** It may happen that not all components of  $y$  are specified at one value of  $x$ . Instead, some of the components of  $y$  may be given in terms of the others at two or more points.

*Approach A.* Perhaps the most obvious approach to this problem is to:

1. Assume initial conditions at  $x_0$ .
2. Solve the problem.
3. Assume other initial conditions.
4. Resolve the problem.

5. Interpolate between the initial conditions for initial conditions which will satisfy one of the other given conditions at some other point. (This is based on the theorem that the solutions of differential equations are, under suitable conditions, continuous functions of their values at particular points.)

6. Reiterate this process until all conditions are satisfied. If there are many conditions to be satisfied by varying the same number of components of  $y$  at  $x_0$  as parameters, the interpolation process becomes quite complicated. If convergence is also slow, it may be necessary to solve the differential equation thousands of times, treating the different equations and distributed conditions as simultaneous equations for all the variables at all the points.

*Approach B* is to consider all the distributed conditions and the approximating equations simultaneously. For instance, consider the second order system:

$$(35) \quad y' = f(x, y, z), \quad z' = g(x, y, z),$$

with the distributed conditions:

$$(36) \quad y(a) = A, \quad ky(b) + lz(b) + my'(b) + z'(b) = 0.$$

One might use Heun's approximating difference equations:

$$(37) \quad \begin{aligned} y_{j+1} - y_j &= (f(x_j, y_j, z_j) + f(x_{j+1}, y_{j+1}, z_{j+1}))(h/2), \\ z_{j+1} - z_j &= (g(x_j, y_j, z_j) + g(x_{j+1}, y_{j+1}, z_{j+1}))(h/2). \end{aligned}$$

Replacing  $a, b$  by  $x_0, x_n$  one would write the side conditions eq. (36) in the form

$$(38) \quad \begin{aligned} y(x_0) &= A, \\ m(y_n - y_{n-1}) + (z_n - z_{n-1}) &= (h/2)[mf_{n-1} + g_{n-1} - ky_n - lz_n], \end{aligned}$$

where the second eq. (36) is replaced by one equivalent to it to third order and  $f_n$  is written for  $f(x_n, y_n, z_n)$ .

Equations (37) and (38) are  $2(n+1)$  simultaneous equations for the  $2(n+1)$  unknowns  $y_j, z_j, j = 0, 1, 2, \dots, n$ . They are not linear; however,  $h$  appears as a factor of the right members and the left members are linear. It is therefore natural and quite practical to define an iterative process, writing  $y_j^i$  and  $z_j^i$  for the  $i$ th approximation to  $y_j$  and  $z_j$ :

$$(39) \quad \begin{aligned} y_0^i &= A, \\ y_{j+1}^i - y_j^i &= (h/2)(f_j^{i-1} + f_{j+1}^{i-1}), \\ z_{j+1}^i - z_j^i &= (h/2)(g_j^{i-1} + g_{j+1}^{i-1}), \\ m(y_n^i - y_{n-1}^i) + z_n^i - z_{n-1}^i & \\ &= (h/2)(mf_{n-1}^{i-1} + g_{n-1}^{i-1} - ky_n^{i-1} - z_n^{i-1}). \end{aligned}$$

*Approach C*, useful if  $f$  and  $g$  are readily differentiable, is to perturb eqs. (35) by introducing  $\eta = y - \bar{y}, \zeta = z - \bar{z}$  where  $\bar{y}, \bar{z}$  is some approximate solution:

$$(40) \quad \eta' = \frac{\partial f}{\partial y} \eta + \frac{\partial f}{\partial z} \zeta, \quad \zeta' = \frac{\partial g}{\partial y} \eta + \frac{\partial g}{\partial z} \zeta.$$

It would be possible to use eqs. (39) to find  $\bar{y}$  and  $\bar{z}$  and then, by evaluating the derivatives  $\partial f/\partial y$ , etc., at  $\bar{y}, \bar{z}$  solve eq. (40) as linear equations for  $\eta, \zeta$  subject to initial conditions  $\eta(x_0) = \zeta(x_0) = 0$ .

**Computer Storage Requirements and Number of Operations.**

The columns of Table 11 provide a guide to the use of the methods listed. Similar remarks apply to this table as in the concluding paragraph of Sect. 2 relative to content and notation of Table 5 in that section.

The presence of past history requirements is an important consideration for digital computers because this generally means programming special starting programs for use at boundary points, at points of discontinuity of the solution, or at points where the grid size changes. For this reason formulas requiring no past history are in general easiest to program.

It is important in evaluating a digital computer procedure to be able to estimate the number of operations, multiplication times, or other index of the computing time. However, in practical problems in differential equations, this time is almost completely dominated by the time to compute the derivative  $f(x, y)$  in eq. (29), and this, of course, cannot be determined except in the context of a specific problem. The next best guide to the volume of computations is the number of times the derivative must be computed per integration step, and this is listed in the last column of Table 11.

TABLE 11. COMPUTER REQUIREMENTS IN SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS

Method	Order of Error	Past History Required	Computer Storage <sup>a</sup>	Derivative Evaluated, Times/Step
Extrapolation Formulas (Tables 9, 10)				
Formula 1 (Euler)	$h^2$	None	$2n + w$	1
Formula 5 (Heun)	$h^3$	None	$3n + w$	1
Adams-Bashforth	Arbitrary	$k$ points, where $k$ is arbitrary	$n(k + 1) + w$	1
Runge Kutta	$h^5$	None	$4n + w$	4
Gill	$h^5$	None	$3n + w$	4
Fourth Order Method	$h^5$	None	$6n + w$	2
(1st deriv. once and 2nd deriv. once)				
Predictor-corrector formulas				
Milne	$h^5$	3 points	$5n + w$	1
Dimsdale-Clippinger (using 3 iterations)	$h^5$	None	$6n + w$	3
Dimsdale-Clippinger (using extrapolator)	$h^5$	1 point	$6n + w$	1
5th order predictor-7th order corrector	$h^8$	1 point	$6n + w$	2
Extrapolation to zero grid size	Order of $h^{n+1}$ , at least, if error of for- mula used is of order $h^n$			3n

<sup>a</sup>  $n$  is dimension of vector  $y$ , and  $w$  is undetermined amount of working storage and program storage.

## 6. PARTIAL DIFFERENTIAL EQUATIONS

J. B. Diaz

R. F. Clippinger

Bernard Friedman

Eugene Isaacson

Robert Richtmyer

**Introduction.** A variety of physical problems, when analyzed from a mathematical point of view, lead to the consideration of boundary value problems for differential equations. In many cases, the physical quantity of interest is found to be represented by a function which satisfies a differential equation in a certain domain of the independent variables. Besides the differential equation (which may be ordinary or partial, depending upon whether the independent variables are one or more than one, respectively) the "unknown" function is required to satisfy certain other conditions, which will be referred to collectively as *boundary conditions*. Generally speaking, these additional boundary conditions select, from the totality of the solutions of the differential equation in question, the solutions which correspond to the actual physical situation under study. *Example.* The determination of the steady-state temperature in a plane circular plate of unit radius, whose periphery is maintained at a given temperature, amounts to the determination of a real-valued function  $u(x, y)$  satisfying the partial differential equation

$$\partial^2 u / \partial x^2 + \partial^2 u / \partial y^2 = 0 \quad \text{for } 0 \leq x^2 + y^2 < 1,$$

and the boundary condition

$$u(x, y) = f(x, y) \quad \text{for } x^2 + y^2 = 1,$$

where  $f$  is a prescribed function. ( $f$  is essentially the preassigned temperature distribution on the periphery.)

An equation involving a function of two or more variables and its partial derivatives is called a *partial differential equation*. The *order* of a partial differential equation is the order of the highest order derivative which actually appears in it. A partial differential equation is *linear*, if it is of the first degree when considered as a polynomial in the unknown function and its partial derivatives (otherwise the equation is called *nonlinear*). *Example.* The equation  $\partial^2 u / \partial x^2 + \partial^2 u / \partial y^2 = 0$  is a linear second order equation, while the equation  $(\partial u / \partial x)^2 + u = 0$  is a nonlinear first order equation.

This section will consider linear and second order partial differential equations starting with some mathematical background and leading to a discussion of numerical methods suitable for digital computer use. The section will conclude with a summary table giving some significant attributes of the methods listed from the point of view of digital computer solution (see Table 12).

**First Order Partial Differential Equations**

Consider

$$(41) \quad \begin{aligned} F(x, y, z, p, q) &= 0, \\ p &= \frac{\partial z}{\partial x}, \quad q = \frac{\partial z}{\partial y}, \end{aligned}$$

a partial differential equation of first order for  $z$  as a function of  $x, y$ . The general solution of this problem depends on an arbitrary function. Lagrange showed that the general solution could be deduced from a "complete" solution, i.e., a two-parameter family of particular solutions. Lagrange and Charpit also showed that such a complete solution could be deduced from the solution of the system of ordinary equations for  $x, y, z, p, q$  in terms of a parameter:

$$(42) \quad \begin{aligned} \frac{dx}{dt} = x' &= P(t) = \frac{\partial F(x, y, z, p, q)}{\partial p}, \\ y' &= Q = \frac{\partial F}{\partial q}, \\ z' &= Pp + Qq, \\ p' &= -\left(\frac{\partial F}{\partial x} + p \frac{\partial F}{\partial z}\right), \\ q' &= -\left(\frac{\partial F}{\partial y} + q \frac{\partial F}{\partial z}\right). \end{aligned}$$

Cauchy showed that any particular solution of eqs. (41) is composed of curves he called characteristics obtained by integrating eqs. (42).

Let

$$(43) \quad x_0 = f(s), \quad y_0 = g(s), \quad z_0 = h(s)$$

be the parametric equations of a curve through which a particular solution

of eqs. (41) is to be found. Then  $p_0(s)$  and  $q_0(s)$  must satisfy the differential equation,

$$(44) \quad F(f, g, h, p_0(s), q_0(s)) = 0,$$

and the condition

$$(45) \quad fp_0(s) + gq_0(s) - h = 0.$$

The solution of eqs. (42) subject to initial conditions (43), (44), and (45) can be represented by

$$x = x(u, t),$$

$$y = y(u, t),$$

$$z = z(u, t),$$

$$p = p(u, t),$$

$$q = q(u, t).$$

Thus the problem of finding the solution of eqs. (41) passing through curve (43) is reduced to the solution of ordinary equations which can be done by the methods of Sect. 5.

To illustrate, consider the linear equation,

$$x + y + z + p + q = 0$$

subject to the conditions

$$y = z = 0,$$

when

$$(46) \quad 0 \leq x \leq 1.$$

When  $y$  is zero and  $x$  is outside the range (46),  $z$  is not defined.

*Cauchy's method* yields the solution

$$x = s + t,$$

$$y = t,$$

$$z = -2t + (s - 2)(e^{-t} - 1),$$

$$p = e^{-t} - 1,$$

$$q = -1 + (1 - s)e^{-t}.$$

$$0 \leq s \leq 1.$$

Eliminating  $s$  and  $t$  yields

$$z = -2y + (x - y - 2)(-1 + e^{-y}).$$

Cauchy's method shows that, in general, if  $z$  is given along some arc of curve  $C$  terminated at points  $A$  and  $B$ , then  $z$  is determined in a strip bounded by a characteristic through  $A$  and a characteristic through  $B$ . Call this strip the region of determinacy, in our example, the strip between  $y = x$  and  $y = x - 1$ . Any method other than Cauchy's must therefore determine these characteristics one way or another to determine the region of determinacy.

Practically, it may be difficult to obtain  $\partial F/\partial x$ ,  $\partial F/\partial y$ , and  $\partial F/\partial z$ . In that case, it is not possible to obtain  $p$  and  $q$  by using the last two equations (42). As an alternative, let the characteristic curves  $s = \text{constant}$  and the curves  $t = \text{constant}$  be used as a curvilinear coordinate network. The transformation from Cartesian coordinates  $x, y$  to coordinates  $s, t$  is governed by the relations

$$\begin{aligned}
 t_x &= y_s/\Delta, \\
 t_y &= -x_s/\Delta, \\
 s_x &= -y_t/\Delta, \\
 s_y &= x_t/\Delta, \\
 \Delta &= x_t y_s - y_t x_s,
 \end{aligned}
 \tag{47}$$

where  $t_x$  is  $\partial t/\partial x$ , etc.

By definition and eqs. (47),

$$\begin{aligned}
 p &= \partial z/\partial x = (z_t y_s - z_s y_t)/\Delta, \\
 q &= \partial z/\partial y = (-z_t x_s + z_s x_t)/\Delta.
 \end{aligned}
 \tag{48}$$

Start along  $t = 0$  and use Euler's method and the first three of eqs. (42) to get  $x, y, z$  at each point on  $t = h$ . Use numerical differentiation to obtain  $x_s, y_s$ , and  $z_s$  at each point on  $t = h$ . Use eqs. (48) to get  $p, q$  on the same curve. If more accuracy is desired, Heun's method may now be used to obtain better values on  $t = h$ . The same process may now be repeated for  $t = -h, 2h, -2h$ , etc.

### Second Order Partial Differential Equations

The general linear partial differential equation of the second order in the two independent real variables  $x$  and  $y$  is

$$a \frac{\partial^2 u}{\partial x^2} + 2b \frac{\partial^2 u}{\partial x \partial y} + c \frac{\partial^2 u}{\partial y^2} + d \frac{\partial u}{\partial x} + e \frac{\partial u}{\partial y} + fu = g,
 \tag{49}$$

where the letters  $a, b, \dots, g$  denote real-valued functions of  $x$  and  $y$ . The equation is called homogeneous if the "nonhomogeneous term"  $g$  is identically zero. The linear homogeneous equation has the property that the

“superposition principle of solutions” holds, i.e., that if  $u$  and  $v$  are solutions, any linear combination  $Au + Bv$ , with constant coefficients  $A$  and  $B$ , is also a solution.

**Classification.** The partial differential eq. (49) can be reduced to certain typical, or canonical forms by means of a suitable change of variables:

$$(50) \quad \xi = \xi(x, y), \quad \eta = \eta(x, y).$$

Consider first the equation with *constant* coefficients

$$(51) \quad A \frac{\partial^2 u}{\partial x^2} + 2B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} = 0,$$

and make the change of variables

$$(52) \quad \xi = \alpha x + \beta y, \quad \eta = \gamma x + \delta y,$$

with  $\alpha, \beta, \gamma, \delta$  real constants. In the new variables  $\xi, \eta$ , one has

$$(53) \quad (A\alpha^2 + 2B\alpha\beta + C\beta^2) \frac{\partial^2 u}{\partial \xi^2} + (A\gamma^2 + 2B\gamma\delta + C\delta^2) \frac{\partial^2 u}{\partial \eta^2} \\ + 2(A\alpha\gamma + B[\alpha\delta + \beta\gamma] + C\beta\delta) \frac{\partial^2 u}{\partial \xi \partial \eta} = 0.$$

Since eq. (51) is assumed to be of second order, not all three *real* constants  $A, B, C$  are zero, i.e.,  $A^2 + B^2 + C^2 > 0$ . It will now be supposed further that  $A \neq 0$ . There is no loss of generality, since if  $A = 0$  and  $C = 0$ , too, then  $B \neq 0$ , and the equation is already in “canonical” form (see eq. 54 below); whereas if  $A = 0$  and  $C \neq 0$  one has merely to interchange the roles of  $x$  and  $y$ . The classification into three types is as follows:

$$B^2 - AC > 0, \quad \textit{hyperbolic type},$$

$$B^2 - AC < 0, \quad \textit{elliptic type},$$

$$B^2 - AC = 0, \quad \textit{parabolic type}.$$

(The reason for the designations elliptic, hyperbolic, and parabolic is obvious from analytic geometry, the reduction of a quadratic bilinear form  $Ax^2 + 2Bxy + Cy^2$  to a sum of squares.)

**HYPERBOLIC CASE.** When  $B^2 - AC > 0$ , by choosing  $\beta = \delta = 1$ ,

$$\alpha = \frac{-B + \sqrt{B^2 - AC}}{A}, \quad \gamma = \frac{-B - \sqrt{B^2 - AC}}{A}$$

in eqs. (52) and (53), and by dividing (53) by a nonzero constant, one obtains the canonical form

$$(54) \quad \frac{\partial^2 u}{\partial \xi \partial \eta} = 0;$$

whereas by choosing

$$\alpha = \frac{-C}{\sqrt{B^2 - AC}}, \quad \beta = \frac{B}{\sqrt{B^2 - AC}}, \quad \gamma = 0, \quad \delta = 1,$$

one obtains similarly the canonical form

$$\frac{\partial^2 u}{\partial \xi^2} - \frac{\partial^2 u}{\partial \eta^2} = 0.$$

ELLIPTIC CASE. When  $B^2 - AC < 0$ , by choosing

$$\alpha = \frac{-C}{\sqrt{AC - B^2}}, \quad \beta = \frac{B}{\sqrt{AC - B^2}}, \quad \gamma = 0, \quad \delta = 1,$$

one obtains the canonical form

$$\frac{\partial^2 u}{\partial \xi^2} + \frac{\partial^2 u}{\partial \eta^2} = 0.$$

PARABOLIC CASE. When  $B^2 - AC = 0$ , by choosing  $\beta = 1$ ,  $\alpha \neq -B/A$  and  $\delta = 1$ ,  $\gamma = -B/A$ , one obtains the canonical form

$$\frac{\partial^2 u}{\partial \xi^2} = 0.$$

In the general case of an eq. (49) with variable coefficients, it is said to be of *elliptic*, *hyperbolic*, or *parabolic type at a given point*  $(x_0, y_0)$  according to whether  $b^2(x_0, y_0) - a(x_0, y_0)c(x_0, y_0)$  is  $< 0$ ,  $> 0$ , or  $= 0$ , respectively. If the coefficients  $a, \dots, g$  are sufficiently smooth in a neighborhood of  $(x_0, y_0)$ , and eq. (49) is elliptic at each point of the neighborhood, there is a sufficiently small subneighborhood of  $(x_0, y_0)$  in which one can introduce new variables by means of eq. (50) (not necessarily a *linear* change of variables as in the case eq. (51) of *constant* coefficients) so that eq. (49) becomes, in this subneighborhood,

$$\frac{\partial^2 u}{\partial \xi^2} + \frac{\partial^2 u}{\partial \eta^2} + \left( \text{Linear terms in } \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}, \text{ and } u \right) = 0.$$

A similar statement applies in the hyperbolic and parabolic cases.

Of course, eq. (49) with variable coefficients may be of different type at different points of a domain, i.e., it may be of "mixed" type, as occurs in the linearized equation for the potential function of a two-dimensional compressible flow. *Example.* The equation  $y\partial^2 u/\partial y^2 + \partial^2 u/\partial x^2 = 0$  is elliptic for  $y > 0$ , parabolic for  $y = 0$ , and hyperbolic for  $y < 0$ .

Representative equations commonly studied are:

(a) Elliptic	Laplace	$\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = 0$
(b) Hyperbolic	Vibrating string	$\partial^2 u/\partial x^2 - \partial^2 u/\partial y^2 = 0$
(c) Parabolic	Heat	$\partial^2 u/\partial x^2 - \partial u/\partial y = 0$

The variable  $t$  is usually written instead of the variable  $y$  in the last two equations. For a more detailed discussion of the canonical forms of eq. (49), as well as for the classification into canonical forms of higher order equations and systems of equations see Refs. 36-38.

**Difference Equations.** In numerical investigations it is often necessary to replace the partial differential equation occurring in a given boundary value problem by a suitable equation involving differences rather than derivatives of the unknown function (see Chap. 4). The basic principle usually employed is none other than the fact that any partial derivative is the limit of a certain difference quotient. For a function of one variable,  $f(x)$ , the difference quotients in the plus  $x$  and minus  $x$  directions,  $f_x$  and  $f_{\bar{x}}$ , are defined by

$$f_x(x) = \frac{f(x+h) - f(x)}{h} \quad \text{and} \quad f_{\bar{x}}(x) = \frac{f(x) - f(x-h)}{h},$$

where  $h > 0$ . The second differences of  $f(x)$  are defined as the differences of the first differences. There are three second differences,  $f_{xx}$ ,  $f_{x\bar{x}} (= f_{\bar{x}x})$ , and  $f_{\bar{x}\bar{x}}$ . The second difference  $f_{x\bar{x}}$  is the most "symmetric" of the three:

$$f_{x\bar{x}}(x) = \frac{f(x+h) + f(x-h) - 2f(x)}{h^2}.$$

The corresponding differences for functions of several independent variables are defined as above, upon holding fixed all the variables but one at a time. For *example*, for a function of two variables  $u(x, y)$ :

$$u_x(x, y) = \frac{u(x+h, y) - u(x, y)}{h}$$

and

$$u_{x\bar{x}}(x, y) = \frac{u(x+h, y) - 2u(x, y) + u(x-h, y)}{h^2}, \text{ etc.}$$

*Laplace equation.* By taking the difference equation  $u_{x\bar{x}} + u_{y\bar{y}} = 0$  as a "difference approximation" to Laplace's differential equation  $\partial^2 u / \partial x^2 + \partial^2 u / \partial y^2 = 0$ , one obtains the difference equation

$$(55) \quad \frac{u(x+h, y) + u(x-h, y) + u(x, y+h) + u(x, y-h)}{4} = u(x, y).$$

*Vibrating string.* By taking the difference equation  $u_{x\bar{x}} - u_{y\bar{y}} = 0$  as a difference approximation to the vibrating string equation  $\partial^2 u / \partial x^2 - \partial^2 u / \partial y^2 = 0$ , one obtains the difference equation

$$u(x+h, y) + u(x-h, y) - u(x, y+h) - u(x, y-h) = 0.$$

*Heat.* In the case of the heat equation  $\partial^2 u / \partial x^2 - \partial u / \partial y = 0$ , one has the alternative difference equations  $u_{x\bar{x}} - u_y = 0$  and  $u_{x\bar{x}} - u_{\bar{y}} = 0$ .

Exactly the same procedure is applicable to first and to higher order partial differential equations, as well as to systems of equations. An alternative approach to the numerical treatment of first order partial differential equations can be based on the fact demonstrated in the previous subsection that the solution of a first order partial differential equation and the solution of the characteristic system of ordinary differential eqs. (42) corresponding to the given first order partial differential equation are equivalent tasks.

*Note.* The following three subsections represent results obtained at the Institute of Mathematical Sciences, New York University, under the sponsorship of the United States Atomic Energy Commission Contract AT(30-1)1480. Reproduction in whole or in part permitted for any purpose of the United States Government.

### Elliptic Partial Differential Equations

Consider a partial differential equation of second order for a function  $u$  of  $n$  variables  $x_1, x_2, \dots, x_n$ . One writes the equation as follows:

$$(56) \quad Lu \equiv \sum_{i,j} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_i b_i \frac{\partial u}{\partial x_i} + cu = f.$$

The coefficients  $a_{ij}, b_i, c$  are assumed to be constant. This equation is called *elliptic* in a region  $R$  if the quadratic form

$$(57) \quad \sum_{i,j} a_{ij} \xi_i \xi_j$$

is non-negative definite for all values of the  $\xi_i$  such that  $(\xi_1, \xi_2, \dots, \xi_n)$  is

a point in  $R$ . A typical example of an elliptic difference equation is *Poisson's equation*, that is,

$$\Delta u \equiv \sum_i \frac{\partial^2 u}{\partial x_i \partial x_i} = f.$$

Note that the condition (57) for ellipticity depends only on the highest order derivative terms of eq. (56).

**Dirichlet and Neumann Problems.** A typical problem involving elliptic differential equations is one that requires the solution of a boundary value problem. For *example*, a typical problem would be to solve

$$Lu = f$$

in a region  $R$  given that  $u$  is a prescribed function  $u_0$  on the boundary  $B$  of the region  $R$ . Such a problem is called a *Dirichlet problem* for eq. (56). If, instead of the values of  $u$ , the values of  $\partial u / \partial \nu$ , the normal derivative of  $u$ , are prescribed on  $B$ , the problem is called a *Neumann problem*. A more general problem is that in which eq. (56) has to be solved, given that the values of

$$h_1 u + h_2 \frac{\partial u}{\partial \nu}$$

are prescribed on  $B$ . Here  $h_1$  and  $h_2$  are known functions. If  $h_2 = 0$ , it is a Dirichlet problem; if  $h_1 = 0$ , a Neumann problem; if neither is identically zero, it is a *mixed problem*.

**Choice of Method.** The standard procedure for solving a partial differential equation numerically is to place a rectangular mesh on  $R$ , to replace the differential equation at each mesh point by a finite difference approximation, and thus obtain a set of linear equations. The main difficulty in this procedure occurs in the process of solving the set of linear equations. Inverting the matrix of this set of linear equations is usually not convenient because the matrix is generally ill conditioned. A "marching" process such as that used for solving hyperbolic equations by which the values in the lines of the mesh are determined in succession from the values on the preceding lines is not feasible because the values on any line depend on the values of *two* preceding lines and the boundary data are not sufficient to determine the values on two successive lines. Because of these considerations, the method most frequently used for solving the linear equations is an *iteration or relaxation* method.

**Iteration Procedure.** In order to discuss iteration methods, some notation is needed. Suppose one wishes to solve a system of equations in  $p$  unknowns. Let  $x$  denote a  $p$ -dimensional vector whose components are the  $p$  unknowns  $x_1, x_2, \dots, x_p$ ; let the  $p \times p$  matrix,  $K$ , of the coefficients

of the unknowns be nonsingular, and let  $b$  be a vector whose components are the  $p$  nonhomogeneous terms in the set of equations. Then write the system of equations as follows:

$$(58) \quad Kx = b.$$

To solve this system by iteration, put

$$K = N - P,$$

where  $N$  and  $P$  are any matrices whose difference is  $K$ , and write (58) as

$$Nx = Px + b.$$

Make an estimate of the value of  $x$  and call it  $x^{(0)}$ . A better estimate is possibly given by the vector  $x^{(1)}$  such that

$$Nx^{(1)} = Px^{(0)} + b.$$

This process can be continued indefinitely. Define the vector  $x^{(n+1)}$  ( $n = 0, 1, 2, \dots$ ) as the solution of

$$(59) \quad Nx^{(n+1)} = Px^{(n)} + b, \quad n = 0, 1, 2, \dots,$$

and hope that the sequence of vectors  $x^{(n)}$  converges in the limit to the desired vector  $x$ .

The iteration method defined here is completely general in that the splitting of the matrix  $K$  into two matrices  $N$  and  $P$  was arbitrary. Each distinct split gives a different iteration procedure. There are, however, two restrictions on the ways of splitting  $K$ .

(a) To find  $x^{(n+1)}$  from eq. (59) more easily than to find  $x$  from (58),  $N$  must be a matrix with an easily found inverse. For *example*,  $N$  might be a diagonal matrix or a lower triangular matrix.

(b) For the iteration scheme to converge, it is required that all the eigenvalues of the matrix  $N^{-1}P$  be in absolute value less than 1. It can be shown that this is a necessary and sufficient condition for the sequence  $x^{(n)}$  to converge to  $x$ , no matter what the original guess  $x^{(0)}$  is.

**Richardson and Liebmann Iteration Methods.** The ideas of the preceding section will be illustrated by applying them to the solution of *Poisson's equation*

$$(60) \quad \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y)$$

inside the unit square when  $0 \leq x \leq 1, 0 \leq y \leq 1$ . Assume that the values of  $u(x, y)$  are given on the boundary of the square.

Put a square mesh of width  $1/p$  over the unit square and let

$$u\left(\frac{i}{p}, \frac{j}{p}\right) = u_{ij},$$

$$f\left(\frac{i}{p}, \frac{j}{p}\right) = f_{ij}, \quad i, j = 0, 1, \dots, p.$$

By the use of the well-known finite difference approximation for the Laplacian (see eq. 55), eq. (60) becomes

$$(61) \quad u_{i-1,j} + u_{i,j-1} - 4u_{i,j} + u_{i+1,j} \\ + u_{i,j+1} = \frac{1}{p^2}f_{ij}, \quad i, j = 1, 2, \dots, p-1.$$

Since the values of  $u_{0j}$ ,  $u_{pj}$  ( $j = 0, 1, \dots, p$ ) and of  $u_{i0}$ ,  $u_{ip}$  ( $i = 0, 1, \dots, p$ ) are given, (61) is a system of  $(p-1)^2$  equations for the  $(p-1)^2$  unknowns  $u_{ij}$  ( $i, j = 1, 2, \dots, p-1$ ).

In *Richardson's method* for solving eq. (61), the following iteration scheme is used:

$$(62) \quad 4u_{i,j}^{(n+1)} = u_{i-1,j}^{(n)} + u_{i,j-1}^{(n)} + u_{i+1,j}^{(n)} \\ + u_{i,j+1}^{(n)} - \frac{1}{p^2}f_{ij}, \quad n = 0, 1, 2, \dots.$$

The values of  $u_{ij}^{(0)}$  are of course the initial guess to the solution of eq. (60). If this method is compared with that in eq. (59), it is apparent that the split is such that  $N$  is a diagonal matrix.

One disadvantage of Richardson's method when an electronic computer is used is that all the previous values of  $u_{ij}^{(n)}$  must be stored until all the new values of  $u_{ij}^{(n+1)}$  are found. This disadvantage is avoided in *Liebmann's method* where the new value of  $u_{ij}^{(n+1)}$  is calculated by using as many new values as are available. Thus, if the values of  $u_{ij}$  are calculated in order along each row from left to right and the rows in order from bottom to top, the following iteration scheme would be used:

$$(63) \quad 4u_{i,j}^{(n+1)} = u_{i-1,j}^{(n+1)} + u_{i,j-1}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j+1}^{(n)} - \frac{1}{p^2}f_{ij}.$$

It can be proved that the method defined by eq. (63) would converge twice as fast as that defined by eq. (62).

The rate of convergence can be still further improved by using an extrapolation parameter  $\alpha$ , thus obtaining what is called *Liebmann's extrapolated*

method. The iteration scheme is now this:

$$4u_{ij}^{(n+1)} = 4(1 - \alpha)u_{ij}^{(n)} + \alpha \left[ u_{i-1,j}^{(n+1)} + u_{i,j-1}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j+1}^{(n)} - \frac{1}{p^2}f_{ij} \right].$$

The value of  $\alpha$  for which *convergence is fastest* is found by solving the equation

$$\alpha^2 t_m^2 - 4\alpha + 4 = 0,$$

where  $t_m$  is the largest eigenvalue of the Richardson scheme eq. (62). For the case considered,  $t_m = \cos(\pi/p)$ . For a rectangular mesh with  $p$  divisions in one direction and  $q$  divisions in the other,

$$t_m = \frac{1}{2}[\cos(\pi/p) + \cos(\pi/q)].$$

In general the use of the extrapolated Liebmann method with the best value of  $\alpha$  will be much faster than the unextrapolated Liebmann method.

**Line Iteration Schemes.** Another iteration method which is useful in many cases is given by the following scheme:

$$(64) \quad 4u_{ij}^{(n+1)} - u_{i,j-1}^{(n+1)} - u_{i,j+1}^{(n+1)} = u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n)} - \frac{1}{p^2}f_{ij}.$$

In this scheme instead of solving for the values of  $u$  at a point  $ij$ , solve for all values of  $u$  on the  $i$ th column in terms of the values of  $u$  on the  $(i - 1)$ -th and  $(i + 1)$ -th column. That is why eq. (64) has been written with the left-hand side containing all the  $u$ -values on the  $i$ th column. Since at each step the value of the right-hand side is known for all values of  $j$ , the three-term relation defined by eq. (64) is solved for the values of  $u$  on the  $i$ th column. (The method of solving the three-term relation is explained in the subsection on Hyperbolic Partial Differential Equations.)

Instead of solving for the values of  $u$  on a column, one may solve for the values of  $u$  on a row. In that case use the following scheme:

$$(65) \quad 4u_{ij}^{(n+1)} - u_{i+1,j}^{(n+1)} - u_{i-1,j}^{(n+1)} = u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n)} - \frac{1}{p^2}f_{ij}.$$

Again this three-term recurrence scheme is solved for the values of  $u$  on the  $j$ th row starting with  $j = 1$ .

**The Method of Peaceman and Rachford** (Ref. 35). This seems to be one of the quickest iterative methods for solving an elliptic differential

equation. It is essentially a line iteration scheme which uses columns and rows alternately. The explicit description of the method is contained in the following formulas:

$$\begin{aligned} u_{i-1,j}^{(2n+1)} - (2 + \rho_n)u_{ij}^{(2n+1)} + u_{i+1,j}^{(2n+1)} \\ = -u_{i,j-1}^{(2n)} + (2 - \rho_n)u_{ij}^{(2n)} - u_{i,j+1}^{(2n)} + \frac{1}{p^2}f_{ij}; \end{aligned}$$

$$\begin{aligned} u_{i,j-1}^{(2n+2)} - (2 + \rho_n)u_{ij}^{(2n+2)} + u_{i,j+1}^{(2n+2)} \\ = -u_{i-1,j}^{(2n+1)} + (2 - \rho_n)u_{ij}^{(2n+1)} - u_{i+1,j}^{(2n+1)} + \frac{1}{p^2}f_{ij}. \end{aligned}$$

Here  $\rho_n$  is an extrapolation parameter which is to be determined so that the method will converge as quickly as possible. In the present case Peaceman and Rachford suggest putting  $\rho_n = \rho_k$  if  $n \equiv k \pmod{p}$ , where

$$\rho_k = 4 \sin^2 \frac{(2k+1)\pi}{4p}.$$

**Variational Principle.** An important characteristic of elliptic differential equations is that they can be obtained as the Euler equations of problems in the calculus of variations. Physically, this implies that the problem possesses an energy integral whose minimum value is given by the solution of the elliptic partial differential equation. For *example*, in Dirichlet's problem the integral

$$(66) \quad \iint_R (\nabla u)^2 dx dy$$

must be a minimum in the domain of all functions  $u$  satisfying the preassigned boundary conditions. In problems with mixed boundary conditions the integral (66) must be modified (for details see Ref. 37). As another *example*, in elasticity problems involving plates, the integral

$$(67) \quad \iint (\Delta u)^2 dx dy$$

must be a minimum in the domain of all functions  $u$  satisfying the preassigned boundary conditions.

For numerical purposes the energy integral can be approximated by a sum involving the values of the unknown function  $u$  at a set of points inside the region  $R$ . Then choose values of  $u$  so that the sum will be a mini-

mum. For *example*, (66) would be approximated by

$$(68) \quad \sum_i \sum_j [(u_{ij} - u_{i-1,j})^2 + (u_{ij} - u_{i,j-1})^2],$$

and (67) by

$$(69) \quad \sum_i \sum_j [u_{ij} - \frac{1}{4}(u_{i,j+1} + u_{i,j-1} + u_{i+1,j} + u_{i-1,j})]^2.$$

As these illustrations show, the sum is a quadratic form in the values  $u_{ij}$ . By differentiation with respect to  $u_{ij}$  a set of linear equations is obtained whose solution will make the sum a minimum. Simple algebra shows that when this method is applied to eqs. (68) and (69) the standard difference equations for Laplace's equation or the biharmonic equation are obtained.

Any iteration method which at each step reduces the value of the sum must automatically converge to a minimum value. Use of this fact easily shows that the various schemes proposed above do converge to a solution. The variational principle is also useful in determining how the boundary conditions should be taken into account.

### Hyperbolic Partial Differential Equations

The equation of the *vibrating string* will be used to illustrate some finite difference methods for solving problems involving a second order hyperbolic partial differential equation.

If the end points of the string are held fixed at  $x = 0$  and  $x = 1$ , the deflection of the string  $u(x, t)$  is determined from the initial deflection  $u(x, 0) = f(x)$ , and the initial velocity  $u_t(x, 0) = g(x)$ .

The conditions describing the motion are:

$$(70) \quad \begin{aligned} \text{P.D.E.} \quad & \frac{1}{c^2} u_{tt} = u_{xx}, \quad \text{for } 0 < x < 1, t > 0. \\ \text{I.C.} \quad & u(x, 0) = f(x) \\ & u_t(x, 0) = g(x), \quad \text{for } 0 < x < 1, t = 0. \\ \text{B.C.} \quad & u(0, t) = u(1, t) = 0, \quad \text{for } t \geq 0. \end{aligned}$$

(Note the abbreviations P.D.E. for partial differential equation, I.C. for initial conditions, B.C. for boundary conditions and  $c^2$  for the ratio of the string tension to the string mass per unit length.)

The finite difference methods are based on the replacement of derivatives by difference quotients in a rectangular lattice over the  $(x, t)$ -plane. The mesh points of the lattice are  $(x_i, t_j)$ , where  $x_i = ih$ ,  $t_j = jk$ , with  $i$  and  $j$  taking on integer values. Let the  $x$  interval be such that  $Nh = 1$  with

some integer  $N$ , and let  $k/h = \lambda$  be the ratio of mesh widths. Figure 1 schematically indicates the lattice work with  $N = 6$ ,  $\lambda = 1$ .

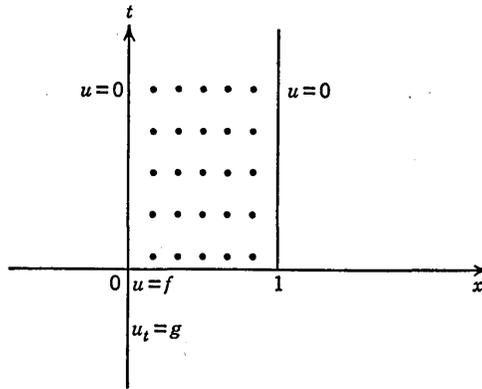


FIG. 1. Mesh points of lattice for  $N = 6$ ,  $\lambda = 1$ .

**Explicit Finite Difference Scheme.** The simplest replacement of eq. (70) by finite difference equations is found by using central differences for derivatives as follows:

$$\begin{aligned} \text{P.D.E.} \quad & \frac{U(x_i, t_{j+1}) - 2U(x_i, t_j) + U(x_i, t_{j-1}))}{c^2 k^2} \\ & = \frac{U(x_{i+1}, t_j) - 2U(x_i, t_j) + U(x_{i-1}, t_j)}{h^2}, \end{aligned}$$

for  $1 \leq i \leq N - 1, j \geq 1$ .

$$\text{I.C.} \quad \begin{cases} U(x_i, 0) = f(x_i) \\ U(x_i, k) = f(x_i) + kg(x_i) + \frac{k^2 c^2}{2h^2} [f(x_{i+1}) - 2f(x_i) + f(x_{i-1}))], \end{cases}$$

for  $1 \leq i \leq N - 1$ .

$$\text{B.C.} \quad U(0, t_j) = U(N, t_j) = 0, \quad \text{for } j \geq 0.$$

The initial condition which determines  $U$  at the first line  $t = k$  is derived from Taylor's expansion

$$u(x, k) = u(x, 0) + ku_t(x, 0) + \frac{k^2}{2} u_{tt}(x, 0)$$

together with the relation

$$u_{tt}(x, 0) = c^2 u_{xx}(x, 0) = c^2 \frac{d^2 f}{dx^2}.$$

With the notation  $f_i = f(x_i)$ ,  $U_{i,j} = U(x_i, t_j)$ , the equation may be re-written to exhibit the explicit manner in which the solution can be calculated.

$$\begin{aligned}
 \text{P.D.E.} \quad & U_{i,j+1} = c^2\lambda^2 U_{i+1,j} + (2 - 2c^2\lambda^2)U_{i,j} \\
 & + c^2\lambda^2 U_{i-1,j} - U_{i,j-1}, \quad \text{for } j \geq 1, 1 \leq i \leq N - 1, \\
 (71) \quad \text{I.C.} \quad & \begin{cases} U_{i,0} = f_i \\ U_{i,1} = \frac{c^2\lambda^2}{2} f_{i+1} + (1 - c^2\lambda^2)f_i + \frac{c^2\lambda^2}{2} f_{i-1} + kg_i, \end{cases} \\
 \text{B.C.} \quad & U_{0,j} = U_{N,j} = 0, \quad \text{with } \lambda = \frac{k}{h}.
 \end{aligned}$$

**Convergence of Finite Difference Scheme.** It is not clear a priori, that the solution  $U$  of (71) will converge to the solution  $u$  of eq. (70) as  $h, k \rightarrow 0$ .

In fact if as  $h, k \rightarrow 0$ ,  $\lambda$  is held fixed, and such that  $\lambda > 1/c$ , the solution  $U$  will *not* converge to  $u$  for all initial displacements. This may be verified by noting that the solution of the initial value problem for the infinite string is given by

$$(72) \quad u(x, t) = \frac{1}{2}[f(x + ct) + f(x - ct)] + \frac{1}{2c} \int_{x-ct}^{x+ct} g(\xi) d\xi.$$

Formula (72) shows that the solution  $u(x, t)$  depends solely on the initial data in the interval  $(x - ct, x + ct)$ . The solution  $U(x, t)$  depends solely on the initial data in the interval  $[x - (t/\lambda), x + (t/\lambda)]$ . Hence if  $\lambda > 1/c$ , it is possible to vary  $f$  and  $g$  in the intervals  $[x - ct, x - (t/\lambda)]$  and  $[x + (t/\lambda), x + ct]$  in such a way that the solution  $u(x, t)$  is changed, but yet  $U(x, t)$  is unaffected. Hence if  $\lambda > 1/c$ , the solution  $U(x, t)$  cannot converge as  $h, k \rightarrow 0$  since it would have to converge to different values.

Hence it is necessary for convergence that  $\lambda \leq 1/c$ , i.e., the "domain of dependence" for the solution of the finite difference equation should contain the domain of dependence of the solution of the differential equation. In fact, if  $\lambda \leq 1/c$ ,  $U$  does converge to  $u$  as  $h, k \rightarrow 0$ . The proof of convergence may be made to rest upon the Fourier series representations of the solutions of eqs. (71) and (70), namely,

$$U(x, t) = \sum_{n=1}^{\infty} (A_n \cos \mu_n t + B_n \sin \mu_n t) \sin nx,$$

where  $\mu_n$  is determined from the condition  $\sin \mu_n k/2 = \lambda c \sin n h/2$ ; and

$$u(x, t) = \sum_{n=1}^{\infty} (a_n \cos nct + b_n \sin nct) \sin nx.$$

**Roundoff and Truncation Errors.** The calculation of  $U$  is in practice effected by rounding to a finite number of decimal places. The equations which determine  $U$  are

$$\begin{aligned} U_{i,0} &= f_i + R_{i,0}, \\ U_{i,1} &= \frac{c^2\lambda^2}{2}f_{i+1} + (1 - c^2\lambda^2)f_i + \frac{c^2\lambda^2}{2}f_{i-1} + kg_i + R_{i,1}, \\ U_{i,j+1} &= c^2\lambda^2 U_{i+1,j} + (2 - 2c^2\lambda^2)U_{i,j} + c^2\lambda^2 U_{i-1,j} - U_{i,j-1} + R_{i,j+1}, \\ U_{0,j} &= U_{N,j} = 0, \end{aligned}$$

where  $R_{p,q}$  is the roundoff error. The truncation error  $T_{p,q}$  is defined by substituting  $u$  into eq. (71) as follows:

$$\begin{aligned} u_{i,0} &= f_i + T_{i,0}, \\ u_{i,1} &= \frac{c^2\lambda^2}{2}f_{i+1} + (1 - c^2\lambda^2)f_i + \frac{c^2\lambda^2}{2}f_{i-1} + kg_i + T_{i,1}, \\ u_{i,j+1} &= c^2\lambda^2 u_{i+1,j} + (2 - 2c^2\lambda^2)u_{i,j} + c^2\lambda^2 u_{i-1,j} - u_{i,j-1} + T_{i,j+1}. \end{aligned}$$

It is easily verified that  $T_{i,0} = 0$ ,  $T_{i,1} = O(k^3)$ ,  $T_{i,j} = O(k^4)$  where  $O(k^n)$  represents a quantity which is bounded in absolute value for all sufficiently small  $k$  by  $Mk^n$  with some constant  $M$ .

It is reasonable to require that the roundoff error be of the same order of magnitude as the truncation error or smaller, in order that the number of digits carried in the calculation be appropriate for the interval size. With this restriction, the total error  $e_{i,j}$  is  $O(T^2k^2)$  for any finite time,  $T$ , where

$$e_{i,j} = u_{i,j} - U_{i,j} \quad \text{and} \quad 0 \leq j \leq T/k.$$

**Implicit Schemes.** The restriction  $k/h = \lambda \leq 1/c$  may be relaxed by using an implicit scheme. That is, it is possible to take larger time steps at the expense of more involved calculations as follows:

$$\begin{aligned} \text{P.D.E.} \quad \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{k^2} &= \frac{c^2}{h^2} [\alpha^2 (U_{i+1,j+1} - 2U_{i,j+1} \\ &+ U_{i-1,j+1}) + (1 - 2\alpha^2)(U_{i+1,j} - 2U_{i,j} + U_{i-1,j}) \\ &+ \alpha^2 (U_{i+1,j-1} - 2U_{i,j-1} + U_{i-1,j-1})]. \end{aligned}$$

The equations when solved for the unknowns at the  $(j + 1)$ -st time step have the form

$$(73) \quad (\alpha\lambda)^2 U_{i+1,j+1} - [1 + 2(\alpha\lambda)^2] U_{i,j+1} + (\alpha\lambda)^2 U_{i-1,j+1} = W, \quad \text{for } i = 1, 2, \dots, N - 1,$$

where  $W$  involves information on the two preceding lines ( $j$  and  $j - 1$ ).

The labor involved in solving eq. (73) is minimal since the  $(N - 1) \times (N - 1)$  matrix of coefficients is in triple diagonal form. At the same time the condition on  $\lambda$  which insures convergence of  $U$  to  $u$  as  $h, k \rightarrow 0$  is

$$\lambda^2 c^2 \leq \frac{1}{1 - 4\alpha^2}, \quad \text{for } 0 \leq \alpha^2 < \frac{1}{4},$$

and no restriction for  $\frac{1}{4} \leq \alpha^2$ .

**Solution of Triple Diagonal Systems.** The equations

$$\begin{array}{rccccccc} b_1 x_1 + c_1 x_2 + 0 & + \dots & & + 0 & & = & y_1 \\ a_2 x_1 + b_2 x_2 + c_2 x_3 + 0 & + \dots & & + 0 & & = & y_2 \\ 0 & + a_3 x_2 + b_3 x_3 + c_3 x_4 + 0 & & + \dots & + 0 & = & y_3 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & + \dots & + 0 & + a_{N-1} x_{N-2} + b_{N-1} x_{N-1} + c_{N-1} x_N & = & y_{N-1} \\ 0 & + \dots & + 0 & + a_N x_{N-1} & + b_N x_N & = & y_N \end{array}$$

may be solved by eliminating the unknowns in succession from the equations. By starting at the top the system can be put in the form

$$\begin{array}{rccccccc} x_1 + C_1 x_2 + 0 & + \dots & + 0 & & = & Y_1 \\ 0 & + x_2 & + C_2 x_3 + 0 \dots & + 0 & & = & Y_2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & + \dots & + 0 & + x_{N-1} + C_{N-1} x_N & = & Y_{N-1} \\ 0 & + \dots & + 0 & + x_N & = & Y_N \end{array}$$

The numbers  $C_K$  and  $Y_K$  may be recursively computed from the formulas

$$(74) \quad \begin{aligned} C_1 &= \frac{c_1}{b_1}; & Y_1 &= \frac{y_1}{b_1}. \\ C_K &= \frac{c_K}{b_K - a_K C_{K-1}}; & Y_K &= \frac{y_K - a_K Y_{K-1}}{b_K - a_K C_{K-1}}, \\ & & & \text{for } K = 2, 3, \dots, N. \end{aligned}$$

It is now easy to solve for the  $x_K$  beginning with  $x_N$  as follows:

$$(75) \quad \begin{aligned} x_N &= Y_N, \\ x_K &= Y_K - C_K x_{K+1}, \quad \text{for } K = N - 1, N - 2, \dots, 1. \end{aligned}$$

**The von Neumann Criterion for Convergence.** A quick method for heuristically testing the convergence of a finite difference method has been attributed to von Neumann.

In the case of linear differential equations with variable coefficients, the method consists in replacing the coefficients by constants and then finding all solutions of the difference equation of the form

$$U(x, t) = e^{\gamma t} e^{i\beta x}, \quad \text{with } \beta \text{ real.}$$

If  $|e^{\gamma t}| \leq 1$  for  $t \geq 0$ , for all  $\beta$  and for all admissible values of the coefficients, the finite difference method is said to be stable, otherwise not.

The von Neumann "test" for convergence is the same as for stability. In practice, this test for convergence is as simple as any a priori calculation could be; in addition, it has been shown to be a sufficient condition for convergence for a large number of cases.

### Parabolic Partial Differential Equations

Finite difference methods for parabolic equations are similar to those for hyperbolic equations. The present discussion will be restricted to equations of the first order in time and second order in one or more variables. Illustrative methods will be given for:

- (a) The linear heat flow equation in one dimension.
- (b) A quasilinear equation in one space variable and time.
- (c) A linear parabolic equation in two space variables and time.

For diffusion or heat flow in one dimension there is an initial value problem consisting of a partial differential equation (P.D.E.), an initial condition (I.C.), and boundary conditions (B.C.) for a function  $u(x, t)$ . In the simplest case, these are:

$$(76) \quad \begin{aligned} \text{P.D.E.} \quad & \frac{1}{\sigma} u_t = u_{xx}, \quad \text{for } 0 < x < 1, t > 0 \\ \text{I.C.} \quad & u(x, 0) = f(x), \quad \text{for } 0 < x < 1, t = 0 \\ \text{B.C.} \quad & u(0, t) = u(1, t) = 0, \quad \text{for } t \geq 0, \end{aligned}$$

where  $\sigma$  is a positive constant.

Just as for the vibrating string problem one introduces a lattice of net points  $(x_i, t_j)$  in the  $(x, t)$ -plane, with  $x_i = ih$ ,  $t_j = jk$ , where  $i$  and  $j$  are inte-

gers,  $h$  and  $k$  are increments of distance and time, respectively, and  $h$  is such that  $Nh = 1$  for some integer  $N$ . By using the notation  $U_{i,j} = U(x_i, t_j)$  as before, where  $U(x, t)$  is the approximation to  $u(x, t)$  to be obtained from the numerical calculation, the system (76) is replaced by the difference equations

$$\text{P.D.E.} \quad \frac{U_{i,j+1} - U_{i,j}}{\sigma k} = \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2},$$

(77) for  $1 \leq i \leq N - 1, j \geq 1$

I.C.  $U_{i,0} = f(x_i), \quad \text{for } 1 \leq i \leq N - 1$

B.C.  $U_{0,j} = U_{N,j} = 0, \quad \text{for } j \geq 0.$

*These equations are explicit and their solution converges to the solution of the problem (76) as the mesh is refined, provided that  $h, k \rightarrow 0$  in such a way that*

$$(78) \quad \frac{\sigma k}{h^2} = \text{Constant} \leq \frac{1}{2}.$$

**Convergence.** By convergence is meant the following. One considers a sequence of calculations with progressively finer meshes and one supposes that  $x, t$  is some fixed point common to the lattices of infinitely many of these calculations. Then  $U(x, t) \rightarrow u(x, t)$  as  $h, k \rightarrow 0$  under the above restrictions. This convergence occurs if the initial function  $f(x)$  is any function continuous in the interval  $0 \leq x \leq 1$ . If the condition (78) is violated by taking  $\sigma k/h^2 = \text{constant} > \frac{1}{2}$  as  $h, k \rightarrow 0$ , the solution of the difference eqs. (77) diverges for all but special cases in which the initial function  $f(x)$  has a terminating Fourier series. One says that the equations are *unstable* under these circumstances and that (78) is a *condition for stability*. For general discussions of convergence and stability, see Refs. 39, 40, and 43.

The convergence as  $k \rightarrow 0$  is slower, at least in a formal sense, than it is for the hyperbolic problem, and its rate depends upon the smoothness of the initial function  $f(x)$ . *If condition (78) is satisfied and  $f(x)$  is analytic for  $0 \leq x \leq 1$ , the error  $e_{i,j}$  of the approximation (77) is  $O(Tk)$  for  $t$  in a finite interval  $0 \leq t \leq T$ . One can of course also write  $e_{i,j} = O(Th^2)$  because of the relation (78).*

The method is more accurate in the special case in which  $h$  and  $k$  are so chosen that

$$(79) \quad \frac{\sigma k}{h^2} = \frac{1}{6}.$$

It is easy to verify by Taylor's series expansions that in this case there is

a cancellation of the first order error terms coming from the two members of the first eq. (77). In consequence, if  $f(x)$  is analytic,  $e_{ij} = O(Tk^2) = O(Th^4)$ .

A condition of the form (78) is perhaps not unexpected from the point of view of the domain of dependence of the differential equation, which is not confined to a small interval as it was for the vibrating string problem. That is,  $u(x, t)$ , for any  $t > 0$ , depends on *all* the initial data, i.e., on the values of  $f(x)$  for the entire interval  $0 \leq x \leq 1$ . For any finite values of  $h$  and  $k$  the difference equations of course possess a restricted domain of dependence, but as the mesh is refined this domain opens out (because eq. 78 requires that  $k$  vary as  $h^2$ ) so as to include all past values of the function.

**Implicit Equations.** Implicit difference equations can be constructed for the heat flow problem in many ways. For *example*, one can replace the first eq. (77) by the equation

$$(80) \quad \frac{U_{i,j+1} - U_{i,j}}{\sigma k} = \frac{1}{h^2} [\alpha(U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1}) + (1 - \alpha)(U_{i+1,j} - 2U_{i,j} + U_{i-1,j})],$$

where  $\alpha$  is a constant. The resulting method reduces to the foregoing explicit method for  $\alpha = 0$ , to the so-called Crank-Nicholson method (Ref. 41) for  $\alpha = \frac{1}{2}$  and to the method of Laasonen (Ref. 42) for  $\alpha = 1$ .

*The condition for convergence of the solutions of eq. (80) as  $h, k \rightarrow 0$  is*

$$(81a) \quad \frac{\sigma k}{h^2} \leq \frac{1}{2 - 4\alpha} \quad \text{if } 0 \leq \alpha < \frac{1}{2}$$

$$(81b) \quad \text{No restriction if } \frac{1}{2} \leq \alpha \leq 1.$$

The *Crank-Nicholson and Laasonen schemes*, in particular, are therefore *unconditionally stable*.

The optimum value of  $\alpha$  from the point of view of accuracy is, surprisingly enough, not  $\alpha = \frac{1}{2}$ , but rather

$$\alpha = \frac{1}{2} - \frac{h^2}{12\sigma k}$$

(which includes eq. 79 as a special case). It is easily verified that the condition eq. (81a) for stability is satisfied in this case.

The *implicit equations* resulting from the use of eq. (80) can be solved by the algorithm described above for the hyperbolic problem, that is, by the use of eqs. (74) and (75).

**Variable Coefficients and Nonlinear Problems.** The procedure can be generalized so as to apply to problems with variable coefficients and

nonlinear problems. For *example*, the quasilinear equation

$$(82) \quad \frac{\partial u}{\partial t} = f(x, t, u) \frac{\partial^2 u}{\partial x^2} + g(x, t, u) \frac{\partial u}{\partial x} + S(x, t, u),$$

where  $f(x, t, u) \geq \text{constant} > 0$  for  $0 < x < 1, 0 < t < T$ , can be replaced by the difference equation

$$(83) \quad \frac{U_{i,j+1} - U_{i,j}}{k} = \frac{f_{i,j}}{h^2} [\alpha(U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1}) + (1-\alpha)(U_{i+1,j} - 2U_{i,j} + U_{i-1,j})] + \frac{g_{i,j}}{h} (U_{i+1,j} - U_{i,j}) + S_{i,j},$$

which was abbreviated by writing  $f_{i,j}$  for  $f(x_i, t_j, U_{i,j})$ , and similarly with respect to  $g$  and  $S$ . Suppose that the initial and boundary conditions are given the same treatment as before. Alternative approximations to eq. (82) can be obtained by replacing the coefficient of  $g_{i,j}/h$  by the backward difference  $U_{i,j} - U_{i-1,j}$  or by expressions such as

$$\text{or} \quad \frac{1}{2}(U_{i+1,j} - U_{i-1,j})$$

$$\frac{1}{4}(U_{i+1,j} + U_{i+1,j+1} - U_{i-1,j} - U_{i-1,j+1}),$$

and in many other ways.

If  $g(x, t, u)$  and  $S(x, t, u)$  are zero, one expects the condition for stability to be

$$f(x, t, u(x, t)) \frac{k}{h^2} \leq \frac{1}{2 - 4\alpha}, \text{ for } 0 \leq x \leq 1, \quad 0 \leq t \leq T, \text{ if } 0 \leq \alpha < \frac{1}{2}.$$

$$\text{No restriction,} \quad \text{if } \frac{1}{2} \leq \alpha \leq 1,$$

in analogy with the simple heat flow problem. Generally the effect of the low order terms on stability is merely to replace the  $\leq$  sign in the first line by the  $<$  sign. This change has of course no significance for machine calculation where roundoff destroys the possibility of distinguishing between exact and near equality. But to be on the safe side, a conservative statement is given. *Calculation based on eq. (83) converges as  $h, k \rightarrow 0$  provided that*

$$(84) \quad \max_{x,t,h,k} f(x, t, u(x, t)) \frac{k}{h^2} \begin{cases} < \frac{1}{2 - 4\alpha} & \text{if } 0 \leq \alpha < \frac{1}{2} \\ < \infty & \text{if } \frac{1}{2} \leq \alpha \leq 1. \end{cases}$$

No mathematically rigorous proof of convergence has been given; the statement made above is based on simple intuitive arguments coupled with

rigorous discussion of the linear cases with constant coefficients, with empirical evidence from many calculations, with Fritz John's rigorous treatment (Ref. 44) of a wide class of difference equations which includes as a special case eq. (83) when  $f$  and  $g$  are independent of  $u$  and  $\alpha$  is taken equal to zero, and with Douglas' rigorous treatment (Ref. 45) of the case in which  $\alpha = 1, g = 0$ .

Note that in the nonlinear case one cannot guarantee stability in advance: rather, the inequality (84) must be verified constantly as the calculation proceeds and, if the solution at any instant threatens to violate the condition, the time interval  $k$  must be reduced.

**Parabolic Equation in Two-Space Variables.** As a last *example*, the simple parabolic equation in two-space variables for a function  $u = u(x, y, t)$

$$(85) \quad \frac{\partial u}{\partial t} = A \frac{\partial^2 u}{\partial x^2} + 2B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2}$$

where  $A, B, C$  are constants such that

$$A > 0, C > 0, \quad AC - B^2 > 0$$

can be approximated by an analog of the general implicit eq. (80). Because of the number of variables it is convenient to introduce a slightly different notational convention for this problem by calling the increments  $\Delta t, \Delta x, \Delta y$  and by relating the time variable  $t$  to a superscript  $n$  as follows. Let

$$U_{j,l}^n = U(j \Delta x, l \Delta y, n \Delta t),$$

$$\begin{aligned} \Phi_{j,l}^n &= \frac{A}{(\Delta x)^2} (U_{j+1,l}^n - 2U_{j,l}^n + U_{j-1,l}^n) \\ &+ \frac{B}{2 \Delta x \Delta y} (U_{j+1,l+1}^n - U_{j-1,l+1}^n - U_{j+1,l-1}^n + U_{j-1,l-1}^n) \\ &+ \frac{C}{(\Delta y)^2} (U_{j,l+1}^n - 2U_{j,l}^n + U_{j,l-1}^n). \end{aligned}$$

With these abbreviations, the approximation to eq. (85) is

$$(86) \quad \frac{U_{j,l}^{n+1} - U_{j,l}^n}{\Delta t} = \alpha \Phi_{j,l}^{n+1} + (1 - \alpha) \Phi_{j,l}^n.$$

If  $\alpha = 0$ , the equations are explicit. If  $\alpha \neq 0$ , they are implicit; in this case the simultaneous equations to be solved at each stage of the calculation are for unknowns  $U_{j,l}^{n+1}$  where, for example,  $j = 1, 2, \dots, J, l = 1, 2, \dots, L$ , with  $n$  fixed. These simultaneous equations are of elliptic character and are usually solved by one of the relaxation methods.

TABLE 12. CHARACTERISTICS OF SOME NUMERICAL METHODS OF SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS

( $w$  = storage requirements for computer program and working storage. Additional storage requirements given in terms of dimension  $N$  of lattice.)

$$\text{Elliptic}^a: \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y)$$

Method	Storage	Evaluation of One $u_{ij}^n$ , Time $b$	Total Calculating Time $b$
Richardson (eq. 62)	$2N^2 + w$	$4\tau$	$11N^4\tau$
Liebmann (eq. 63)	$N^2 + w$	$4\tau$	$5.6N^4\tau$
Extrapolated Liebmann (optimum $\alpha$ )	Same as Liebmann	$7\tau$	$15N^3\tau$
Peaceman and Rachford	$N^2 + w$	$9\tau$	$(34N^2 \log_e N)\tau$

$$\text{Hyperbolic: } \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \quad (c = \text{constant})$$

Method	Storage	Stability Condition	Truncation Error for 1 Step $c$
Explicit (eq. 71)	$2N + w$	$\lambda \leq \frac{1}{c}$	$O(k^4)$
Implicit (eq. 73)	$4N + w$	$\begin{cases} \lambda^2 c^2 \leq \frac{1}{1 - 4\alpha^2} \\ \text{if } 0 \leq \alpha^2 < \frac{1}{4} \\ \text{Unconditionally stable} \\ \text{if } \frac{1}{4} \leq \alpha^2 \end{cases}$	$O(k^4)$

$$\text{Parabolic: }^d \frac{1}{\sigma} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad (\sigma = \text{constant})$$

Explicit (eq. 77, or eq. 80 with $\alpha = 0$ )	$N + w$	$\frac{\sigma k}{h^2} \leq \frac{1}{2}$	$\begin{cases} O(k) \\ O(k^2) \text{ if } \frac{\sigma k}{h^2} = \frac{1}{6} \end{cases}$
Crank-Nicholson (eq. 80 with $\alpha = \frac{1}{2}$ )	$4N + w$	Unconditionally stable	$O(k)$
Laasonen (eq. 80 with $\alpha = 1$ )	$4N + w$	Unconditionally stable	$O(k)$
Generalized implicit method (eq. 80)	$4N + w$	$\begin{cases} \frac{\sigma k}{h^2} < \frac{1}{2 - 4\alpha} \\ \text{if } 0 \leq \alpha < \frac{1}{2} \\ \text{Unconditionally stable} \\ \text{if } \frac{1}{2} \leq \alpha \leq 1 \end{cases}$	$\begin{cases} O(k); O(k^2) \\ \text{if } \alpha = \frac{1}{2} - \frac{h^2}{12\sigma k} \end{cases}$ $O(k)$

<sup>a</sup>(1) An  $N \times N$  lattice is assumed. (2) One has the option, of course, of computing  $f_{ij}$  or storing a table of  $f_{ij}$ . In the latter case the storage is increased by  $N^2$ , whereas in the

[Footnote continued on p. 14-88.]

The stability condition for eq. (86) is

$$(87) \quad \frac{A \Delta t}{(\Delta x)^2} + \frac{C \Delta t}{(\Delta y)^2} \leq \frac{1}{2 - 4\alpha}, \quad \text{if } 0 \leq \alpha < \frac{1}{2},$$

$$\text{No restriction,} \quad \text{if } \frac{1}{2} \leq \alpha \leq 1.$$

Equation (86) can be generalized so as to apply to problems with variable coefficients and to quasilinear problems, as well as to problems in three or more space variables.

### Summary Table

Table 12 gives some significant attributes of methods from the point of view of digital computer solution.

### REFERENCES

#### *Section 1. Interpolation, Curve Fitting, Differentiation, and Integration*

1. L. M. Milne-Thompson, *The Calculus of Finite Differences*, Macmillan, London, 1951.
2. A. S. Householder, *Principles of Numerical Analysis*, McGraw-Hill, New York, 1953.
3. E. T. Whittaker and G. Robinson, *The Calculus of Observations*, Blackie, London, 1937.
4. R. E. von Holdt and R. J. Brousseau, *Weighted Least Square Polynomial Approximation to a Continuous Function of a Single Value*, University of California Radiation Laboratory, Livermore, 1956.
- 4a. M. Ascher and G. A. Forsythe, SWAC experiments on the use of orthogonal polynomials for data fitting, *J. Assoc. Computing Machinery*, 5, (1958).
5. K. Levenberg, A method for the solution of certain non-linear problems in least squares, *Quart. Appl. Math.*, 2, 164-168 (1944).
6. E. W. Hobson, *Spherical and Ellipsoidal Harmonics*, Cambridge University Press, Cambridge, England, 1931.

#### *Section 2. Matrix Inversion and Simultaneous Linear Equations*

7. G. B. Dantzig, A. Orden, and P. Wolfe, *The Generalized Simplex Method for Mini-*

---

former case, the computation time is increased. The entries in this table include neither storage nor computing time for  $f_{ij}$ . (3) The criteria for evaluating the methods given for solving elliptic equations differ from those for hyperbolic and parabolic, because the methods given for elliptic equations are iterative, whereas those given for parabolic and hyperbolic are not.

<sup>b</sup> See, for example, Refs. 32 and 35. In these columns,  $\tau$  is the mean time per arithmetic operation.

<sup>c</sup> The degree of accuracy shown in this column would not in general hold for nonconstant  $c$ .

<sup>d</sup> See Refs. 41 and 42.

mizing a Linear Form under Linear Inequality Restraints, Rand Paper P-392, Santa Monica, Calif., 1954.

8. G. E. Forsythe, *Theory of Selected Methods of Finite Matrix Inversion and Decomposition*, Lecture notes by D. G. Aronson and K. E. Iverson, Inst. Numerical Anal. Rept. 52-5, Natl. Bur. of Standards, Los Angeles, Calif., multilithed typescript, 1951.

9. A. S. Householder, *Principles of Numerical Analysis*, McGraw-Hill, New York, 1953.

10. E. Bodewig, *Matrix Calculus*, North Holland, Amsterdam, Interscience, New York, 1956.

11. M. R. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Research Natl. Bur. Standards*, **49**, [6], 409-436 (1952).

12. G. E. Forsythe, *Tentative Classification of Methods and Bibliography on Solving Systems of Linear Equations*, Inst. Numerical Anal. Rept. 52-7, Natl. Bur. Standards, Los Angeles, Calif., multilithed typescript, 1951 (Applied Math. Ser. 29, pp. 1-28, Natl. Bur. Standards, Washington, D. C., 1953).

13. G. E. Forsythe, *Solving Linear Algebraic Equations Can Be Interesting*, Inst. Numerical Anal. Rept. 2076, Natl. Bur. Standards, Los Angeles, Calif., 1952.

14. C. Lanczos, *Applied Analysis*, Chaps. II and III, Prentice-Hall, Englewood Cliffs, N. J., 1956.

#### Section 3. Eigenvalues and Eigenvectors

15. O. Taussky, *Bibliography on Bounds for Characteristic Roots of Finite Matrices*, Natl. Bur. Standards Rept. 1162, Washington, D. C., 1951.

See also Refs. 8-10, and 12-14 above.

#### Section 4. Digital Techniques in Statistical Analysis of Experiments

16. J. R. Bainbridge, A. M. Grant, and U. Radok, Tabular analysis of factorial experiments and the use of punch cards, *J. Am. Stat. Assoc.*, **51**, 149-158 (1956).

17. R. C. Bose, W. H. Clatworthy, and S. S. Shrikhande, Tables of partially balanced designs with two associate classes, *North Carolina Agr. Expt. Sta. Tech. Bull.* 107, Raleigh, N. C., 1954.

18. W. G. Cochran and G. M. Cox, *Experimental Designs*, Wiley, New York, 1950.

19. O. L. Davies, Editor, *The Design and Analysis of Industrial Experiments*, Oliver and Boyd, London, 1954.

20. D. B. DeLury, *Values and Integrals of the Orthogonal Polynomials*, Ontario Research Foundation, University of Toronto Press, Toronto, Canada, 1950.

21. R. A. Fisher and F. Yates, *Statistical Tables*, 4th edition, Hafner, New York, 1953.

22. H. O. Hartley, Programming analysis of variance for general purpose computers, *Biometrics*, **12**, 110-122 (1956).

23. O. Kempthorne, *The Design and Analysis of Experiments*, Wiley, New York, 1952.

#### Section 5. Ordinary Differential Equations

24. W. E. Milne, *Numerical Solution of Differential Equations*, p. 65, Wiley, New York, 1953.

25. F. Bashforth and J. C. Adams, *An Attempt to Test the Theories of Capillary Action*, etc., pp. 15-16, Cambridge University Press, Cambridge, England, 1883.

26. C. Runge, Über die numerische Auflösung von Differentialgleichungen, *Math. Ann.*, **46**, 167-178 (1895).

27. W. Kutta, Beitrag zur näherungsweise Integration totaler Differentialgleichungen, *Z. Math. Phys.*, **46**, 435-453 (1901).

28. S. Gill, A process for the step-by-step integration of differential equations in an

automatic digital computing machine, *Proc. Roy. Soc. (London)*, **A193**, 407-433 (1948).

29. B. Dimsdale and R. F. Clippinger, unpublished notes of course presented in 1948 at Computing Laboratory, Ballistic Research Laboratories, Aberdeen Proving Ground, Aberdeen, Md.

30. L. F. Richardson and J. A. Gaunt, The deferred approach to the limit, *Trans. Roy. Soc. (London)*, **A226**, 299 (1927).

#### Section 6. Partial Differential Equations

31. P. D. Lax, Approximations to solutions of differential equations, an operator theoretic approach, *1955 Berkeley Symposium on Partial Differential Equations*, Interscience, New York, 1957.

32. S. P. Frankel, Convergence of iterative treatments of partial differential equations, in *Mathematical Tables and Other Aids to Computation*, Vol. 4, pp. 65-75, Natl. Research Council, Washington, D. C., 1950.

33. D. Young, Iterative methods for solving partial differential equations of elliptic type, *Trans. Am. Math. Soc.*, **76**, 92-111 (1954).

34. J. D. Riley, Iteration procedures for the Dirichlet difference problem, in *Mathematical Tables and Other Aids to Computation*, Vol. 8, pp. 124-131, Natl. Research Council, Washington, D. C., 1954.

35. D. W. Peaceman and H. H. Rachford, Jr., Numerical solution of parabolic and elliptic differential equations, *J. Soc. Ind. and Appl. Math.*, **3**, 28-41 (1955).

36. E. Goursat, *Cours d'Analyse Mathématique*, 5th edition, Vol. II, Chap. XXII, pp. 573-664, Gauthier-Villars, Paris, 1929.

37. R. Courant and D. Hilbert: *Methods of Mathematical Physics*, Vol. I, Interscience, New York, 1953.

38. I. G. Petrovsky, *Lectures on Partial Differential Equations*, Interscience, New York, 1954.

39. R. Courant, K. Friedrichs, and H. Lewy, Über die partiellen Differenzgleichungen der mathematischen Physik, *Math. Ann.*, **100**, 32-74 (1928).

40. G. G. O'Brien, M. A. Hyman, and S. Kaplan, A study of the numerical solution of partial differential equations, *J. Math. Phys.*, **29**, 223-251 (1951).

41. J. Crank and P. Nicholson, A practical method for numerical integration of solutions of partial differential equations of heat-conduction type, *Proc. Cambridge Phil. Soc.*, **43**, 50 (1947).

42. P. Laasonen, Über eine Methode zur Lösung der Wärmeleitungsgleichung, *Acta Math.*, **31**, 309-317 (1949).

43. P. D. Lax and R. D. Richtmyer, Survey of the stability of linear finite difference equations, *Communs. Pure and Appl. Math.*, **9**, 267-293 (1956).

44. F. John, On integration of parabolic equations by difference methods, *Communs. Pure and Appl. Math.*, **5**, 155-211 (1952).

45. J. Douglas, On the numerical integration of quasi-linear parabolic differential equations, *Pacific J. Math.*, **6**, 35-42 (1956).

46. R. Courant, E. Isaacson, and M. Rees, On the solution of nonlinear hyperbolic differential equations by finite differences, *Communs. Pure and Appl. Math.*, **5**, 243-255 (1952).

47. C. Lanczos, Solution of systems of linear equations by minimized iterations, National Bureau of Standards, NAML Rept. 52-13, 1951.

48. P. Brock and F. J. Murray, The use of exponential sums in step by step integration I, II, in *Mathematical Tables and Other Aids to Computation*, Vol. 6, pp. 63-78, 138-150, National Research Council, Washington, D. C., April, July 1952.

# OPERATIONS RESEARCH

## C. OPERATIONS RESEARCH

15. Operations Research, *by E. L. Arnoff*



## Operations Research

*E. Leonard Arnoff*

1. Operations Research and Mathematical Models	15- 02
2. Solution of the Model	15- 10
3. Inventory Models	15- 21
4. Allocation Models	15- 31
5. Waiting Time Models	15- 73
6. Replacement Models	15- 86
7. Competitive Problems	15- 99
8. Data for Model Testing	15-115
9. Controlling the Solution	15-120
10. Implementation	15-123
References	15-124

*Note.* For an extensive discussion of the general nature of O.R., its development, structure, prototype problems, methods, techniques, and tools, the reader is referred to *Introduction to Operations Research* by C. W. Churchman, R. L. Ackoff, and E. L. Arnoff (Ref. 1). The material presented here quite naturally reflects the influence of several years of close association and collaboration with these colleagues and summarizes all too briefly material developed and presented in their textbook. Tables and material included here are reproduced with the permission of the publisher, John Wiley and Sons.

## 1. OPERATIONS RESEARCH AND MATHEMATICAL MODELS

### General Nature of Operations Research

**Development.** Each branch of science has emerged as a result of a concentration on a particular class of problems and on the associated development of scientific methods, techniques, and tools for the solution of these problems. One such class of problems, spotlighted early in the century by such farsighted men as F. W. Taylor, is concerned with the management function. Recent concentration of scientific effort on this set of problems has given rise to the science called "Operations Research," often abbreviated to "O.R." For an extensive discussion of the general nature of O.R., its development, structure, prototype problems, methods, techniques, and tools, see Ref. 1.

Historically, the development of industrial organization has been the growth and transition of industry from one composed of countless small businesses into one characterized by large enterprises. Whereas the small companies were managed by single individuals, these individuals were no longer able to perform effectively all the managerial functions required for large enterprises. Consequently, there resulted a division of the management function so that responsibilities were shared by managers of production, sales and marketing, finance and control, personnel, and so forth.

With the further growth of industry, functional operations, such as production and marketing, were decentralized and resulted in a further division of the management function. For *example*, the production function was now carried on by a combination of subfunctions such as production planning, material control, maintenance, purchasing, stock, and quality control. The production manager was elevated to an executive position somewhat removed from the actual production scene.

This subdivision of the management function focused attention on problems of various subfunctional areas of the industrial organization. Thus, a concentration on production problems by scientists and engineers led to the establishment of industrial engineering, chemical engineering, mechanical engineering, electrical engineering, and statistical quality control. Sciences such as industrial psychology and industrial sociology arose as a result of a concentration of interest on the solution of personnel problems, etc.

However, with this increased concentration of attention to subfunctional problems (e.g., those of production or marketing), and the corresponding emergence of sciences devoted to the solution of these problems, a new type of problem arose—that concerned with the solution of industrial problems

in terms of the overall objectives of the organization rather than in terms of the objectives of each individual subfunctional unit.

For *example*, in production scheduling and inventory control the production department wishes to minimize the cost of production and strives to achieve this by means of long production runs of a relatively few types of items. This leads to a large inventory (in order to achieve savings in setup costs) spread over just a few types of items.

The sales department, on the other hand, is usually interested in sales volume and in being able to fill "every" customer order. They, too, would like to have a large inventory, but one which is composed of many types of items. Furthermore, they are usually in favor of a highly flexible production shop, one able to produce quickly to customer order rather than one committed to long production runs.

The finance department is interested in the production scheduling and inventory control problem in terms of the capital investments and, consequently, is all too apt to seek *low* inventory levels.

The personnel department is also concerned with the production scheduling and inventory control problem. It would like to use inventories as a means of stabilizing labor requirements, thereby reducing costs associated with hiring and layoff, fringe benefits, retraining, employee malcontent, etc.

*The Executive Problem.* The objectives of each of these departments are not mutually consistent but, rather, are in conflict with one another. Therefore, the policy established by any one of these departments will seldom be favorable to the others. Consequently, there exists an *executive problem*—an *overall* problem of mediating among these conflicts of interest and selecting that course of action which is best for the organization as a whole. *It is to the solution of this type of overall problem that the methods, techniques, and tools of O.R. are directed.* The executive problem arises anywhere that there is a need for such a mediation of conflicts of interest in terms of the effectiveness of the overall system. Hence O.R. is not restricted to the solution of problems at the level of, say, the corporation president but rather at any functional or subfunctional level where such a conflict of interests exists.

For many years there have been instances of the application of science to the solution of executive problems. However, the most significant stimulus was provided as a result of the military emergencies of World War II, especially in the Battle of Britain. Here, in a sense somewhat of desperation, scientists were called upon to solve "executive type" problems of strategy and tactics for the most effective utilization of very limited resources (Ref. 2). Interdisciplinary teams of scientists and engineers were formed to bring to bear the widest range of scientific knowledge and methodology to the solution of these problems. These teams, engaged in what

became known as *Operational Research* in Great Britain, were so successful that O.R. units were soon established for military services in the United States.

**Definition of Operations Research.** One working definition that has been stated is: "*Operations Research* is the application of scientific methods, techniques, and tools to problems involving the operations of a system so as to provide those in control of the system with optimum solutions to the problems" (Ref. 1, p. 18).

**Methodology of Operations Research.** By the use of teams (members drawn from different disciplines), a variety of scientific methods, techniques, and tools is made available. O.R. personnel have developed a methodological approach to the solution of system problems consisting of the following steps:

1. Formulating the problem.
2. Constructing a mathematical model to represent the system under study.
3. Deriving a solution from the model.
4. Testing the model and the solution derived from it.
5. Establishing controls over the solution.
6. Implementing the solution.

O.R. establishes a model to represent the system under study in order to be able to perform an abstract manipulation of the system (i.e., via the model) and determines how the system can best be changed.

**Prototype Problems of Operations Research.** In applying O.R. methods, techniques, and tools toward the solution of problems of the industrial system, several classes of problems have arisen with great frequency. These recurrent problems gave rise to the development of special techniques for their solution. The principal classes of recurrent problems are: (a) inventory, (b) allocation, (c) waiting time or queuing, (d) replacement, (e) competition.

For general articles on O.R. prior to 1954, see the bibliography of Ref. 2. See also Refs. 3 and 4.

### Phases of Operations Research

**Formulation of the Problem.** This is usually a sequential process. An initial formulation is completed and research proceeds, but, in proceeding, the problem is subjected to almost continuous and progressive reformulation and refinement. This continues until a solution is reached. In a sense, one never knows until the end of the research whether the problem was correctly formulated, and perhaps not even then.

Anxiety to get the research under way frequently leads to reduction of the time and effort devoted to formulating the problem. This is likely to

be very costly. Consequently, some systematic way of formulating the problem should be a standard procedure of an O.R. team, and a specific allocation of time for formulating the problem should be made.

*The procedure for formulating a problem* may be summarized in the following outline:

- A. Analyze the relevant operations and the communication system by which they are controlled.
  1. Identify and trace each communication related to these operations.
  2. Identify each transformation of information and decision process.
  3. Identify each step in the operations.
  4. Drop from consideration each communication or transformation which has no effect on operations (e.g., billing in production operations).
  5. Group operations between control points.
  6. Prepare a flow chart showing: (a) control points and decisions made, (b) flow of pertinent information between control points and time consumed, and (c) flow of materials and time of grouped operations.
- B. Formulate management's problem.
  1. Identify the decision-makers and the decision-making procedure.
  2. Determine the decision-makers' relevant objectives.
  3. Identify other participants and the channels of their influence on a solution.
  4. Determine the objectives of the other participants.
  5. Determine the alternative courses of action available to the decision-makers.
  6. Determine the counteractions available to other participants.
- C. Formulate the research problem.
  1. Edit and condense the relevant objectives.
  2. Edit and condense the relevant courses of action.
  3. Define the measure of effectiveness to be used: (a) define the measure of efficiency to be used relative to each objective, (b) weight objectives (if qualitative) or units of objectives (if quantitative), and (c) define the criterion of best decision as some function of the sum of weighted efficiencies (e.g., maximum expected return, minimum expected loss).

**The Components of a Problem.** These are: (1) the decision-maker, (2) objectives, (3) the system, and (4) alternative courses of action.

**The Decision-Maker.** The following questions may serve to direct study of the decision-making process in the problem area:

1. Who has the responsibility for making recommendations concerning modification of policies?

2. Whose approval is required and how is this approval expressed?
3. What constitutes final approval? (A majority vote in group deliberation, approval by a final authority in a sequence of reviews, etc.)
4. Does anyone have absolute veto power? If not, how can a recommendation be rejected?
5. Who has the responsibility for carrying out recommendations once they are approved?
6. Who has responsibility for evaluating the action taken?

**The Decision-Maker's Objectives.** Direct questioning of decision-makers seldom reveals all their pertinent objectives relative to a problem. Such questioning provides a start, but it seldom provides enough information for a complete formulation of objectives.

For *example*, objectives to be obtained may be: (a) to decrease production costs, (b) to render better customer service, (c) to increase a share of the market. Objectives to be retained may be: (a) to maintain stable employment, (b) to retain product leadership, (c) to preserve good relations with the community.

**The System.** Most organized systems involve the following components: (1) management, which directs (2) men, who control and operate (3) machines, which convert (4) materials into products or services made available to (5) consumers, whose purchases are also sought by (6) competitors, (7) government, and the public.

**Alternative Courses of Action.** A number of possible alternative courses of action are ordinarily disclosed in the process of going through the earlier steps in formulating the problem. The list of alternatives disclosed in this way may not be exhaustive. The researchers should get as complete a list of alternatives as possible, even to the extent of including possibilities that are not thought to be feasible.

The derivation of a list of alternatives comes about by asking and answering the following questions. For each phase of the system, would the efficiency be affected by a change in (1) personnel, (2) operations, (3) the materials and/or machines, or (4) environment? Whenever an affirmative answer is obtained, the specific alternatives at this stage can be explored.

**The Research Problem.** Transformation of the decision-maker's problem into a research problem involves the following steps: (1) editing the list of objectives obtained in the first stage of problem formulation, (2) editing the list of alternative courses of action, (3) defining the measure of effectiveness to be used.

Before proceeding to a detailed discussion of each of these three steps, some remarks concerning the logic of decision-making (i.e., decision theory) are in order.

**EXAMPLE.** Consider the following simplified abstract problem. Only two objectives are involved,  $O_1$  and  $O_2$ ; and only two courses of action are possible,  $C_1$  and  $C_2$ . Now suppose the efficiency of each course of action for each objective has been determined (on a scale from 0 to 1) and the results are shown in the following matrix:

	$O_1$	$O_2$
$C_1$	0.8	0.4
$C_2$	0.2	0.6

Which course of action should be selected? It is a mistake to answer either " $C_1$ " or " $C_2$ ." The question cannot be answered without information concerning the relative importance of the objectives. If  $O_1$  is much more important than  $O_2$ , it seems clear that  $C_1$  should be selected, but if  $O_2$  is much more important than  $O_1$ ,  $C_2$  should be selected. How can the criterion of selection be made explicit? If the relative importance of  $O_1$  and  $O_2$  could be measured, such a criterion could be provided. Suppose that *relative importance* could be measured along a scale running from 0 to 1, and is 0.3 for  $O_1$ , and is 0.7 for  $O_2$ . Now the efficiency of each course of action for each objective can be weighted as follows:

	$O_1$	$O_2$	Total
$C_1$	$0.3 \times 0.8 = 0.24$	$0.7 \times 0.4 = 0.28$	0.52
$C_2$	$0.3 \times 0.2 = 0.06$	$0.7 \times 0.6 = 0.42$	0.48

The sum of the weighted efficiencies (efficiency times relative importance) of a course of action is its *relative effectiveness*.

**Editing the Objectives and Courses of Action.** The purpose of editing the objectives is to simplify and condense the list obtained in the first stage of formulating the problem. The editing procedure can be considered in three steps.

1. An examination of the list to determine whether the attainment of any one objective is important only because it is a means to the attainment of another objective or objectives. If so, it may be eliminated. For *example*, suppose one of the objectives is "to increase the company's annual net profit," and another is to "decrease production costs." It is likely that there is interest in decreased production costs only to the extent that it leads to increased net profits. If so, "decreased production costs" can be considered as a "means," not as an objective.

2. Examination of each objective relative to the alternative courses of action to determine whether attainment of any of the objectives would be

unaffected by a choice from among the alternatives. If an objective is so unaffected, it should be eliminated from the list. For *example*, suppose one objective listed is "to maintain a high quality product" and the alternative courses of action involve only determination of production lot sizes. Then, if quality is unaffected by lot size, the "quality maintenance" objective can be dropped from the list.

3. *Combine objectives* of different participants that are essentially the same. For *example*, both employer and employee may be interested in stable employment, and both manufacturer and consumer may be interested in low price and high quality.

*Editing the Courses of Action.* The list of possible *alternative actions* available to the decision-maker should be examined to determine whether there are reasons for eliminating any of these from further consideration.

**Defining the Measure of Effectiveness.** (See Ref. 1, Chap. 5.) Measures of effectiveness consist of two component measures: (1) the importance of the objectives and (2) the efficiency of the courses of action.

The procedure for establishing an overall measure of effectiveness is a very complex one. However, for *quantifiable objectives*, this procedure may be summarized as follows:

1. Develop a measure of efficiency relative to each objective.
2. Where the measures of efficiency obtained in step 1 differ, develop a way of transforming the measures into one common or standard measure of efficiency.
3. For each course of action and each objective, determine the probability of attaining each possible level of efficiency. This is the *efficiency function* of each course of action for each objective.
4. For each course of action, "add" the efficiency functions so as to obtain a combined efficiency relative to all objectives. The result is an *effectiveness function* for each course of action relative to the entire set of objectives.
5. State the objective of the decision process in terms of maximizing or minimizing expected return, gain, or loss.
6. Construct a "return function" for each course of action. A return function expresses the expected outcome (outcome times its probability of occurrence) in terms of gains and losses.

For establishing measures of effectiveness for mixed quantitative and qualitative objectives or for purely qualitative objectives, see Ref. 1, Chap. 5.

**Constructing a Model** (Ref. 1, Chap. 7). A mathematical model expresses the effectiveness of the system under study as a function of a set of variables, at least one of which is subject to control.

The general form of an O.R. model can be expressed as

$$E = f(x_i, y_j),$$

where  $E$  represents the *effectiveness of the system*,  $x_i$  those variables of the system which are subject to control, and  $y_j$  those variables which are not subject to control. Restrictions, if any, on values of the variables may be expressed in a supplementary set of equations and/or inequations. *The O.R. problem is to determine those values of the control variables,  $x_i$ , which best achieve the given measure of effectiveness,  $E$ .*

The procedure for constructing a mathematical model may be summarized as follows:

1. Itemize those components of the system which affect its effectiveness.
2. Edit this list of components (combining and eliminating factors).
3. Assign a symbol to each factor retained.
4. Construct one (or more) equations which express the effectiveness of the system as a function of the pertinent variables.
5. Identify the *control variables*, i.e., variables which can be manipulated by management decision.

**Deriving a Solution from the Model** (see Sect. 2). There are essentially two types of procedures for deriving an optimum solution from a model, *analytic* and *numerical*.

*Analytic procedures* consist of the use of mathematical deduction. Through the application of various branches of mathematics such as calculus and matrix algebra, analytic solutions are obtained in the abstract; i.e., the substitution of numbers for symbols is generally made after the solution has been obtained.

*Numerical procedures* consist of trying various values of the control variables in the model, comparing the results obtained, and selecting that set of values of the control variables which yield the best solution. Such procedures vary from pure trial-and-error to iterative procedures in which successive intermediate trials converge to an optimum solution.

In many instances, especially in probabilistic models, one may not be able to evaluate certain terms or expressions in the model because of mathematical or practical considerations. In many such cases, a particular application of random sampling, called the *Monte Carlo technique*, can be used to obtain accurate evaluations of the expressions.

**Testing the Model and the Solution** (see Sect. 8). A model is never more than a partial representation of the system under study. It is a good model if, despite its incompleteness, it can accurately predict the effect of changes in the system on the overall effectiveness of the system.

The solution can be evaluated by comparing results obtained without applying the solution with results obtained when it is used. These evalua-

tions may be performed retrospectively by the use of past data or by a trial run or pretest. Testing requires careful analysis as to the validity of the data.

**Establishing Controls over the Solution** (see Sect. 9). A solution derived from a model remains a solution only as long as the uncontrolled variables retain their values and the relationship between the variables in the model remains constant. The solution itself goes "out of control" when the value of one or more of the uncontrolled variables and/or one or more of the relationships between variables have changed significantly. The significance of the change depends on (1) the amount by which the solution is made to deviate from the true optimum under the changed conditions and (2) the cost of changing the solution in operation. To establish controls over the solution, then, one must develop tools for determining when significant changes occur and rules must be established for modifying the solution so as to take these changes into account.

**Putting the Solution to Work** (see Sect. 10). The tested solution must be translated into a set of operating procedures capable of being understood and applied by the personnel who will be responsible for their use. Required changes in existing procedures and reserves must be specified and carried out.

## 2. SOLUTION OF THE MODEL

**Types of Models.** A *scientific model* is a representation of some subject of inquiry (such as objects, events, processes, systems) and is used for purposes of *prediction* and *control*. The primary function of a scientific model is explanatory rather than descriptive. It is intended to make possible or to facilitate determination of how changes in one or more aspects of the modeled entity may affect other aspects, or the whole. In the employment of models, this determination is made by manipulating the model rather than by imposing changes on the modeled entity itself.

Models may be classified into three types (see Ref. 1, Chap. 7):

1. *Iconic* models pictorially or visually represent certain aspects of a system. *Examples:* a photograph, a scaled model plant layout.
2. *Analog* models represent one set of (system) properties by another set of properties. *Examples:* graphs, analog computers.
3. *Symbolic* models, by means of mathematical equations, employ symbols to designate properties of the system. *Examples:* the force of gravity,  $F = mg$ .

*Note.* This section will deal only with symbolic models, since most problems in O.R. are solved by means of such models.

The model is an instrument which helps one evaluate alternative policies efficiently. The selection of a procedure for deriving a solution to the

problem from the model depends upon the characteristics of the model. These procedures can be classified into two types, analytic and numerical.

*Analytic procedures* are essentially deductive in character, whereas *numerical procedures* are essentially inductive. In some cases (especially probabilistic situations), neither of these procedures can be applied until one or more terms in the equation have been evaluated by what is called the Monte Carlo technique.

### Analytic Solutions

Symbolic models can assume a variety of mathematical forms and consequently require many different types of mathematical analyses for solution. The primary means of obtaining analytic solutions is through the use of calculus (see standard college textbooks). The methods of calculus are especially applicable to the solution of inventory problems (see Sect. 3).

**One Control Variable, No Restrictions.** To illustrate the application of the methods of calculus, consider the elementary inventory situation described in Sect. 3 and represented by the model given there in eq. (1). See Sect. 3 for definition of symbols.

The total expected relevant cost per year,  $TEC$ , is given by

$$(1) \quad TEC = \frac{1}{2}C_1Tq + \frac{1}{q}C_sR,$$

where  $q$  is the controllable variable which designates the number of items to be produced per run. The problem is to determine the most economic lot size,  $q^*$ . Take the derivative of  $TEC$ , eq. (1), with respect to  $q$ :

$$(2) \quad \frac{d(TEC)}{dq} = \frac{1}{2}C_1T - \frac{C_sR}{q^2}.$$

Set eq. (2) equal to zero and solve for  $q$ :

$$(3) \quad \frac{1}{2}C_1T - \frac{C_sR}{q^2} = 0.$$

Then

$$(4) \quad q^* = \sqrt{2\frac{RC_s}{TC_1}}$$

represents the most economic number of items per run, i.e., that value of  $q$  which minimizes  $TEC$ .

**Two or More Control Variables, No Restrictions.** When the model contains more than one control variable, one must take *partial* derivatives. Thus, if one wishes to determine the economic lot size for *two* independent

items by means of one equation, the total estimated relevant cost per year becomes

$$(5) \quad TEC = \left( \frac{1}{2}C_{11}Tq_1 + \frac{1}{q_1}C_{s1}R_1 \right) + \left( \frac{1}{2}C_{12}Tq_2 + \frac{1}{q_2}C_{s2}R_2 \right),$$

where  $q_1$  and  $q_2$  are the respective control variables.

By taking the *partial* derivatives of  $TEC$  with respect to  $q_1$  and  $q_2$ ,

$$(6) \quad \frac{\partial(TEC)}{\partial q_1} = \frac{1}{2}C_{11}T - C_{s1}\frac{R_1}{q_1^2} \quad \text{and} \quad \frac{\partial(TEC)}{\partial q_2} = \frac{1}{2}C_{12}T - C_{s2}\frac{R_2}{q_2^2},$$

which, when equated to zero, yield the most economic lot sizes:

$$(7) \quad q_1^* = \sqrt{2\frac{R_1C_{s1}}{TC_{11}}} \quad \text{and} \quad q_2^* = \sqrt{2\frac{R_2C_{s2}}{TC_{12}}}.$$

More generally, for any number of control variables,

$$(8) \quad q_i^* = \sqrt{2\frac{R_iC_{si}}{TC_{1i}}}.$$

**Two or More Control Variables with Restrictions.** When one has an objective function which is to be maximized or minimized subject to a set of restrictions, a number of methods or techniques for solution may be applicable, but only one is discussed here.

**Lagrangian Multipliers.** The method of Lagrangian multipliers enables one to solve the following problem.

PROBLEM. Determine values of  $x_1, x_2, \dots, x_n$  which maximize

$$(9) \quad u = f(x_1, x_2, \dots, x_n)$$

subject to

$$(10) \quad \begin{aligned} g_1(x_1, x_2, \dots, x_n) &= 0 \\ g_2(x_1, x_2, \dots, x_n) &= 0 \\ \dots & \\ g_m(x_1, x_2, \dots, x_n) &= 0 \end{aligned}$$

By the method of Lagrange, a new function is formed,

$$(11) \quad \phi = f(x_1, x_2, \dots, x_n) + \lambda_1g_1(x_1, x_2, \dots, x_n) + \dots + \lambda_mg_m(x_1, x_2, \dots, x_n),$$

where  $\lambda_1, \lambda_2, \dots, \lambda_m$  are undetermined multipliers called *Lagrangian multipliers*.

Then, in order to determine the extremal values of  $u = f(x_1, x_2, \dots, x_n)$ , all that is necessary is to obtain the solution of the system of eqs. (10) for the unknowns  $x_1, x_2, \dots, x_n, \lambda_1, \lambda_2, \dots, \lambda_m$ . (See Ref. 5.)

**EXAMPLE.** Find the point in the plane  $x + 2y + 3z = 14$  nearest the origin. The problem may be converted to that of finding values of  $x, y$ , and  $z$  which minimize the square of the sphere diameter

$$u = D^2 = x^2 + y^2 + z^2$$

subject to

$$g = x + 2y + 3z - 14 = 0.$$

Form:

$$\phi = x^2 + y^2 + z^2 - \lambda(x + 2y + 3z - 14).$$

Take partial derivatives of  $\phi$  with respect to  $x, y, z$ , and  $\lambda$ .

$$\frac{\partial \phi}{\partial x} = 2x - \lambda, \quad \frac{\partial \phi}{\partial y} = 2y - 2\lambda,$$

$$\frac{\partial \phi}{\partial z} = 2z - 3\lambda, \quad \frac{\partial \phi}{\partial \lambda} = x + 2y + 3z - 14.$$

Setting these four partial derivatives equal to zero and solving the system of four simultaneous equations then yields:

$$x = 1, \quad y = 2, \quad z = 3, \quad \text{and} \quad \lambda = -2;$$

that is, the point in the plane  $x + 2y + 3z = 14$  nearest the origin is (1, 2, 3).

Other examples of the use of Lagrangian multipliers can be found in Ref. 6 and in texts on advanced calculus.

**Modified Lagrangian Multiplier Method** (Ref. 1, Chap. 10). Many practical extremal problems have the added restriction that all variables must be non-negative; for example, it makes no sense to produce  $-n$  units of a given product. Furthermore, the restrictions may be given in the form of *inequalities* instead of equalities. Since the Lagrangian multiplier method does not guarantee the non-negativity of the solution variables, a modification must be made.

**EXAMPLE.** Consider an economic lot size inventory problem (of the type described above) involving two products, with a restriction on the total available warehouse space. If  $W_1$  and  $W_2$  are the respective unit storage requirements, and an average inventory level is assumed equal to one-half the lot size  $q$ , the total space requirement can be written as

$$(12) \quad \frac{1}{2} \sum W_i q_i = \frac{1}{2} W_1 q_1 + \frac{1}{2} W_2 q_2 \leq S.$$

If  $W_1 = 5$  cu ft,  $W_2 = 35$  cu ft, and  $S = 14,000$  cu ft, eq. (12) becomes

$$(13) \quad \frac{5}{2}q_1 + \frac{35}{2}q_2 \leq 14,000,$$

or, equivalently,

$$(14) \quad 5q_1 + 35q_2 \leq 28,000.$$

The problem (for two products) can then be stated as:

*Problem.* Determine *non-negative* values of  $q_1$  and  $q_2$  which minimize

$$TEC = \left( \frac{1}{2}C_{11}Tq_1 + \frac{1}{q_1}C_{s1}R_1 \right) + \left( \frac{1}{2}C_{12}Tq_2 + \frac{1}{q_2}C_{s2}R_2 \right)$$

subject to the restriction of eq. (14).

*Solution.* Define an undetermined multiplier  $\lambda$  such that

$$(15) \quad \begin{aligned} \lambda < 0 & \text{ when } S - \frac{1}{2}\Sigma W_i q_i = 0; \\ \lambda = 0 & \text{ when } S - \frac{1}{2}\Sigma W_i q_i > 0. \end{aligned}$$

Form

$$(16) \quad \phi = TEC + \lambda(S - \frac{1}{2}\Sigma W_i q_i),$$

that is,

$$(17) \quad \phi = \left( \frac{1}{2}C_{11}Tq_1 + \frac{1}{q_1}C_{s1}R_1 \right) + \left( \frac{1}{2}C_{12}Tq_2 + \frac{1}{q_2}C_{s2}R_2 \right) + \lambda(S - \frac{1}{2}W_1q_1 - \frac{1}{2}W_2q_2).$$

Since  $\lambda(S - \frac{1}{2}W_i q_i)$  is *always* identically zero by eq. (15),  $\phi = TEC$ . Taking partial derivatives of  $\phi$  with respect to  $q_1$  and  $q_2$  yields

$$(18a) \quad \frac{\partial(TEC)}{\partial q_1} = \frac{1}{2}C_{11}T - \frac{1}{q_1^2}C_{s1}R_1 - \frac{1}{2}\lambda W_1,$$

and

$$(18b) \quad \frac{\partial(TEC)}{\partial q_2} = \frac{1}{2}C_{12}T - \frac{1}{q_2^2}C_{s2}R_2 - \frac{1}{2}\lambda W_2.$$

Setting eqs. (18) equal to zero yields

$$(19a) \quad q_1^* = \sqrt{\frac{2C_{s1}R_1}{C_{11}T - \lambda W_1}},$$

and

$$(19b) \quad q_2^* = \sqrt{\frac{2C_{s2}R_2}{C_{12}T - \lambda W_2}}.$$

For each product, the quantities  $R_i$ ,  $C_{si}$ ,  $C_{1i}$ ,  $W_i$ , and  $T$  are known, but  $\lambda$  is still unknown. However, for any arbitrarily assigned value of  $\lambda$ ,  $q_i$  and, hence,  $\frac{1}{2}\sum W_i q_i$  can be calculated. If  $\frac{1}{2}\sum W_i q_i$  exceeds  $S$  (see eqs. 12 and 13), the lot sizes are too large. In this case, decrease  $\lambda$  repeatedly and recompute until  $\frac{1}{2}\sum W_i q_i = S$  has been obtained. If  $\frac{1}{2}\sum W_i q_i < S$  for all negative  $\lambda$ , set  $\lambda = 0$  in eq. (19). The resulting  $q_i$ 's will allow the smallest possible total costs for the company with existing warehouse space  $S$ .

TABLE 1. STORAGE SET BY VARIOUS  $\lambda$  VALUES<sup>a</sup>

$\lambda$	$q_1^*$	$q_2^*$	$\frac{1}{2}(5q_1 + 35q_2)$
-0.0000	816	756	15,270
-0.0012	813	721	14,650
-0.0024	810	690	14,100
-0.0036	806	663	13,618
-0.0060	800	617	12,790
-0.0084	794	580	12,135
-0.0120	784	535	11,323

<sup>a</sup> Assumes:  $T = 12$  months and

Product	$R_i$	$C_{si}$	$C_{1i}$
$X_1$	2400	\$100	0.060
$X_2$	4800	\$ 25	0.035

Values of  $\frac{1}{2}(5q_1 + 35q_2)$  are calculated in Table 1 in order to determine the correct value of  $\lambda$ . As indicated in Table 1,  $\lambda$  should be approximately equal to  $-0.0024$  so that

$$q_1^* = 810 \quad \text{and} \quad q_2^* = 690.$$

*Note.* For this example, without any restriction on storage space (minimizing *TEC*, rather than  $\phi$ ),

$$q_1^* = 816 \quad \text{and} \quad q_2^* = 756.$$

Another approach to a modified Lagrangian multiplier technique can be found in Ref. 7. Such modified Lagrangian multiplier techniques, when applicable, are most cumbersome and impractical for a large number of variables. (See Ref. 1, Chap. 10.) Where the objective function and the restrictions are linear (and for some other special cases), the techniques of linear programming are applicable (see Sect. 4).

**Other Analytic Methods of Solution.** There are many other analytic methods of solution much more sophisticated than those presented here. A number of the models arising in specific problems require the development of special methods for their solution. For these more sophisticated

and special methods, see journals such as *Operations Research*, *Management Science*, *Econometrica*, *Naval Research Logistics Quarterly*, and the publications of the RAND Corporation (Santa Monica, Calif.). See also Refs. 7, 8, and 9.

### Numerical Solutions

Numerical techniques of deriving a solution from a model consist of substituting numbers for the symbols in the model and finding that set of substituted numbers which yields the maximum effectiveness. Some numerical procedures are trial-and-error procedures into which one seeks to build some rationale for the selection of subsequent trials. Others are so-called iterative procedures in which one converges to an optimum solution through successively better steps.

**Newton's Method.** An example of a quite useful trial-and-error procedure is Newton's method for solving equations, which is a procedure for determining, within any desired degree of accuracy, the roots of an algebraic equation. The method is based on the fact that, for a short distance, the tangent to a smooth curve is a good approximation to the curve.

Newton's method may be formulated as follows. Let  $f(X) = 0$  be the equation under consideration. A root of this equation is the abscissa of a point at which the curve  $Y = f(X)$  crosses the  $X$ -axis.

Start with a trial solution, say  $X_0$  (see Fig. 1). This value  $X_0$  determines a point  $P$  on the curve whose coordinates are  $(X_0, Y_0)$ . The tangent to the

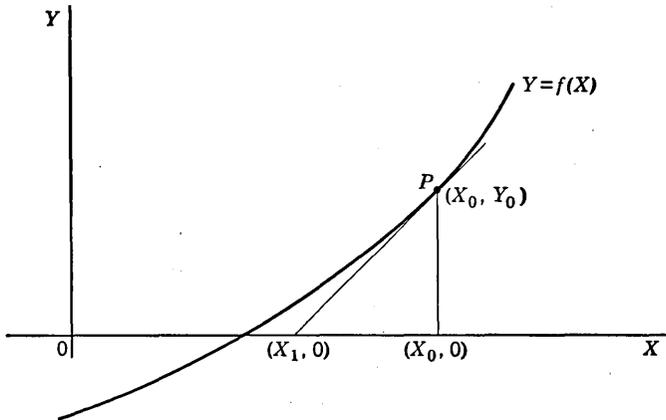


FIG. 1. Figure for Newton's method (Ref. 1).

curve at  $P$  is then drawn and will intersect the  $X$ -axis at  $(X_1, 0)$ . If the curve and the tangent are nearly coincident over the range  $(X_0, X_1)$ , the value  $X_1$  will be the first approximate root of the equation. Furthermore,

using the fact that the slope of the tangent at  $P$  is given by  $f'(X_0)$ , namely the derivative of  $f(X)$  evaluated at  $X = X_0$ , yields

$$(20) \quad X_1 = X_0 - \frac{f(X_0)}{f'(X_0)}.$$

The procedure may be repeated as many times as necessary where, in general,

$$(21) \quad X_{n+1} = X_n - \frac{f(X_n)}{f'(X_n)}.$$

Whether and how fast the process will converge depends on the function  $f(X)$  and the initial value  $X_0$ . Conditions favorable to convergence are evidently that  $f(X_0)$  be small and  $f'(X_0)$  be large.

To illustrate Newton's method, consider

$$f(X) = X^3 - 3X^2 + 4X - 2.$$

Although there are many devices which can be used to locate integers between or at which roots will lie, arbitrarily take  $X_0 = 2$  as the trial solution. For the particular  $f(X)$  chosen,  $X = 1$  is obviously a solution, and that is what we wish to approximate by Newton's method. The deviation from the value  $X = 1$  will, of course, measure the degree of accuracy of this approximation.

Now

$$f'(X) = 3X^2 - 6X + 4,$$

so that

$$f(2) = 8 - 12 + 8 - 2 = 2,$$

$$f'(2) = 12 - 12 + 4 = 4.$$

Hence, using eq. (21) yields

$$X_1 = 2 - \frac{2}{4} = 1.5.$$

By continuing in this manner,

$$f(1.5) = \left(\frac{3}{2}\right)^3 - 3\left(\frac{3}{2}\right)^2 + 4\left(\frac{3}{2}\right) - 2 = \frac{5}{8},$$

and

$$f'(1.5) = 3\left(\frac{3}{2}\right)^2 - 6\left(\frac{3}{2}\right) + 4 = \frac{7}{4},$$

so that

$$X_2 = 1.5 - \frac{5}{14} = 1.143.$$

Continuing once more gives

$$f(1.143) = 0.147,$$

$$f'(1.143) = 1.060,$$

so that

$$X_3 = 1.143 - \frac{0.147}{1.060} = 1.004.$$

One could continue in this manner, measuring at each stage of the iterative procedure the value of  $f(X_i)$  to indicate how quickly one is converging to a solution [obviously, at a point of solution  $X^*$ ,  $f(X^*) = 0$ ], and, hence, obtain this solution within any prescribed degree of accuracy.

Excellent examples of converging iterative procedures are to be found in the several techniques of linear programming. These are discussed in Sect. 4.

### The Monte Carlo Technique

In many mathematical models, it is necessary to evaluate certain terms in the model before a solution can be derived. Especially where probability concepts are involved, it may not be possible or practical to evaluate a given function (within a model) by mathematical analysis. Such expressions, however, can be evaluated by the Monte Carlo technique. Specifically, the Monte Carlo technique is a procedure by which one can obtain approximate evaluations of mathematical expressions which are built up of one or more probability distribution functions.

The Monte Carlo technique consists of simulating an experiment to determine some probabilistic property of a population of objects or events by the use of random sampling applied to the components of the objects or events. This statement can best be clarified by means of examples.

**The Random Walk Problem.** The discovery of the Monte Carlo technique is said to be due to a legendary mathematician observing the perambulation of a saturated drunk. The mathematician wondered how many steps the drunkard would have to take, *on the average*, to reach a specified distance from his starting point, if it were assumed that, at each step, there was an equal probability of the drunkard stepping off in any direction.

**EXAMPLE.** To illustrate how the Monte Carlo technique can be applied to this problem of the "random walk," an estimate can be obtained of the probable distance traveled after five steps of equal size. (It is further assumed, for simplicity of presentation, that these steps are at  $45^\circ$ ,  $135^\circ$ ,  $225^\circ$ , or  $315^\circ$ .) To do this, refer to Table 2, which is a portion of a two-digit random number table.

TABLE 2. RANDOM NUMBERS (Ref. 1)

09 73 25 33	76 53 01 35 86	34 67 35 48 76	80 95 90 90 17	39 29 27 49
54 20 48 05	64 89 47 42 96	24 80 52 40 37	20 63 61 04 02	00 82 29 16
42 26 89 53	19 64 50 93 03	23 20 90 25 60	15 95 33 47 64	35 08 03 36
01 90 25 29	09 37 67 07 15	38 31 13 11 65	88 67 67 43 97	04 43 62 76
80 79 99 70	80 15 73 61 47	64 03 23 66 53	98 95 11 68 77	12 17 17 68
06 57 47 17	34 07 27 68 50	36 69 73 61 70	65 81 33 98 85	11 19 92 91
06 01 08 05	45 57 18 24 06	35 30 34 26 14	86 79 90 74 39	23 40 30 97
26 97 76 02	02 05 16 56 92	68 66 57 48 18	73 05 38 52 47	18 62 38 85
57 33 21 35	05 32 54 70 48	90 55 35 75 48	28 46 82 87 09	82 49 12 56
79 64 57 53	03 52 96 47 78	35 80 83 42 82	60 93 52 03 44	35 27 38 84
52 01 77 67	14 90 56 86 07	22 10 94 05 58	60 97 09 34 33	50 50 07 39
80 50 54 31	39 80 82 77 32	50 72 56 82 48	29 40 52 42 01	52 77 56 78
45 29 96 34	06 28 89 80 83	13 74 67 00 78	18 47 54 06 10	68 71 17 78
68 34 02 00	86 50 75 84 01	36 76 66 79 51	90 36 47 64 93	29 60 91 01
59 46 73 48	87 51 76 49 69	91 82 60 89 28	93 78 56 13 68	23 47 83 41
48 11 76 74	17 46 85 09 50	58 04 77 69 74	73 03 95 71 86	40 21 81 65
12 43 56 35	17 72 70 80 15	45 31 82 23 74	21 11 57 82 53	14 38 55 37
35 09 98 17	77 40 27 72 14	43 23 60 02 10	45 52 16 42 37	96 28 60 26
91 62 68 03	66 25 22 91 48	36 93 68 72 03	76 62 11 39 90	94 40 05 64
89 32 05 05	14 22 56 85 14	46 42 75 67 88	96 29 77 88 22	54 38 21 45
49 91 45 23	68 47 92 76 86	46 16 28 35 54	94 75 08 99 23	37 08 92 00
33 69 45 98	26 94 03 68 58	70 29 73 41 35	53 14 03 33 40	42 05 08 23
10 48 19 49	85 15 74 79 54	32 97 92 65 75	57 60 04 08 81	22 22 20 64
55 07 37 42	11 10 00 20 40	12 86 07 46 97	96 64 48 94 39	28 70 72 58
60 64 93 29	16 50 53 44 84	40 21 95 25 63	43 65 17 70 82	07 20 73 17
19 69 04 46	26 45 74 77 74	51 92 43 37 29	65 39 45 95 93	42 58 26 05
47 44 52 66	95 27 07 99 53	59 36 78 38 48	82 39 61 01 18	33 21 15 94
55 72 85 73	67 89 75 43 87	54 62 24 44 31	91 19 04 25 92	92 92 74 59
48 11 62 13	97 34 40 87 21	16 86 84 87 67	02 07 11 20 59	25 70 14 66
52 37 83 17	73 20 88 98 37	68 93 59 14 16	26 25 22 96 63	05 52 28 25
49 35 24 94	75 24 63 38 24	45 86 25 10 25	61 96 27 93 35	65 33 71 24
54 99 76 54	64 05 18 81 59	96 11 96 38 96	54 69 28 23 91	23 28 72 95
96 31 53 07	26 89 80 93 54	33 35 13 54 62	77 97 45 00 24	90 10 33 93
80 80 83 91	45 42 72 68 42	83 60 94 97 00	13 02 12 48 92	78 56 52 01
05 88 52 36	01 39 09 22 86	77 28 14 40 77	93 91 08 36 47	70 61 74 29
17 90 02 97	87 37 92 52 41	05 56 70 70 07	86 74 31 71 57	85 39 41 18
23 46 14 06	20 11 74 52 04	15 95 66 00 00	18 74 39 24 23	97 11 89 63
56 54 14 30	01 75 87 53 79	40 41 92 15 85	66 67 43 68 06	84 96 28 52
15 51 49 38	19 47 60 72 46	43 66 79 45 43	59 04 79 00 33	20 82 66 85
86 43 19 94	36 16 81 08 51	34 88 88 15 53	01 54 03 54 56	05 01 45 11
08 62 48 26	45 24 02 84 04	44 99 90 88 96	39 09 47 34 07	35 44 13 18
18 51 62 32	41 94 15 09 49	89 43 54 85 81	88 69 54 19 94	37 54 87 30
95 10 04 06	96 38 27 07 74	20 15 12 33 87	25 01 62 52 98	94 62 46 11

Use the following symbolism:

1. The lamppost is represented by the origin of the  $X$ - and  $Y$ -axis. See Fig. 2.

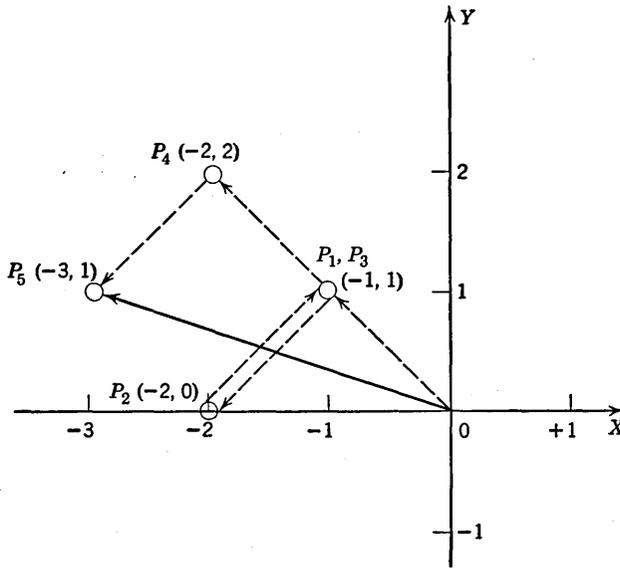


FIG. 2. Plotting of points  $(x_n, y_n)$  (Ref. 1).

2. The first digit of the two-digit random number selected from the table represents *one unit* of  $X$ , *positive if even or zero, negative if odd*.

3. The second digit of the same two-digit random number selected represents *one unit* of  $Y$ , *positive if even or zero, and negative if odd*.

4.  $(x_n, y_n)$  represents the position of the drunkard at the end of the  $n$ th phase.

5.  $d_n$  equals the distance of the drunkard from the lamppost at the end of the  $n$ th phase; that is,  $d_n^2 = x_n^2 + y_n^2$ .

To start at random, select the two-digit number, say in column 10 and row 6 of Table 2, and, by reading down, obtain the following five numbers: 36, 35, 68, 90, and 35. These numbers may then be arranged and the drunkard's moves obtained as shown in Table 3. The points  $(x_n, y_n)$  may also be plotted as in Fig. 2.

TABLE 3

Phase $n$	First Digit	Second Digit	Point Location $(x_n, y_n)$
1	3	6	$(-1, 1)$
2	3	5	$(-2, 0)$
3	6	8	$(-1, 1)$
4	9	0	$(-2, 2)$
5	3	5	$(-3, 1)$

In this example, then, *one* estimate is that the drunkard will be 3.16 units from the lamppost at the end of the fifth phase. This is obtained as follows:

$$d_5^2 = x_5^2 + y_5^2,$$

$$d_5^2 = 9 + 1,$$

$$d_5 = \sqrt{10} = 3.16.$$

This procedure must be repeated for different random numbers in the table so that a group of estimates of the desired distance is obtained. The estimates in this group can then be averaged to yield an average estimated distance from the lamppost. In general, the estimates will improve as the number of such samples is increased. The accuracy of the estimate will be proportional to the square root of the number of samples.

More generally, from many such simulated trials, the probability of the drunkard's being a specified distance from the lamppost for any number  $n$  of irregular zigzag phases is estimated.

As a point of interest and as a basis for the reader comparing his own Monte Carlo solutions, it might be pointed out that, for this example, an analytic solution is obtainable and is given by

$$d_n = a\sqrt{n},$$

i.e., the most probable distance of the drunkard from the lamppost, *after a large number of irregular phases of his walk*, is equal to the average length  $a$  of each straight track he walks, times the square root of the number  $n$  of phases of his walk.

For an illustration of the use of the Monte Carlo technique for the solution of problems involving normal distributions, see Sect. 6.

The use of the Monte Carlo technique for any probability distribution function can be found in Ref. 1, Chap. 7. Reference 1 discusses only the normal distribution, but the treatment is general and applicable to any probability distribution function. For a discussion of the nature of tables of random numbers and a bibliography of tables and works on this subject, see Ref. 10. Examples of other uses of the Monte Carlo technique can be found in Ref. 1, Chaps. 7, 14, and 17. See also Refs. 5, 11-13.

### 3. INVENTORY MODELS

**Problem Statement.** Inventory problems are concerned with minimizing the sum of costs such as those due to (a) carrying inventory, (b) setup, (c) shortage, (d) obsolescence, and (e) change of work force level. Inventory problems require the determination of (a) how many (or much) to order (i.e., produce or purchase) and/or (b) when to order.

This section will introduce the kind of analysis that yields symbolic models of inventory processes. The mathematical models and solutions presented here pertain to specific inventory situations and progress from the most elementary to somewhat complex ones. For a complete definition and classification of the characteristics of inventory problems, see Ref. 1, Chap. 7.

*Decisions.* The general class of inventory problems to be considered involves decisions concerning inventory levels. These decisions can be classified as follows: (1) The *time* at which orders for goods are to be placed is fixed. The *quantity* to be ordered must be determined. (2) Both the order *quantity* and order *time* must be determined.

*Cost.* The costs associated with inventory are of three types: (1) *setup cost*, the fixed cost per lot of obtaining goods (purchasing or manufacturing); (2) *inventory holding cost*, including cost of money spent in obtaining the part, storage, obsolescence, handling, taxes, and insurance; (3) *shortage cost*, cost resulting from a delay in supplying the goods or an inability to fill the order at the time of request.

*Variables.* The three major classes of variables in an inventory problem are: (1) *cost* variables, (2) *demand* variables, i.e., relative to customer demand for goods; (3) *order* variables, i.e., relative to obtaining the necessary goods.

### Elementary Inventory Models (see Ref. 1, Chap. 8)

*Symbols.* The following symbols are used throughout the discussion of the elementary inventory models.

$q$	input, or quantity ordered
$q_i$	input which occurs at the beginning of the $i$ th time interval
$q^*$	optimum order quantity
$r$	requirements per time interval
$r_i$	requirements for the $i$ th time interval
$S$	inventory level
$S_i$	inventory level at beginning of $i$ th interval
$s_i$	inventory level at end of $i$ th interval. Note. $s_i = S_i - r_i$ , and $S_i = s_{i-1} + q_i$
$S^*$	optimum inventory level at the beginning of a time interval
$t$	an interval of time
$t_s$	interval between placing orders, in units of time
$t_s^*$	optimum interval between placing orders
$T$	period for which a policy is being established
$R$	total requirement for period $T$
$C_1$	holding cost per unit of goods for a unit of time
$C_2$	shortage cost per unit of goods for a specified period
$C_s$	setup cost per production run
$TEC$	total expected relevant cost

- $TEC^*$  minimum (optimum) total expected relevant cost
- $P(r)$  probability of requiring  $r$  units, where  $r$  is a discrete variable
- $f(r)$  probability density function of  $r$ , where  $r$  is a continuous variable
- $P(r \leq S)$  probability of requiring  $S$  units or less, where  $r$  is a discrete variable
- $F(r)$  cumulative probability function of  $r$ , where  $r$  is a continuous variable
- $F(S) = \int_0^S f(r) dr$ , probability of requiring  $S$  or less units, where  $r$  is a continuous variable.

**Model I.** (See Fig. 3.) Given: (a) Demand is fixed and known. (b) Withdrawals from stock are continuous and at a constant rate. (c) No shortages are permitted.

The variable costs are:  $C_1$  and  $C_s$  (see Symbols above).

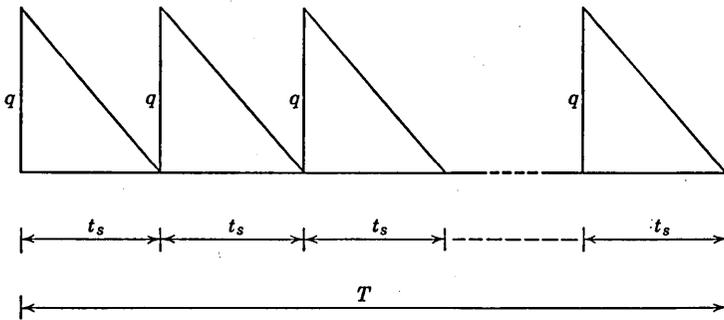


FIG. 3. Inventory situation for Model I (Ref. 1).

*Problem.* To determine: (1) how often to make a production run; (2) how many units should be made per run.

*Cost Equation.*

$$(22) \quad TEC = \frac{1}{2}C_1Tq + \frac{1}{q}C_sR. \bullet$$

*Solution.*

$$(23) \quad q^* = \sqrt{2 \frac{RC_s}{TC_1}}$$

$$(24) \quad t_s^* = \sqrt{2 \frac{TC_s}{RC_1}}$$

$$(25) \quad TEC^* = \sqrt{2RTC_1C_s}.$$

Note that Model I is a special case of Model II, wherein  $C_2 = \infty$ . Accordingly, by letting  $C_2 \rightarrow \infty$  in eqs. (27-30), one readily obtains eqs. (23-25).

**Model II.** (See Fig. 4.) Given: (a) Demand is known and fixed. (b) Shortages are permitted.

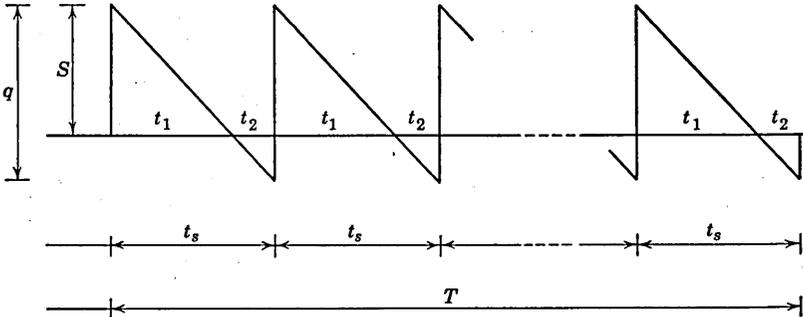


FIG. 4. Inventory situation for Model II (Ref. 1).

*Cost Equation.*

$$(26) \quad TEC(q, S) = \frac{1}{2q} S^2 C_1 T + \frac{(q - S)^2 C_2 T}{2q} + \frac{C_s R}{q}.$$

*Solution.*

$$(27) \quad q^* = \sqrt{2 \frac{RC_s}{TC_1}} \sqrt{\frac{C_1 + C_2}{C_2}},$$

$$(28) \quad S^* = \sqrt{2 \frac{RC_s}{TC_1}} \sqrt{\frac{C_2}{C_1 + C_2}},$$

$$(29) \quad t_s^* = \sqrt{2 \frac{TC_s}{RC_1}} \sqrt{\frac{C_1 + C_2}{C_1}},$$

$$(30) \quad TEC^* = \sqrt{2RTC_1 C_s} \sqrt{\frac{C_2}{C_1 + C_2}}.$$

**Model III.** Given: (a) Estimated variable demands and inputs, (b) discrete units, (c) shortages permitted (finite cost of shortage), (d) discontinuous distribution over time of withdrawals and input at a discontinuous rate, (e) known and constant reorder cycle time.

In this model and in Model VI, the cost of carrying an inventory of parts until they are used is not taken into consideration. Rather, in this elementary inventory situation, the cost of having excess parts that are never used is balanced against the cost of being short of parts when needed.

*Problem.* To determine how many units of a given part should be ordered at the time of the initial purchase order. Here, one is balancing the cost of

having excess parts that are *never* used against the cost of being short of parts when needed. No consideration is given to the cost of carrying the inventory of parts until they are used.

*Cost Equation.*

$$(31) \quad TEC(S) = C_1 \sum_{r=0}^S P(r)(S-r) + C_2 \sum_{r=S+1}^{\infty} P(r)(r-S).$$

*Solution.* The optimum value,  $S^*$ , is given by that value of  $S$  which satisfies the inequalities:

$$(32) \quad P(r \leq S-1) < \frac{C_2}{C_1 + C_2} < P(r \leq S).$$

For further discussion, derivation of this solution, and an example of its use, see Ref. 1, Chap. 8.

**Model IV.** Given: (a) Estimated variable demand and inputs; (b) continuous (rather than discrete) units; (c) shortages permitted, i.e., finite cost of shortage; (d) continuous distribution over time of withdrawals and input at a continuous rate; (e) known and constant reorder cycle time; (f) negative orders, i.e., returns, not considered.

*Problem.* To determine the initial order quantity, where one balances the holding cost against the shortage cost.

*Cost Equation.*

$$(33) \quad TEC(S) = C_1 \int_0^S (S-r)f(r) dr + C_2 \int_S^{\infty} (r-S)f(r) dr.$$

*Solution.* (See Ref. 1, Chap. 8.) The total expected cost is minimum for that value  $S$  which satisfies the condition

$$(34) \quad F(S) \equiv \int_0^S f(r) dr = \frac{C_2}{C_1 + C_2}.$$

**Model V.** Given: Conditions of Model IV *plus* a significant reorder lead time, i.e., one must take into account the lapse of time between the placing of an order and the receipt of the goods.

*Problem.* To determine how much (many) should be ordered for the  $k$ th day hence (where the reorder lead time is  $k$  days).

*Cost Equation.*

Let  $k$  = number of days in the order lead time,  
 $s_0$  = the stock level at the end of the period preceding the placing of the order,

$q_1, q_2, \dots, q_{k-1}$  = quantities already ordered and due to arrive on the 1st, 2nd,  $\dots$ ,  $(k-1)$ st days hence,

$q_k$  = quantity to be ordered for delivery  $k$  days hence,

$R' = \sum_{i=1}^k r_i$ , the total requirement over the order lead time,

$S'$  = total of amount available in stock at end of previous period and amounts ordered over the present  $k$ -day period; i.e.,

$$S' = s_0 + \sum_{i=1}^{k-1} q_i + q_k.$$

The problem is to determine the value of  $q_k$  which will minimize the total expected cost over the lead time period, i.e.,  $k$  days. However, since orders in the amounts  $q_1, q_2, \dots, q_{k-1}$  have already been placed, the total expected cost for the first  $k-1$  days has already been determined and is no longer subject to control. Hence, equivalently, the problem is one of determining the value of  $q_k$  which will minimize the total expected cost for the  $k$ th day only.

*Solution.* The stock at the end of the  $k$ th period can be expressed as

$$(35) \quad s_k = s_0 + \sum_{i=1}^{k-1} q_i + q_k - \sum_{i=1}^k r_i.$$

Then, since

$$(36) \quad S' = s_0 + \sum_{i=1}^{k-1} q_i + q_k,$$

and

$$(37) \quad R' = \sum_{i=1}^k r_i,$$

the total expected cost for the  $k$ th day will be given by

$$(38) \quad TEC(S') = C_1 \int_0^{S'} (S' - R')f(R') dR' + C_2 \int_{S'}^{\infty} (R' - S')f(R') dR'.$$

Equation (38) is equivalent to eq. (33); therefore the optimum value of  $S'$  is given by (see eq. 34)

$$(39) \quad F(S') \equiv \int_0^{S'} f(R') dR' = \frac{C_2}{C_1 + C_2}.$$

Once having determined the optimum value of  $S'$ , namely  $S'^*$ ,  $q_k^*$  can be determined from eq. (36), i.e.,

$$(40) \quad q_k^* = S'^* - \left( s_0 + \sum_{i=1}^{k-1} q_i \right).$$

See Ref. 1, Chap. 8, for further discussion and an example of the use of this solution.

**Model VI.** (See Fig. 5.) Given: Conditions of Model III except that withdrawals from stock are continuous and at a constant rate.

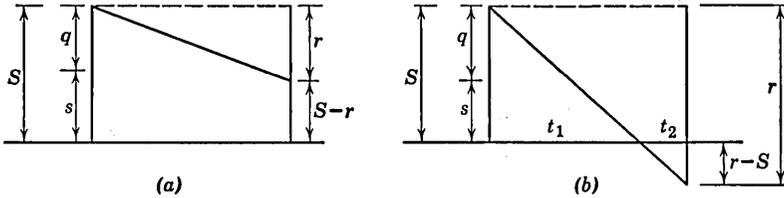


FIG. 5. Illustration for Model VI (Ref. 1).

*Problem.* To determine how many parts should be ordered at the time of the initial purchase order.

*Cost Equation.* (a) For  $r \leq S$ .

For a given value of  $r$ , the average number of units in stock over the order cycle period is given by

$$(41) \quad \frac{1}{2}[S + (S - r)] = S - \frac{r}{2}.$$

The expected cost, for a specific value of  $r$ , ( $r \leq S$ ), will then be

$$(42) \quad C_1 P(r)[S - \frac{1}{2}r].$$

Therefore, the total expected cost for all  $r \leq S$  will be

$$(43) \quad C_1 \sum_{r=0}^S P(r)(S - \frac{1}{2}r).$$

*Cost Equation.* (b) For  $r > S$ .

Here, as seen from Fig. 5, there will be no shortages  $t_1/(t_1 + t_2)$  part of the time, while shortages will occur  $t_2/(t_1 + t_2)$  part of the time. Now

$$(44) \quad \frac{t_1}{t_1 + t_2} = \frac{S}{r} \quad \text{and} \quad \frac{t_2}{t_1 + t_2} = \frac{r - S}{r}.$$

Furthermore, the average amount stocked is  $\frac{1}{2}S$ , and the average amount short is  $\frac{1}{2}(r - S)$ . Therefore, the holding cost for each value of  $r$  over the period  $S/r$  is given by

$$(45) \quad C_1 \left(\frac{S}{2}\right) \left(\frac{S}{r}\right) = \frac{1}{2r} S^2 C_1,$$

while the shortage cost for each  $r$ , over the period  $(r - S)/r$ , is given by

$$(46) \quad C_2 \left( \frac{r - S}{2} \right) \left( \frac{r - S}{r} \right) = C_2 \frac{(r - S)^2}{2r}.$$

Therefore, the total expected cost will be given by

$$(47) \quad TEC(S) = C_1 \sum_{r=0}^S P(r) \left( S - \frac{1}{2}r \right) + C_1 \sum_{r=S+1}^{\infty} P(r) \frac{S^2}{2r} \\ + C_2 \sum_{r=S+1}^{\infty} P(r) \frac{(r - S)^2}{2r}.$$

*Solution.* The optimum value of  $S$  is that which satisfies the condition

$$(48) \quad \left\{ P[r \leq (S - 1)] + \left( S - \frac{1}{2} \right) \sum_{r=S}^{\infty} \frac{P(r)}{r} \right\} < \frac{C_2}{C_1 + C_2} \\ < \left\{ P[r \leq S] + \left( S + \frac{1}{2} \right) \sum_{r=S+1}^{\infty} \frac{P(r)}{r} \right\}.$$

See Ref. 1, Chap. 8, for further discussion, an example of the use of this solution, and a case study employing this model.

### Inventory Models with Price Breaks

In this section, decision rules are given for the optimum lot size (or optimum purchase quantity) as derived for a class of inventory problems in which the unit manufacturing (or purchase) cost is variable, that is, subject to quantity discounts or price breaks. Specifically, this section will generalize on Model I (see Elementary Inventory Models), which describes a system in which demand is fixed and known, withdrawals from stock are continuous and at a constant rate, and no shortages are permitted. (See Fig. 3.)

**Symbols.** The following symbols are used:

$k_i$	cost per unit of manufacturing or purchasing for range $i$
$P$	monthly holding cost expressed as a decimal fraction of the value of the unit
$C_s$	setup cost per production run or, when for purchased parts, the setup cost associated with the procurement of the purchased items
$TEK$	total expected cost
$TEK^*$	minimum (optimum) total expected cost

As before,

$T$	the period of time for which the decision rules are being determined
$R$	total requirement during period $T$
$t_s$	interval between placing orders
$q$	input, or quantity ordered
$q^*$	optimum order quantity, i.e., economic lot size or economic purchase quantity

Finally, let the price break situation be described by the following:

Range	Quantity	Unit Purchase Price
$R_1$	$1 \leq q_1 < b_1$	$k_1$
$R_2$	$b_1 \leq q_2 < b_2$	$k_2$
...	...	...
$R_n$	$b_{n-1} \leq q_n$	$k_n$

where  $b_j$  ( $j = 1, 2, \dots, n - 1$ ) are those quantities which determine the price breaks.

*Problem.* The problem can be stated as one of determining: (1) how often should parts be purchased; (2) how many units should be purchased at any one time.

*Basic Cost Equation.* The basic cost equation for the period  $T$  for any one value of the unit purchase cost  $k_1$  is given by

$$(49) \quad TEK = \frac{C_s R}{q} + k_1 R + \frac{1}{2} C_s T P + \frac{1}{2} k_1 T P q,$$

while the basic solution is given by

$$(50) \quad q^* = \sqrt{\frac{2C_s R}{k_1 T P}},$$

and

$$(51) \quad TEK^* = \sqrt{2k_1 T P C_s R} + k_1 R + \frac{1}{2} C_s T P.$$

**Solution.** *Decision Rules.* (See Ref. 1, Chap. 9.)

*One Price Break.*

1. Compute  $q_2^*$  from eq. (50), by using  $k_2$ . If  $q_2^* \geq b$ , then the optimum purchase quantity is  $q_2^*$ , that is,  $q^* = q_2^*$ .

2. If  $q_2^* < b$ , compute  $TEK^*(k_1)$  from eq. (51) [or, equivalently,  $TEK(q_1^*)$  from eq. (49)] and compare this with  $TEK(b_1)$  as given by eq. (49).

If  $TEK(q_1^*) < TEK(b_1)$ , then  $q^* = q_1^*$ .

If  $TEK(q_1^*) > TEK(b_1)$ , then  $q^* = b_1$ .

*Two Price Breaks.*

1. Compute  $q_3^*$ . If  $q_{3,0} \geq b_2$ , then  $q^* = q_3^*$ .

2. If  $q_3^* < b_2$ , compute  $q_2^*$ . If  $q_3^* < b_2$  and  $b_1 \leq q_2^* < b_2$ , proceed as in the case of one price break, i.e., compare  $TEK^*(k_2)$  with  $TEK(b_2)$  to determine the optimum purchase quantity.

3. If  $q_3^* < b_2$  and  $q_2^* < b_1$ , compute  $TEK^*(k_1)$  and compare it with  $TEK(b_1)$  and  $TEK(b_2)$  to determine the optimum purchase quantity.

$(n - 1)$  Price Breaks.

1. Compute  $q_n^*$ . If  $q_n^* \geq b_{n-1}$ , then  $q^* = q_n^*$ .
2. If  $q_n^* < b_{n-1}$ , compute  $q_{n-1}^*$ . If  $q_{n-1}^* \geq b_{n-2}$ , i.e.,  $b_{n-2} \leq q_{n-1} < b_{n-1}$ , proceed as for one price break, i.e., compare  $TEK^*(k_{n-1})$  with  $TEK(b_{n-1})$  to determine  $q^*$ .
3. If  $q_{n-1}^* < b_{n-2}$ , compute  $q_{n-2}^*$ . If  $q_{n-2}^* \geq b_{n-3}$ , proceed as for two price breaks, i.e., compare  $TEK^*(k_{n-2})$  with  $TEK(b_{n-2})$  and  $TEK(b_{n-1})$  to determine  $q^*$ .
4. If  $q_{n-2}^* < b_{n-3}$ , compute  $q_{n-3}^*$ . If  $q_{n-3}^* \geq b_{n-4}$ , compare  $TEK^*(k_{n-3})$  with  $TEK(b_{n-3})$ ,  $TEK(b_{n-2})$ , and  $TEK(b_{n-1})$ .
5. Continue in this manner until  $q_{n-j}^* \geq b_{n-j-1}$ , ( $0 \leq j \leq n - 1$ ), and then compare  $TEK^*(k_{n-j})$  with  $TEK(b_{n-j})$ ,  $TEK(b_{n-j+1})$ ,  $\dots$ ,  $TEK(b_{n-1})$  to determine the economic purchase quantity  $q^*$ . *Note.* Define  $b_0 = 1$  for this step.

### Inventory Models with Restrictions

In some inventory situations it is necessary to consider restrictions on production facilities, storage space, time, or money. When such restrictions are introduced in situations involving more than one product, it is necessary to allocate the limited available resources among the products. Models have been developed which enable one to determine how much of each item to produce (or purchase) under the specified restrictions. Such models are developed and solved in Ref. 1, Chap. 10. A brief description of the approach to the solution of such models is given in Sect. 2, Modified Lagrangian Multipliers. See also Refs. 14-16.

### Other Inventory Models

Arrow, Harris, and Marschak (Ref. 17), Eisenhart (Ref. 18), Tompkins (Ref. 19), and others have treated the problem of determining the optimum buffer stock needed to protect against shortages, where demand is uncertain. Whitin (Ref. 20) has investigated the interaction between buffer stocks and lot sizes. Dvoretzky, Kiefer, and Wolfowitz (Refs. 21 and 22) have shown the conditions under which optimum inventory levels can be found.

**Multistorage Points.** Berman and Clark (Ref. 23) have developed and solved specific models for systems in which a central warehouse supplies a number of field warehouses which, in turn, supply distributors.

**Dynamic Models.** The dynamic problem of inventory is one in which consideration must be given to the effect of a decision in the current period on subsequent periods.

A servomechanism approach to the dynamic inventory problem which utilizes feedback rules to adjust production to sales has been developed

and applied at Carnegie Institute of Technology (Refs. 8 and 24) for situations of uncertain demand. This procedure applies Norbert Wiener's autocorrelation methods (see Chap. 17). A related method has been developed by Vassian (Ref. 26).

A number of persons have developed approaches with linear programming techniques. Such linear programming models are designed primarily for situations with important seasonal fluctuations in demands. Charnes, Cooper, and Farr (Ref. 27) have treated this case while further assuming that demand is known. See also Dannerstedt (Ref. 28).

Bellman (Refs. 29-32) has developed "dynamic programming" which makes it possible to approach these problems through the calculus of variations. See also Bellman, Glicksberg, and Gross (Ref. 25). Holt, Modigliani, and Simon (Ref. 8) have developed "quadratic programming" and applied it to setting overall production levels for cases in which the cost functions are quadratic.

For excellent summaries of the great amount of pertinent research and application in the inventory area, see Whitin (Refs. 33 and 34) and Simon and Holt (Ref. 35). See also Ref. 1, Chaps. 8-10.

#### 4. ALLOCATION MODELS

**Types of Problems.** *Allocation models* are used to solve a class of problems which arise when (a) a number of *activities* are to be performed and there are alternative ways of doing them, and (b) *resources* or *facilities* are not available for performing each activity in the most effective way. The problem is to combine activities and resources in such a way as to maximize overall effectiveness. These problems are divisible into two types:

1. An amount of work to be done is specified. Certain resources are available; i.e., a fixed capacity and/or material for doing the job is available and, hence, constitutes a restriction or limitation. The problem is to use these limited facilities and/or materials to accomplish the *required work* in the most economical manner.

2. The facilities and/or materials which are to be used are considered to be fixed. The problem is to determine *what work*, if performed, will yield the maximum return on use of the facilities and/or materials.

**Linear Programming.** Generally speaking, linear programming techniques can be used to solve a special class of allocation problems for which the following *conditions* are satisfied:

1. There must exist an *objective*, such as profit, cost, or quantities, which is to be optimized and which can be expressed as, or represented by, a *linear* function.

2. There must be *restrictions* on the amount or extent of attainment of



### Solution of Linear Programming Problems

Among the several techniques which can be used to solve linear programming problems, the most important ones are the *simplex technique* and the *transportation technique*. There is also a special linear programming problem, called the *assignment problem*, for which special techniques greatly reduce the tremendous amount of computation that would otherwise follow from the use of either the transportation or simplex techniques. The assignment problem is discussed in Ref. 1, Chap. 12.

### Solution of Linear Programming Problems by the Simplex Technique

The solution of linear programming problems by the simplex technique may best be illustrated by the solution of a specific problem. The problem, simplified for purposes of illustration, may be stated as follows.

**PROBLEM.** A manufacturer wishes to maximize the profits associated with producing two products,  $R$  and  $S$ . Products  $R$  and  $S$  are manufactured by a two-stage process in which all initial operations are performed in machine center I and all final operations may be performed in either machine center IIA or in machine center IIB. Machine centers IIA and IIB are different from each other in the sense that, in general, for any given product they yield different unit rates and different unit profits. In addition, a certain amount of overtime has been made available in machine center IIA for the manufacture of products  $R$  and  $S$ . Since the use of overtime results in changes (decreases) in unit profits (but not in unit rates), let us denote separately, by machine center IIAA, any overtime use of machine center IIA.

The unit times required to manufacture products  $R$  and  $S$ , the hours available in each machine center, and the unit profits are given in Table 4. In this table,  $R_1$ ,  $R_2$ , and  $R_3$  denote the three possible combinations for producing  $R$ , and similarly,  $S_1$ ,  $S_2$ , and  $S_3$  are defined for product  $S$ .

TABLE 4. UNIT TIMES REQUIRED TO MANUFACTURE PRODUCTS  $R$  AND  $S$

Operation	Machine Center	Product $R$			Product $S$			Hours Available
		$R_1$	$R_2$	$R_3$	$S_1$	$S_2$	$S_3$	
1	I	0.01	0.01	0.01	0.03	0.03	0.03	850
2	IIA	0.02			0.05			700
	IIAA		0.02			0.05		100
	IIB			0.03			0.08	900
Profit per part (in dollars)		0.40	0.28	0.32	0.72	0.64	0.60	

The problem is to determine how much of each product should be made through the use of each possible combination of machine centers so as to maximize the total profits, and to keep in mind the prescribed limitations on the capacities of the machine centers. The assumption here is that one can sell all that one can produce. This is a simplification which may be removed very easily by imposing additional restrictions in the form of maximum permissible quantities of each product. (See Ref. 36.)

**Simplex Solution.** The simplex technique is a procedure which, through a series of repetitive arithmetic operations, progressively approaches, and ultimately reaches, an optimum solution. The procedure may be summarized briefly as follows:

1. The problem is first set up in mathematical form in which all relevant initial relationships and restrictions are stated.
2. The problem is then set up in tabular form.
3. An initial (feasible) solution is determined.
4. Alternative changes to this solution are evaluated.
5. A new solution is determined by introducing the "most favorable" alternative change.
6. Steps 4 and 5 are repeated to derive successively better solutions.
7. When, at any stage, step 4 evaluates no alternative choice favorably, the procedure is complete and gives an optimal solution.

More explicitly, the simplex technique is carried out as indicated in the following steps:

*Step 1.* Rephrase the problem in mathematical form. Let  $X_1, X_2, X_3, X_4, X_5, X_6$  denote the amounts to be made of products  $R_1, R_2, R_3, S_1, S_2, S_3$ , respectively. Then the total profit  $Z$  will be given by

$$(56) \quad Z = 0.40X_1 + 0.28X_2 + 0.32X_3 + 0.72X_4 + 0.64X_5 + 0.60X_6.$$

Furthermore, the restrictions to the problem will be given by

$$(57) \quad \begin{aligned} 0.01X_1 + 0.01X_2 + 0.01X_3 + 0.03X_4 + 0.03X_5 + 0.03X_6 &\leq 850 \\ 0.02X_1 + 0.05X_4 &\leq 700 \\ 0.02X_2 + 0.05X_5 &\leq 100 \\ 0.03X_3 + 0.08X_6 &\leq 900. \end{aligned}$$

Therefore, the problem may now be restated as follows: Determine the values of  $X_j \geq 0$  (where  $j = 1, 2, \dots, 6$ ) which maximize eq. (56) subject to the restrictions of eqs. (57).

The restrictions  $X_j \geq 0, j = 1, 2, \dots, 6$ , arise from the fact that, since the manufacturing process is irreversible, one must preclude the appearance of negative values for these variables.



TABLE 5. COLUMN VECTORS FOR SIMPLEX SOLUTION

$P_1$	$P_2$	$P_3$	$P_4$	$P_5$	$P_6$	$P_7$	$P_8$	$P_9$	$P_{10}$	$P_0$
0.01	0.01	0.01	0.03	0.03	0.03	1				850
0.02			0.05				1			700
	0.02			0.05				1		100
		0.03			0.08				1	900

or *identity* matrix. The set of vectors which form the identity matrix are, in turn, said to be a *unit basis* of the particular space of interest, which is, in this problem, a four-dimensional space. The basis vectors are linearly independent vectors in terms of which every point in the  $n$ -dimensional (here,  $n = 4$ ) space may be uniquely expressed and in terms of which a solution (or solutions) will be stated.

*Step 5.* The columns of Table 5 are now rearranged as shown in Table 6a. Then, a column labeled "Basis" is inserted to the left of the  $P_0$  column and, in this column, the basis vectors are listed. For this example, the slack vectors form the unit basis. In some problems for which some of the restrictions are stated either in terms of equalities or in terms of inequalities which impose minimum limits, so-called artificial vectors will have to be introduced in order to form a unit basis (see Ref. 36). It should be noted also that structural vectors may be such that they may be included in the unit basis.

Next, a row of  $C_j$ 's is added, where the  $C_j$ 's are defined as the coefficients of the corresponding  $X_j$ 's in the expression for  $Z$  given in eq. (56). Then, a column of  $C_i$ 's is added, these corresponding to the  $C_j$ 's, but having the subscript  $i$  to denote the row, rather than the subscript  $j$ , which is used to denote the column. The expression for  $Z$  can now be written as

$$(59) \quad Z = \sum_{j=1}^{10} C_j X_j.$$

*Step 6.* Next, add a row of numbers labeled  $Z_j$ , where  $j$  denotes the appropriate column. Letting  $X_{ij}$  denote the element in the  $i$ th row and  $j$ th column of the table, the  $Z_j$ 's (including  $Z_0$ ) are defined by

$$(60) \quad Z_j = \sum_i C_i X_{ij}.$$

TABLE 6. SIMPLEX METHOD (Ref. 1)

(a) First Feasible Solution

$C_i \backslash C_j$							0.40	0.28	0.32	0.72	0.64	0.60
	Basis	$P_0$	$P_7$	$P_8$	$P_9$	$P_{10}$	$P_1$	$P_2$	$P_3$	$P_4$	$P_5$	$P_6$
	$P_7$	850	1				0.01	0.01	0.01	0.03	0.03	0.03
←	$P_8$	700		1			0.02			0.05		
	$P_9$	100			1			0.02			0.05	
	$P_{10}$	900				1			0.03			0.08
	$Z_j$											
	$Z_j - C_j$						-0.40	-0.28	-0.32	-0.72*	-0.64	-0.60

(b) Second Feasible Solution

	$P_7$	430	1	-0.6			-0.002	0.01	0.01		0.03	0.03
→	0.72 $P_4$	14,000		20			0.4			1		
←	$P_9$	100			1			0.02			0.05	
	$P_{10}$	900				1			0.03			0.08
	$Z_j$	10,080		14.4			0.288			0.72		
	$Z_j - C_j$	10,080		14.4			-0.112	-0.28	-0.32		-0.64*	-0.60

TABLE 6. SIMPLEX METHOD (Ref. 1)—Continued  
(c) Third Feasible Solution

$C_i \backslash C_j$							0.40	0.28	0.32	0.72	0.64	0.60
	Basis	$P_0$	$P_7$	$P_8$	$P_9$	$P_{10}$	$P_1$	$P_2$	$P_3$	$P_4$	$P_5$	$P_6$
	$P_7$	370	1	-0.6	-0.6		-0.002	-0.002	0.01			0.03
	0.72 $P_4$	14,000		20			0.4			1		
→	0.64 $P_5$	2,000			20			0.4			1	
←	$P_{10}$	900				1			0.03			0.08
	$Z_j - C_j$	11,360		14.4	12.8		-0.112	-0.024	-0.32			-0.60*

(d) Fourth Feasible Solution

	$P_7$	32.5	1	-0.6	-0.6	$-\frac{3}{8}$	-0.002	-0.002	$-\frac{1}{800}$			
←	0.72 $P_4$	14,000		20			0.4			1		
	0.64 $P_5$	2,000			20			0.4			1	
→	0.60 $P_6$	11,250				$12\frac{1}{2}$			$\frac{3}{8}$			1
	$Z_j - C_j$	18,110		14.4	12.8	$7\frac{1}{2}$	-0.112*	-0.024	-0.095			

(e) Fifth Feasible Solution

$C_i \backslash C_j$							0.40	0.28	0.32	0.72	0.64	0.60
	Basis	$P_0$	$P_7$	$P_8$	$P_9$	$P_{10}$	$P_1$	$P_2$	$P_3$	$P_4$	$P_5$	$P_6$
	$P_7$	102.5	1	$-\frac{1}{2}$	-0.6	$-\frac{3}{8}$		-0.002	$-\frac{1}{800}$	0.005		
→	0.40 $P_1$	35,000		50			1			$\frac{5}{2}$		
	0.64 $P_5$	2,000			20			0.4			1	
←	0.60 $P_6$	11,250				$\frac{100}{8}$			$\frac{3}{8}$			1
	$Z_j - C_j$	22,030		20	12.8	$7\frac{1}{2}$		-0.024	-0.095*	0.28		

(f) Sixth Feasible Solution

	$P_7$	140	1	$-\frac{1}{2}$	-0.6	$-\frac{1}{3}$		-0.002		0.005		$\frac{1}{300}$
	0.40 $P_1$	35,000		50			1			$\frac{5}{2}$		
←	0.64 $P_5$	2,000			20			$\boxed{0.4}$			1	
→	0.32 $P_3$	30,000				$\frac{100}{3}$			1			$\frac{5}{3}$
	$Z_j - C_j$	24,880		20	12.8	$\frac{32}{3}$		-0.024*		0.28		$0.25\frac{1}{3}$

TABLE 6. SIMPLEX METHOD (Ref. 1)—Continued  
(g) Maximum Feasible Solution

$C_i \backslash C_j$							0.40	0.28	0.32	0.72	0.64	0.60
	Basis	$P_0$	$P_7$	$P_8$	$P_9$	$P_{10}$	$P_1$	$P_2$	$P_3$	$P_4$	$P_5$	$P_6$
	$P_7$	150	1	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{3}$				$\frac{1}{200}$	$\frac{1}{200}$	$\frac{1}{300}$
0.40	$P_1$	35,000		50			1			$\frac{5}{2}$		
0.28	$P_2$	5,000			50			1			$\frac{5}{2}$	
0.32	$P_3$	30,000				$\frac{100}{3}$			1			$\frac{8}{3}$
	$Z_j - C_j$	25,000		20	14	$10\frac{2}{3}$				0.28	0.06	$0.25\frac{1}{3}$

*Step 7.* A row labeled  $Z_j - C_j$  is entered into the table and for any column, say  $j_0$ , consists of the corresponding  $C_{j_0}$  subtracted from the value of  $Z_{j_0}$  which was entered in the previous row.

Steps 1 through 7 complete the first phase of the simplex technique calculations and result in what is known as a *feasible* solution to the problem, namely a solution which satisfies all the restrictions but which does not necessarily yield the optimum result. This feasible solution is given by the column vector  $P_0$  (Table 6a) in terms of the basis vectors  $P_7, P_8, P_9, P_{10}$ , namely,

$$(61) \quad X_7 = 850; \quad X_8 = 700; \quad X_9 = 100; \quad X_{10} = 900.$$

That is, the initial feasible program consists of "Do not use any of the time available in any of the machine centers; i.e., do nothing," thus resulting in a net profit of  $Z = 0$ .

**Optimum Solution Criteria.** Having obtained a feasible solution, one can proceed to the optimum solution by considering the following mutually exclusive and collectively exhaustive possibilities:

M1. Maximum  $Z = \infty$  (i.e., maximum  $Z$  is infinitely large) and has been obtained by means of the present program.

M2. Maximum  $Z$  is finite and has been obtained by means of the present program.

M3. An optimum program has not yet been achieved and a higher value of  $Z$  may be possible.

The simplex technique is such that possibilities M1 or M2 must be reached in a finite number of steps. Furthermore, if one remembers that  $X_{ij}$  denotes the element in the  $i$ th row and  $j$ th column of the table, the technique is such that, for a given tableau (i.e., table or matrix):

C1. If there exist any  $Z_j - C_j < 0$ , either M1 or M3 holds: (a) if all  $X_{ij} \leq 0$  in that column (for which  $Z_j - C_j < 0$ ), then M1 is true; (b) if some  $X_{ij} > 0$ , further calculations are required, i.e., M3 holds.

C2. If all  $Z_j - C_j \geq 0$ , a maximal  $Z$  has been obtained (M2).

**Iterative Procedure to an Optimum Solution.** In the example (Table 6a),  $Z_1 - C_1 < 0$  (as are  $Z_2 - C_2$  through  $Z_6 - C_6$ ) and, furthermore, some of the coefficients under  $P_1$  are greater than zero. Hence, by condition C1b, further calculations are required (i.e., condition M3 holds).

To discover new solutions, it is possible to proceed in a purely systematic fashion by the simplex technique. Furthermore, any new solution so obtained will never decrease the value of the objective functional (although an increase need not occur), and, as stated earlier, the optimal solution, if one exists, must be reached in a finite number of steps. Hence, the simplex technique is a converging iterative procedure.

*Step 8.* Of all the  $Z_j - C_j < 0$ , choose the most negative. (In the particular example, this is  $Z_4 - C_4 = -0.72$  and is so indicated by an asterisk in Table 6a.) This determines a particular  $P_j$  (namely  $P_4$ ) which will be introduced into the column labeled "Basis" in Table 6b.

*Step 9.* Determine the vector which this  $P_j$  will replace by dividing all the positive  $X_{ij}$  appearing in the  $P_j$  column into the corresponding  $X_{i0}$  which appears in the same row under  $P_0$ . (Since all the components of  $P_0$  must be non-negative, all these ratios must, in turn, be non-negative.) The smallest of these ratios then determines the vector to be replaced. In the present example,  $P_4$  is to replace one of the vectors  $P_7, P_8, P_9$ , or  $P_{10}$ . Under  $P_4$ , there are two positive  $X_{ij}$ , namely  $X_{7,4} = 0.03$  and  $X_{8,4} = 0.05$ . The division of these  $X_{ij}$ 's into the corresponding  $X_{i0}$ 's which appear under  $P_0$  gives a minimum of 14,000 (i.e.,  $700/0.05$ ). Thus,  $P_8$  is the vector to be replaced by  $P_4$ , so that a new basis is formed consisting of the vectors  $P_7, P_4, P_9$ , and  $P_{10}$  (see Table 6b).

*Step 10.* Let subscript  $k$  denote "coming in," subscript  $r$  denote "going out,"  $X'_{ij}$  denote the elements of the new matrix, and

$$(62) \quad \phi = \min_i \frac{X_{i0}}{X_{ik}}, \quad X_{ik} > 0,$$

[i.e.,  $\phi$  is the minimum of all ratios ( $X_{i0}/X_{ik}$ ) for  $X_{ik} > 0$ ]. The elements of the new matrix ( $X'_{ij}$ ) are calculated as follows. The elements,  $X'_{kj}$ , of the row corresponding to the vector just entered into the unit basis are calculated by

$$(63) \quad X'_{kj} = \frac{X_{rj}}{X_{rk}}.$$

The other elements ( $X'_{ij}$ ) of the new matrix are calculated by

$$(64) \quad X'_{ij} = X_{ij} - \left( \frac{X_{rj}}{X_{rk}} \right) X_{ik},$$

where eq. (64) also applies to the  $X_{i0}$ 's appearing under  $P_0$  and to the  $Z_j - C_j$  in the entire bottom row (but not to the  $Z_j$ 's in the second to the last row).

The new value of the profit function will be given by

$$(65) \quad Z'_0 = Z_0 - \phi(Z_k - C_k),$$

or, since  $C_0 = 0$ , the profit function will be given by

$$(66) \quad (Z_0 - C_0)' = (Z_0 - C_0) - \phi(Z_k - C_k).$$

For *example*, starting with Table 6a and proceeding to Table 6b, the most negative  $Z_j - C_j$  is  $Z_4 - C_4 = -0.72$ . Therefore  $k = 4$ . Hence, from eq. (62),

$$\phi = \min_i \frac{X_{i0}}{X_{i4}} \quad \text{for all } X_{i4} > 0,$$

i.e.,

$$\phi = \min \left( \frac{850}{0.03} = 28,333; \quad \frac{700}{0.05} = 14,000 \right) = 14,000.$$

Therefore,  $P_4$  will replace  $P_8$ ; or, in our notation,  $k = 4$ ,  $r = 8$ .

The elements in the  $P_4$  row of Table 6b are then computed by eq. (63)

$$X'_{4j} = \frac{X_{8j}}{X_{84}} = \left( \frac{X_{8j}}{0.05} \right).$$

Therefore,

$$X'_{40} = \left( \frac{X_{80}}{0.05} \right) = \left( \frac{700}{0.05} \right) = 14,000,$$

$$X'_{41} = \left( \frac{X_{81}}{0.05} \right) = \left( \frac{0.02}{0.05} \right) = 0.4, \text{ etc.}$$

For the elements of the other rows, where  $k = 4$ ,  $r = 8$  are substituted into eq. (64),

$$X'_{ij} = X_{ij} - \left( \frac{X_{8j}}{X_{84}} \right) X_{i4} = X_{ij} - \left( \frac{X_{8j}}{0.05} \right) X_{i4}.$$

Therefore,

$$X'_{70} = X_{70} - \left( \frac{X_{80}}{0.05} \right) (X_{74}) = 850 - \left( \frac{700}{0.05} \right) (0.03)$$

$$= 850 - (14,000)(0.03) = 850 - 420 = 430.$$

and

$$(Z_1 - C_1)' = (Z_1 - C_1) - \left( \frac{X_{81}}{X_{84}} \right) (Z_4 - C_4)$$

$$= (-0.40) - \left( \frac{0.02}{0.05} \right) (-0.72)$$

$$= (-0.4) - (0.4)(-0.72) = -0.4 + 0.288 = -0.112, \text{ etc.}$$

Finally, the new value of the profit functional will be given (see Table 6b) by

$$\begin{aligned}(Z_0 - C_0)' &= (Z_0 - C_0) - \phi(Z_4 - C_4) \\ &= 0 - 14,000(-0.72) = +10,080.\end{aligned}$$

The results are shown in Table 6b.

*Step 11.* The process is then repeated until such time as either condition M1 or condition M2 holds. For the present example, the solution is obtained after six iterations, i.e., six tableaux or matrices after the first (see Tables 6a-g). The final tableau, Table 6g, yields the optimum solution. (If any other optimum solutions existed, they would be indicated by  $Z_j - C_j = 0$  for  $j$ 's other than those appearing in the basis. Here,  $Z_j - C_j = 0$  for  $j = 1, 2, 3,$  and  $7$  only. Hence no other optimum solutions exist.) This optimum solution is also stated, both in terms of the number of parts and hours required, in Tables 7 and 8.

TABLE 7. OPTIMUM PROGRAM (NUMBER OF PARTS)

Product <i>R</i>		Product <i>S</i>	
$R_1$ (Centers I-IIA)	35,000 parts	$S_1$	0 parts
$R_2$ (I-IIAA)	5,000 parts	$S_2$	0 parts
$R_3$ (I-IIB)	30,000 parts	$S_3$	0 parts
Total	$R = 70,000$ parts	$S = 0$	
Total profit	\$25,000	+	0 = \$25,000

TABLE 8. OPTIMUM PROGRAM (HOURS)

Operation	Machine Center	Product <i>R</i>			Product <i>S</i>			Hours Used	Hours Avail.	Surplus Hours
		$R_1$	$R_2$	$R_3$	$S_1$	$S_2$	$S_3$			
1	I	350	50	300	0	0	0	700	850	150
2	IIA	700			0			700	700	0
	IIAA		100			0		100	100	0
	IIB			900			0	900	900	0

Thus, one readily sees that the optimum (most profitable) program under the prescribed conditions consists of manufacturing 70,000 units of product *R* to the complete exclusion of product *S*. Furthermore, by eq. (56) and also by  $(Z_0 - C_0)$  in the optimum tableau, the total profits will be

$$\begin{aligned}Z &= 0.40(35,000) + 0.28(5,000) + 0.32(30,000) + 0.72(0) \\ &\quad + 0.64(0) + 0.60(0) \\ &= \$25,000.\end{aligned}$$

*Alternate Step 8.* One should note at this point that the improvement from one tableau to the next is given by  $-\phi(Z_k - C_k)$ , see eq. (65). Furthermore, in practice, *one need not select the most negative number ( $Z_j - C_j$ ) but, rather, that negative number which yields the greatest improvement.* Thus, in the example,

$$-\phi(Z_1 - C_1) = -\left(\frac{700}{0.02}\right)(-0.40) = 14,000, \quad k = 1$$

$$-\phi(Z_2 - C_2) = -\left(\frac{100}{0.02}\right)(-0.28) = 1,400, \quad k = 2$$

$$-\phi(Z_3 - C_3) = -\left(\frac{900}{0.03}\right)(-0.32) = 9,600, \quad k = 3$$

$$-\phi(Z_4 - C_4) = -\left(\frac{700}{0.05}\right)(-0.72) = 10,080, \quad k = 4$$

$$-\phi(Z_5 - C_5) = -\left(\frac{100}{0.05}\right)(-0.64) = 1,280, \quad k = 5$$

$$-\phi(Z_6 - C_6) = -\left(\frac{900}{0.08}\right)(-0.60) = 6,750, \quad k = 6.$$

Therefore, instead of introducing  $P_4$  into the basis, a greater gain is achieved at this step through the introduction of  $P_1$ . In this particular example, following alternate Step 8 enables one to reach the optimum solution, Table 6g, in three less iterations.

**Further Restrictions in Linear Programming Problems.** Once having established the solution to a given linear programming problem, one may wish to consider (or evaluate) further restrictions on the variables. Thus, by referring to the preceding example, these restrictions may be in the form of: (1) minimum requirements for product  $S$ , (2) changes in the amount of time available in the machine centers, (3) changes in the prices of the various products, (4) changes in the unit production rates, e.g., due to the "introduction" of new equipment.

The simplex technique is such that, in general, new optimum solutions can easily be constructed in terms of such added restrictions by making use of the optimum solution to the original problem. For a full discussion of this point, see Ref. 1, Chap. 11.

**Solution of Minimization Problems by the Simplex Technique.** To solve minimization problems by the simplex technique one may, in Step 8, select either (1) the most *positive*  $Z_j - C_j$ , or alternately (2) the

most negative  $C_j - Z_j$ , and then proceed as before to the solution of the problem.

### The Transportation Problem

A linear programming problem for which a special technique has been developed is the so-called transportation problem which may be stated as follows.

PROBLEM. Determine  $X_{ij} \geq 0$  which minimize

$$(67) \quad Z = \sum_{i=1}^m \sum_{j=1}^n C_{ij} X_{ij},$$

such that

$$(68) \quad \sum_{j=1}^n X_{ij} = A_i \quad (i = 1, 2, \dots, m)$$

and

$$(69) \quad \sum_{i=1}^m X_{ij} = B_j \quad (j = 1, 2, \dots, n).$$

The transportation problem is obviously a special case of the general linear programming problem; hence, it can be solved by the simplex technique.

However, a special solution technique, far simpler than the simplex technique, has been developed for solving transportation problems and, quite appropriately, it is called the transportation technique (Ref. 39).

The *procedure in the transportation technique* is outlined as follows:

1. The problem is set up in tabular form.
  - a. All requirements are explicitly stated.
  - b. All permissible slack in the system is explicitly stated.
  - c. All appropriate costs and/or revenues are determined.
  - d. An objective function is determined.
  - e. The computational framework is established.
2. An initial solution is determined.
  - a. The initial solution must be technically feasible, i.e., it must meet all restrictions.
3. Alternative choices are evaluated.
  - a. Changes in the solution are made one at a time.
  - b. The evaluation is of the *complete* effect of each unit change.
4. The "most favorable" alternative is selected.
5. The number of units to be included in this change is determined.
  - a. Owing to the linear nature of the model, each unit contributes the same cost or profit difference.

- b. The limit on the number of units involved in the particular change is technical feasibility (non-negativity requirements).
6. A new solution is determined.
  - a. The elements to change and the number of units to include have been previously determined.
7. Steps 3 through 6 are repeated. The process is a converging iterative one.
8. When Step 3 evaluates no alternative favorably, the procedure is complete and one has an optimal solution.

EXAMPLE. This example, taken from Ref. 37a, deals with the problem of moving empty freight cars from three "excess" origins to five "deficiency" destinations in such a manner that, subject to the given restrictions, the total cost of the required movement will be a minimum. The specific conditions of the problem and the unit (per freight car) shipping costs are given in Tables 9 and 10.

Table 9 states that origins  $S_1$ ,  $S_2$ , and  $S_3$  have surpluses of 9, 4, and 8 empty freight cars, respectively, while destinations  $D_1$ ,  $D_2$ ,  $D_3$ ,  $D_4$ , and  $D_5$

TABLE 9. PHYSICAL PROGRAM REQUIREMENTS

Destinations Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	Surpluses
$S_1$	$X_{11}$	$X_{12}$	$X_{13}$	$X_{14}$	$X_{15}$	9
$S_2$	$X_{21}$	$X_{22}$	$X_{23}$	$X_{24}$	$X_{25}$	4
$S_3$	$X_{31}$	$X_{32}$	$X_{33}$	$X_{34}$	$X_{35}$	8
Deficiencies	3	5	4	6	3	21

are in need of 3, 5, 4, 6, and 3 cars, respectively. For simplicity, it has been assumed that the problem is self-contained, i.e., that the number of excess cars is equal to the number of deficiencies. Any transportation problem can be made self-contained through the introduction of dummy origins or destinations.

Table 10 lists the unit costs  $C_{ij}$  of sending an empty freight car from the  $i$ th origin to the  $j$ th destination.

TABLE 10. UNIT SHIPPING COSTS

Destinations Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$
$S_1$	$C_{11}$ -10	$C_{12}$ -20	$C_{13}$ -5	$C_{14}$ -9	$C_{15}$ -10
$S_2$	$C_{21}$ -2	$C_{22}$ -10	$C_{23}$ -8	$C_{24}$ -30	$C_{25}$ -6
$S_3$	$C_{31}$ -1	$C_{32}$ -20	$C_{33}$ -7	$C_{34}$ -10	$C_{35}$ -4

The solution to this problem by the transportation technique is obtained as indicated in the following steps.

*Step 1.* Set up the tables listing the physical program requirements (Table 9) and unit shipping costs (Table 10).

*Step 2. Obtaining a First Feasible Solution.* Write down an initial (feasible) solution, namely one which satisfies the movement requirements. (If a feasible solution also minimizes the total cost, it is then called an optimum feasible or, in this case, a minimal feasible solution). This can easily be done by applying a technique which has been developed by Dantzig, Ref. 38, and which Charnes and Cooper, Ref. 39, refer to as "the northwest corner rule."

The northwest corner rule may be stated as follows:

1. Start in the upper left-hand corner of Table 9 (requirements) and compare the amount available at  $S_1$  with the amount required at  $D_1$ . (a) If  $D_1 < S_1$ , i.e., if the amount needed at  $D_1$  is less than the number of units available at  $S_1$ , set  $X_{11}$  equal to  $D_1$  and proceed to cell  $X_{12}$ , i.e., proceed horizontally. (b) If  $D_1 = S_1$ , set  $X_{11}$  equal to  $D_1$  and proceed to cell  $X_{22}$ , i.e., proceed diagonally. (c) If  $D_1 > S_1$ , set  $X_{11}$  equal to  $S_1$  and proceed to  $X_{21}$ , i.e., proceed vertically.

2. Continue in this manner, step by step, away from the upper left corner until, finally, a value is reached in the lower right corner. Thus, in the present example (see Table 11), proceed as follows:

(a) Set  $X_{11}$  equal to 3, namely, the smaller of the amount available at  $S_1$  (9) and that needed at  $D_1$  (3).

TABLE 11. FIRST FEASIBLE SOLUTION

Destinations Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	Total Surpluses
$S_1$	③	⑤	①			9
$S_2$			③	①		4
$S_3$				⑤	③	8
Total deficiencies	3	5	4	6	3	21

(b) Proceed to  $X_{12}$  (rule 1a). Compare the number of units still available at  $S_1$  (namely 6) with the amount required at  $D_2$  (5) and, accordingly, let  $X_{12} = 5$ .

(c) Proceed to  $X_{13}$  (rule 1a), where there is but one unit left at  $S_1$  while four units are required at  $D_3$ . Thus set  $X_{13} = 1$ .

(d) Then proceed to  $X_{23}$  (rule 1c). Here  $X_{23} = 3$ .

(e) Continue and set  $X_{24} = 1$ ,  $X_{34} = 5$ , and, finally, in the southeast corner, set  $X_{35} = 3$ .

The feasible solution obtained by this northwest corner rule is shown in Table 11 by the circled values of the  $X_{ij}$ . That this set of values is a feasible solution is easily verified by checking the respective row and column requirements. The corresponding total cost of this solution is obtained by multiplying each circled  $X_{ij}$  in Table 11 by its corresponding  $C_{ij}$  in Table 10 and summing the products. For any cell in which no circled number appears, the corresponding  $X_{ij}$  is equal to zero. That is, the total cost is given by:

$$(70) \quad \text{Total cost} = \sum_{j=1}^5 \sum_{i=1}^3 C_{ij}X_{ij} = \sum_{i=1}^3 \sum_{j=1}^5 C_{ij}X_{ij}.$$

The total cost associated with the first feasible solution is computed as follows:

$$\begin{aligned} \text{T.C.} &= X_{11}C_{11} + X_{12}C_{12} + X_{13}C_{13} + X_{23}C_{23} + X_{24}C_{24} + X_{34}C_{34} \\ &\quad + X_{35}C_{35} \end{aligned}$$

$$\begin{aligned}
 \text{T.C.} &= (3)(-10) + (5)(-20) + (1)(-5) + (3)(-8) + (1)(-30) \\
 &\quad + (5)(-10) + (3)(-4) \\
 &= -\$251 \text{ (minus sign means "cost" rather than "profit").}
 \end{aligned}$$

*Step 3. Evaluation of Alternative Possibilities.* Evaluate alternative possibilities, i.e., evaluate the opportunity costs associated with not using the cells which do not contain circled numbers. Such an evaluation is illustrated by means of the program given in Table 11 and is exhibited in Table 12 (noncircled numbers only). This evaluation is obtained as follows. (For an alternative method of evaluation, see Ref. 1.)

TABLE 12. FIRST FEASIBLE SOLUTION (WITH EVALUATIONS):  $C = 251$

Destinations Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	Total
$S_1$	③	⑤	①	-18	-11	9
$S_2$	-11	-13	③	①	-18	4
$S_3$	8	17	19	⑤	③	8
Total	3	5	4	6	3	21

1. For any cell in which no circled number appears, describe a path in this manner. Locate the nearest circled-number cell in the same row which is such that another circled value lies in the same column.

Thus, in Table 12, if one starts with cell  $S_3D_1$  (row 3, column 1), the value ⑤ at  $S_3D_4$  (row 3, column 4) satisfies this requirement; i.e., it is the closest circled-number cell in the third row which has another circled value, ① at  $S_2D_4$ , in the same column (column 4). The circled number ③ in position  $S_3D_5$  fails to meet this requirement.

2. Make the horizontal and then the vertical moves so indicated. In the example, move from  $S_3D_1$  to  $S_3D_4$  (see Table 12).

3. Having made the prescribed horizontal and vertical moves, repeat the procedure outlined in Steps 1 and 2. For the example, this now gives cells  $S_2D_3$  and  $S_1D_3$  respectively; accordingly, one moves from ① at  $S_2D_4$  to ① at  $S_1D_3$  by way of ③ at  $S_2D_3$ .

4. Continue in this manner, moving from one circled number to another by, first, a horizontal move and then a vertical move until, by only a horizontal move, that column is reached in which the cell being evaluated is located. (The fewest steps possible should be used in this circumambulatory procedure.) Thus, to continue the example, this step is from ① at  $S_1D_3$  to ③ at  $S_1D_1$ .

5. Finally, move to the cell being evaluated (here,  $S_3D_1$ ). This completes the path necessary to evaluate the given cell. (*Note.* For the purposes of evaluation, the path ends, rather than starts, with the cell being evaluated.)

6. Form the sum, with alternate plus and minus signs, of the unit costs associated with the cells being traversed. (These unit costs are given in Table 10.) This is the (noncircled) evaluation to be entered into the appropriate cell in Table 12. Thus, for the example, one has for the evaluation of cell  $S_3D_1$ :

Path (Table 12)	$S_3D_4$	$S_2D_4$	$S_2D_3$	$S_1D_3$	$S_1D_1$	$S_3D_1$
Unit cost (Table 10)	-10	-30	-8	-5	-10	-1
Evaluation ( $S_3D_1$ )	$+(-10)$	$-(-30)$	$+(-8)$	$-(-5)$	$+(-10)$	$-(-1) = +8$

Accordingly, one enters +8 in cell  $S_3D_1$  of Table 12.

7. Repeat the procedure outlined until all cells not containing circled numbers are evaluated.

*Step 4. Iterative Procedure toward an Optimum Solution.* If the noncircled numbers (the evaluations) are all non-negative, an optimum has been achieved. If one or more noncircled numbers are negative, further improvement with respect to the objective function is possible (e.g., the negative numbers in  $S_1D_4$ ,  $S_2D_2$ , etc., in Table 12). (At this stage, it should be quite apparent that one must be careful to circle the values of  $X_{ij}$  obtained in a feasible solution in order to distinguish them from the "evaluation" numbers which are also in the same table.)

Improvement is obtained by an iterative procedure in which one proceeds as follows:

(a) Of the one or more negative values which appear, select the most negative one, say  $-N$ . If there are more than one such values, any one of these may be selected arbitrarily.

(b) Retrace the path used to obtain this most negative value.

(c) Select those circled values which were preceded by a plus sign in the alternation between plus and minus and, of these, choose the one with the smallest value written in its circle, say  $m$ .

(d) One is now ready to form a new table, wherein one replaces the most negative value,  $-N$ , by this smallest value,  $m$ .

(e) Circle the number  $m$  and then enter all the other circles (except the

one which contained the value  $m$  in the previous program) in their previous cells, but without any numbers inside.

The improvement in cost from one program to the next will then be equal to  $mN$ . Furthermore, as with the simplex technique, one need not select the most negative number. It is permissible, and sometimes advantageous, to select the first negative number which appears. Since the improvement from one program to the next is given by  $mN$ , a study of Table 12 shows that selections of  $S_2D_1$ ,  $S_2D_5$ ,  $S_2D_5$ , or  $S_1D_5$  would have resulted in improvements of 33, 39, 18, and 11 respectively, as compared with the

TABLE 13. ITERATIVE PROCEDURE TOWARD SOLUTION

(a) Value to Be Moved

Destinations Origins	Destinations					Total
	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	
$S_1$	○	○	○	①		9
$S_2$			○			4
$S_3$				○	○	8
Total	3	5	4	6	3	21

(b) Second Feasible Solution:  $C = 233$

Destinations Origins	Destinations					Total
	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	
$S_1$	③	⑤	⑦	①	7	9
$S_2$	-11	<b>-13</b>	④	18	0	4
$S_3$	-10	-1	1	⑤	③	8
Total	3	5	4	6	3	21

improvement of 18 resulting from the selection of  $S_1D_4$ . Another alternative is to examine all products,  $mN$ , and select that negative numbered cell which results in the greatest improvement, in this case,  $S_2D_5$ .

Thus in Table 12 the most negative number is  $-18$  and appears in both cells  $S_1D_4$  and  $S_2D_5$  (i.e.,  $-N = -18$ ). For such ties, one may arbitrarily select either of the cells containing this most negative number. Here, cell  $S_1D_4$  is chosen. Retracing the path used to obtain the  $-18$  value in cell  $S_1D_4$ , one then obtains  $+S_1D_3$ ,  $-S_2D_3$ ,  $+S_2D_4$ ,  $-S_1D_4$ . Of those preceded by a plus sign, namely  $S_1D_3$  and  $S_2D_4$ , both have the circled value ① in their cells. Consequently, either one of these may be chosen as the circled value to be moved. In this case, cell  $S_2D_4$  is arbitrarily chosen. The circled value ① is then entered into cell  $S_1D_4$  (see Table 13a, i.e., that cell where  $-18$  appeared in Table 12). (Therefore, the improvement over the program given in Table 12 will be  $1 \times 18 = 18$  cost units. That is, the next program (Table 13b) will cost  $251 - 18 = 233$  cost units.) The other circles (without numbers) are then entered in the same positions as before (see Table 13a).

*Step 5.* A new feasible solution is obtained by filling in the circles according to the given surplus-deficiency (input-output) specifications. This solution is given by the circled values in Table 13b.

*Step 6.* The program is then evaluated, as before, and negative (non-circled) numbers still appear.

*Step 7.* The process is successively repeated (Tables 14, 15, and 16) until, finally, in Table 16 the evaluation of the corresponding program given therein results in all (noncircled) numbers being non-negative. An optimum feasible solution, or program, therefore, has been reached.

TABLE 14. THIRD FEASIBLE SOLUTION:  $C = 181$

Destinations Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	Total
$S_1$	③	①	④	①	7	9
$S_2$	2	④	13	31	13	4
$S_3$	$-10$	-1	1	⑤	③	8
Total	3	5	4	6	3	21

TABLE 15. FOURTH FEASIBLE SOLUTION:  $C = 151$ 

Destinations Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	Total
	$S_1$	10	①	④	④	7
$S_2$	12	④	13	31	13	4
$S_3$	③	④-1	1	②	③	8
Total	3	5	4	6	3	21

TABLE 16. OPTIMUM FEASIBLE SOLUTION:  $C = 150$ 

Destinations Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	Total
	$S_1$	10	1	④	⑤	7
$S_2$	11	④	12	30	12	4
$S_3$	③	①	1	①	③	8
Total	3	5	4	6	3	21

**Alternate Optimum Programs.** If any of the evaluation numbers in the optimum tableau are zero, alternate optimum tableaux exist. These alternate optimum solutions are obtained by essentially the same procedure as that which was just given. The only variation is that the zeros (if any) which appear in the optimum feasible solutions are now treated in exactly the same manner as were the negative values.

Furthermore, given such alternate optimum programs, say  $\{P_1\}$ ,  $\{P_2\}$ ,  $\dots$ ,  $\{P_n\}$ , where  $\{P_n\}$  refers to the set of  $X_{ij}$  which form the  $n$ th optimum program, then

$$(71) \quad \{P_{n+1}\} = a_1\{P_1\} + a_2\{P_2\} + \dots + a_n\{P_n\}$$

is also an optimum program provided the  $a_i$  are non-negative constants such that

$$(72) \quad \sum_{i=1}^n a_i = a_1 + a_2 + a_3 + \dots + a_n = 1.$$

For *example*, the cost minimization problem represented by Table 17 has two optimum programs, namely those given in Tables 18 and 19. Table 19 is obtained from Table 18 (and vice versa) by treating the zero in cell  $S_3D_5$  of Table 18 (or cell  $S_3D_4$  of Table 19) as the "most negative number" and proceeding as before.

TABLE 17. UNIT COST MATRIX

Destina- tions Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	Total
	$S_1$	-2	-1	-4	-3	0
$S_2$	+1	-3	-5	-2	-1	7
$S_3$	-1	-4	-3	-2	-1	6
Total	2	2	5	4	5	18

TABLE 18. OPTIMUM PROGRAM FOR TABLE 17

Destina- tions Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	Total
	$S_1$	4	②	2	2	③
$S_2$	②	1	2	③	②	7
$S_3$	2	2	⑤	①	0	6
Total	2	2	5	4	5	18

TABLE 19. ALTERNATE OPTIMUM PROGRAM FOR TABLE 17

Destinations Origins	Destinations					Total
	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	
$S_1$	4	②	2	2	③	5
$S_2$	②	1	2	④	①	7
$S_3$	2	2	⑤	0	①	6
Total	2	2	5	4	5	18

An infinite number of derived optimum programs can now be obtained by forming what are called "convex linear combinations" of the two basic optimum programs. Thus, if we select two positive fractions whose sum is unity, e.g.,  $\frac{1}{4}$  and  $\frac{3}{4}$ , we can obtain a new optimum program by multiplying every element of the first program by  $\frac{1}{4}$  and every element of the second program by  $\frac{3}{4}$  and then adding corresponding cells. This yields the derived optimum program of Table 22 and is obtained as shown in Tables 20 and 21. Similarly, other optimum programs could be derived for other non-negative fractions whose sum is equal to 1. *Note.* In general, derived optimum programs will involve fractional answers. These programs are for use only where nonintegral answers are realistic.

TABLE 20.  $\frac{1}{4}$  TABLE 18 =

Destinations Origins	Destinations					Total
	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	
$S_1$		①/2			③/4	5
$S_2$	①/2			③/4	①/2	7
$S_3$			⑤/4	①/4		6
Total	2	2	5	4	5	18

TABLE 21.  $\frac{3}{4}$  TABLE 19 =

Destinations Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	Total
	$S_1$		$\left(\frac{3}{2}\right)$			$\left(\frac{9}{4}\right)$
$S_2$	$\left(\frac{3}{2}\right)$			$(3)$	$\left(\frac{3}{4}\right)$	7
$S_3$			$\left(\frac{1.5}{4}\right)$		$\left(\frac{3}{4}\right)$	6
Total	2	2	5	4	5	18

TABLE 22. A DERIVED OPTIMUM PROGRAM:  $\frac{1}{4}$  TABLE 20 +  $\frac{3}{4}$  TABLE 21

Destinations Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	Total
	$S_1$		$(2)$			$(3)$
$S_2$	$(2)$			$\left(3\frac{3}{4}\right)$	$\left(\frac{5}{4}\right)$	7
$S_3$			$(5)$	$\left(\frac{1}{4}\right)$	$\left(\frac{3}{4}\right)$	6
Total	2	2	5	4	5	18

**Solution of Maximization Problems by the Transportation Technique.** Although the exposition just given treats only a (linear) minimization problem, it should be obvious that the transportation technique is equally applicable to (linear) maximization problems. The only difference in solving maximization problems lies in the preparation of the "profit" matrix. Whereas in the minimization problem all costs are entered with a negative sign, here all profits (or whatever units are involved in the maximization problem) are entered without any modification of signs. Once the initial datum matrix is obtained, one proceeds to the solution exactly as previously outlined.

**Variations on the Transportation Technique.** Many variations on the transportation technique are available for the solution of transportation problems. One has already been cited with respect to the selection of the cell to be introduced into the new basis. A second variation, designed to decrease the number of iterations, involves a rearrangement of the cost matrix. By using the problem cited in Tables 9 and 10, this may be illustrated as follows.

1. Form a new matrix in which the first row and first column correspond to the cell yielding the least cost. In the example, this is  $S_3D_1$ . Enter the totals of 8 for  $S_3$  and 3 for  $D_1$  in the new matrix. Place the smallest of these two numbers in that cell,  $S_3D_1$ .

/	$D_1$					Total
$S_3$	3					8
Total	3					

2. This satisfies the requirement for  $D_1$ , but still leaves 5 units available at  $S_3$ . Hence, select the next least unit cost which involves  $S_3$ . In the example, this is  $-4$  in cell  $S_3D_5$ . Therefore, list  $D_5$  in the second column and enter the corresponding total (requirement) of 3. Compare the requirement of 3 units at  $D_5$  with the remaining availability of 5 units at  $S_3$ , and assign 3 units to cell  $S_3D_5$ .

/	$D_1$	$D_5$				Total
$S_3$	3	3				8
Total	3	3				

3. Since 2 units are still available at  $S_3$ , select the third highest cost, namely  $-7$  in  $S_3D_3$ . Enter  $D_3$  in the third column along with its total requirement of 4 units. Comparing the requirement of 4 units at  $D_3$  with the remaining availability of 2 ( $8 - 3 - 3$ ) units at  $S_3$ , assign 2 units to cell  $S_3D_3$ , thereby using all available units at  $S_3$  but leaving 2 units still to be assigned to  $D_3$ .

	$D_1$	$D_5$	$D_3$			Total
$S_3$	3	3	2			8
Total	3	3	4			

4. Compare the costs associated with  $D_3$  ( $C_{13} = -5$  and  $C_{23} = -8$ ) and select  $S_1$  as the entry for the second row and, with it, enter the availability at  $S_1$ , namely 9. Compare this availability at  $S_1$  (i.e., 9) with the remaining requirement at  $D_3$  (i.e.,  $2 = 4 - 2$ ) enter 2 units in cell  $S_1D_3$ , and thereby satisfy the requirement at  $D_3$ .

	$D_1$	$D_5$	$D_3$			Total
$S_3$	3	3	2			8
$S_1$						9
Total	3	3	4			

5. By proceeding in this fashion, the following matrix is obtained:

	$D_1$	$D_5$	$D_3$	$D_4$	$D_2$	Total
$S_3$	3	3	2			8
$S_1$			2	6	1	9
$S_2$					4	4
Total	3	3	4	6	5	21

The cost for this initial feasible solution is given by

$$3(-1) + 3(-4) + 2(-7) + 2(05) + 6(-9) + 1(-20) + 4(-10),$$

i.e., neglecting the minus sign which indicates cost,

$$\text{T.C.} = \$153,$$

as compared with the first feasible solution of \$251 obtained by the north-west corner rule (and with the optimum solution of \$150). Such a reshuffling of the cost matrix generally leads to a better (i.e., lower cost or higher profit) first feasible solution so that the optimum solution is usually reached after a smaller number of alterations.

The reader should note that this first feasible solution costing \$153 could have been obtained without reshuffling the matrix. One simply starts in the cell of lowest cost (here  $S_3D_1$ ) and proceeds accordingly.

For further details of the transportation technique, including a discussion of so-called degenerate cases, see Ref. 39. The mathematical derivation of the transportation technique is given in Ref. 40, Chap. 23.

**Alternate Method of Evaluating Cells in Transportation Technique.** An alternate evaluation technique (or procedure) is presented by means of the problem represented by Tables 23 and 24, namely the unit cost table and the table listing the first feasible solution of the transportation problem given earlier (see Tables 10 and 11). The evaluation technique presented here is a variation of that originally designed by Dantzig in Koopmans (Ref. 40, Chap. XXI), and is part of the procedure described in Henderson and Schlaifer (Ref. 41). The discussion of determining the costs of deviating from the optimum solution is given in Ref. 41. The first

TABLE 23. UNIT SHIPPING COSTS

Destinations Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$
	$S_1$	-10	-20	-5	-9
$S_2$	-2	-10	-8	-30	-6
$S_3$	-1	-20	-7	-10	-4

TABLE 24. FIRST FEASIBLE SOLUTION

Destinations Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	Total
	$S_1$	3	5	1		
$S_2$			3	1		4
$S_3$				5	3	8
Total	3	5	4	6	3	21

part of the evaluation procedure is to form a new table (Table 25) corresponding to Table 24, but listing the unit costs rather than the amounts to be shipped. These costs are given by the boldface numbers in Table 25.

Add to Table 25 a column labeled "Row Values" and a row labeled "Column Values" and calculate these values as follows:

1. Assign an arbitrary value to some one row or some one column. For purposes of illustration, let us assign the value 0 to row  $S_1$ .

2. Next, for every cell in row  $S_1$  which contains a circled number representing part of the feasible solution, assign a corresponding column value (which may be positive, negative, or zero) which is such that the sum of the column value and row value is equal to the unit cost rate.

More generally, if  $r_i$  is the row value of the  $i$ th row,  $c_j$  the column value of the  $j$ th column, and  $C_{ij}$  the unit cost for the cell in the  $i$ th row and  $j$ th

TABLE 25. UNIT COSTS AND FICTITIOUS COSTS CORRESPONDING TO FIRST FEASIBLE SOLUTION

Destina- tions Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	Row Values
$S_1$	-10	-20	-5	-27	-21	0
$S_2$	-13	-23	-8	-30	-24	-3
$S_3$	7	-3	12	-10	-4	17
Column Values	-10	-20	-5	-27	-21	

column which contains a circled number, then all row and column values are obtained by the equation

$$(73) \quad r_i + c_j = C_{ij}.$$

Thus, by assuming  $r_1 = 0$ , it can be immediately determined from eq. (73) that

$$c_1 = -10; \quad c_2 = -20; \quad c_3 = -5.$$

3. Next, since  $c_3 = -5$  and  $C_{23} = -8$ , determine that  $r_2 = -3$ .
4. Since  $r_2 = -3$  and  $C_{24} = -30$ , then  $c_4 = -27$ .
5. From  $c_4 = -27$  and  $C_{34} = -10$ , then  $r_3 = +17$  is obtained.
6. Finally, for  $r_3 = +17$  and  $C_{35} = -4$ ,  $c_5 = -21$  is obtained.

This procedure for assigning row and column values can be used for any solution-matrix which is nondegenerate, i.e., given a matrix of  $m$  rows and  $n$  columns, where the solution consists of exactly  $m + n - 1$  nonzero elements. (Any solution consisting of less than  $m + n - 1$  nonzero elements is said to be *degenerate*. Simple methods for dealing with degeneracy may be found in Charnes and Cooper (Ref. 39), Henderson and Schlaifer (Ref. 41), and Dantzig (Ref. 38).)

After all row and column values for Table 25 have been computed, the table can be completed by filling in the remaining cells according to eq. (73). This results in the lightface figures given in Table 25.

After Table 25 has been completed, the cell evaluations may be obtained as follows. Form a new table (Table 26) which consists of the unit cost

rates of Table 23 subtracted from the number in the corresponding cell of Table 25. That is, in symbolic notation,

$$\{\text{Table 26}\} = \{\text{Table 25}\} - \{\text{Table 23}\}.$$

The cells corresponding to movements which are part of the solution will contain zeros. These zeros are given in boldface type in Table 26. The

TABLE 26. CELL EVALUATIONS FOR THE FIRST FEASIBLE SOLUTION

Destinations Origins	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$
$S_1$	<b>0</b>	<b>0</b>	<b>0</b>	-18	-11
$S_2$	-11	-13	<b>0</b>	<b>0</b>	-18
$S_3$	8	17	19	<b>0</b>	<b>0</b>

resulting numbers for the remaining cells are given in lightface type and are the cell evaluations to be used in determining a better program or solution. (Comparison with Table 12 will show this to be true.)

When these cell evaluations have been determined, proceed as previously outlined in the section.

### Geometric Interpretation of the Linear Programming Problem

A geometric interpretation of the linear programming problem may be given by means of the following specific two-dimensional example.

PROBLEM 1. To determine  $X, Y \geq 0$  which maximize  $Z = 2X + 5Y$  subject to

$$(74) \quad \begin{aligned} X &\leq 4, \\ Y &\leq 3, \\ X + 2Y &\leq 8. \end{aligned}$$

The system of linear inequalities which constitute the restrictions results in the convex set of points given by polygon  $OABCD$  of Fig. 6. That is, any point  $(X, Y)$  on or within the polygon satisfies the entire system of inequalities (74). Hence, there exist an infinite number of solutions to system (74). The linear programming problem then is to select, from this

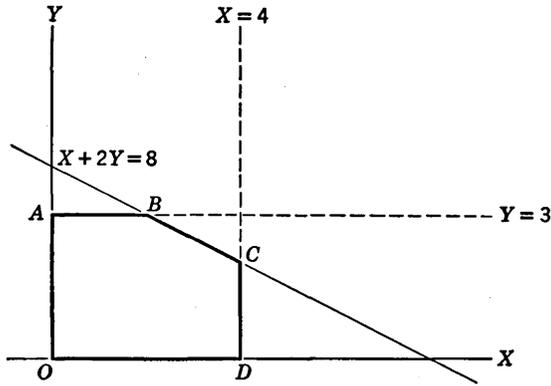


FIG. 6. Region satisfying restrictions (23) for non-negative  $X$  and  $Y$  (Ref. 1).

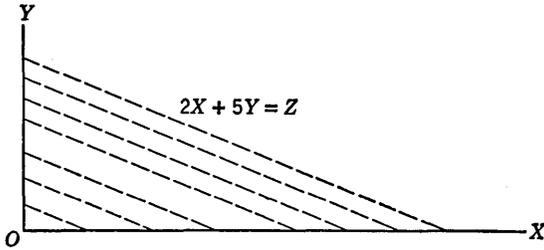


FIG. 7. Family of parallel straight lines,  $Z = 2X + 5Y$  (Ref. 1).

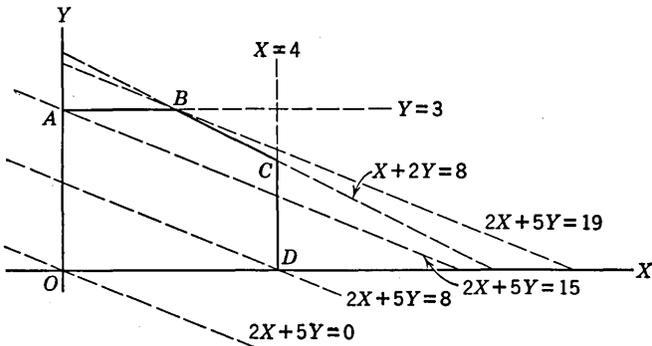


FIG. 8. Figure for geometric solution of linear programming problem (Ref. 1).

infinite number of points, the one or more points which will maximize the function  $Z = 2X + 5Y$ .

The function  $Z = 2X + 5Y$  is a one-parameter family of parallel lines; i.e., the function represents a family of parallel straight lines (of slope  $-\frac{2}{5}$ ) such that  $Z$  increases as the line gets farther removed from the origin; see Fig. 7. The problem may then be thought of as one of determining that line of the family,  $2X + 5Y = Z$ , which is farthest away from the origin but which still contains at least one point of the polygon  $OABCD$ .

Figure 8 shows how several members of the family  $Z = 2X + 5Y$  are related to the polygon  $OABCD$  and, in particular, shows that the solution is given by the coordinates of point  $B$ . Point  $B$  is the intersection of  $Y = 3$  and  $X + 2Y = 8$ . Hence,  $B$  is given by  $(2, 3)$  and, in turn,  $Z_{\max} = 2(2) + 5(3) = 19$ .

**Geometric Interpretation of the Simplex Method.** In order to exhibit, geometrically, what happens when one solves the problem by means of the simplex technique, the simplex solution of the example of Fig. 8 is given in Tables 27a-c. We see from those tables that the solution progresses from the point ( $X \equiv X_1 = 0, Y \equiv X_2 = 0$ ) to the point ( $X \equiv X_1 = 0, Y \equiv X_2 = 3$ ) to the point ( $X \equiv X_1 = 2, Y \equiv X_2 = 3$ ); i.e., referring to Fig. 8, from point  $O$  (origin) to point  $A$  to point  $B$ .

Mathematically, polygon  $OABCD$  (Fig. 6) constitutes a *convex set of points*; i.e., given any two points in the polygon, the line segment joining them is also in the polygon. An *extreme point* of a convex set is any point in the convex set which does not lie on a line segment joining some two other points of the set. Thus, the extreme points of polygon  $OABCD$  are points  $O, A, B, C,$  and  $D$ . The optimum solution to the linear programming problem will be at an extreme point and this optimum (extreme) point is reached by proceeding from one extreme point to another. Note that, in the example discussed here, the solution proceeded from extreme point  $O$  (Table 27a) to extreme point  $A$  (Table 27b) and, finally, to extreme point  $B$  (Table 27c).

**More Than One Optimum Solution.** If the example is now changed slightly to read:

**PROBLEM 2.** To determine  $X, Y \geq 0$  which maximize  $Z = X + 2Y$  subject to the restrictions

$$X \leq 4,$$

$$Y \leq 3,$$

$$X + 2Y \leq 8,$$

then Fig. 9 shows that the solution is given by either extreme point  $B$  or extreme point  $C$ . This is because  $X + 2Y = 8$  is *both* a boundary line of

TABLE 27. SIMPLEX METHOD

(a) Feasible Solution Corresponding to  $X = 0, Y = 0$  in Fig. 8

$C_j$		0	0	0	0	2	5
$C_i$							
	Basis	$P_0$	$P_3$	$P_4$	$P_5$	$P_1$	$P_2$
0	$P_3$	4	1	0	0	1	0
0	$P_4$	3	0	1	0	0	1
0	$P_5$	8	0	0	1	1	2
	$Z_j$	0	0	0	0	0	0
	$Z_j - C_j$	0	0	0	0	-2	-5

(b) Feasible Solution Corresponding to  $X = 0, Y = 3$  in Fig. 8

$C_j$		0	0	0	0	2	5
$C_i$							
	Basis	$P_0$	$P_3$	$P_4$	$P_5$	$P_1$	$P_2$
0	$P_3$	4	1	0	0	1	0
5	$P_2$	3	0	1	0	0	1
0	$P_5$	2	0	-2	1	1	0
	$Z_j - C_j$	15	0	5	0	-2	0

(c) Maximum Feasible Solution Corresponding to  $X = 2, Y = 3$  in Fig. 8

$C_j$		0	0	0	0	2	5
$C_i$							
	Basis	$P_0$	$P_3$	$P_4$	$P_5$	$P_1$	$P_2$
0	$P_3$	2	1	2	-1	0	0
5	$P_2$	3	0	1	0	0	1
2	$P_1$	2	0	-2	1	1	0
	$Z_j - C_j$	19	0	1	2	0	0

the polygon  $OABCD$  and also a member of the family of parallel lines  $Z = X + 2Y$ . Hence  $B = (2, 3)$  and  $C = (4, 2)$  both constitute solutions and yield the answer  $Z_{\max} = 8$ .

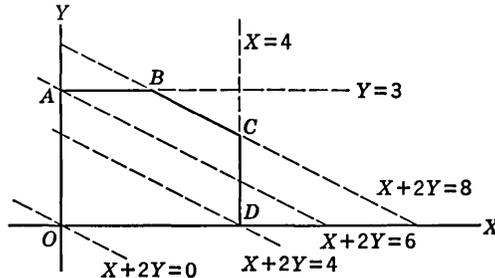


FIG. 9. Geometric solution of linear programming problem with more than one optimum solution (Ref. 1).

Furthermore, any convex linear combination of  $B$  and  $C$  will also be a solution, namely, any point on the line segment  $BC$ .

**The Dual Problem of Linear Programming**

By a *dual* theorem of linear programming, one has a choice of two problems to solve instead of just one. This is because every linear programming problem has a dual problem such that one involves maximizing a linear function and the other involves minimizing a linear function. Furthermore, if one solves a linear programming problem by the simplex technique, a tableau corresponding to an optimum solution automatically contains a solution to the dual problem. Thus, one is free to work with either the stated problem or its dual.

The dual problem of linear programming is illustrated by the example given earlier (eq. 74), namely:

**PROBLEM.** Determine  $X, Y \geq 0$  which maximize  $Z = 2X + 5Y$  subject to

$$(74) \quad \begin{aligned} X &\leq 4, \\ Y &\leq 3, \\ X + 2Y &\leq 8. \end{aligned}$$

This problem may be displayed in tabular form as is done in Table 28, that is, the restrictions may be read off by interpreting a light vertical line as  $+$  and the heavy vertical line as  $\leq$ . Furthermore, the function to be maximized is given by the bottom row, namely  $2X + 5Y$ . To obtain the dual problem, extend Table 28 as is done in Table 29. Then, by reading down each column as indicated, obtain the dual problem, namely:

TABLE 28. TABULAR FORM OF PROBLEM

$$\cong$$

	X	Y	
	1	0	4
	0	1	3
	1	2	8
Max	2	5	

TABLE 29. DUAL PROBLEM IN TABULAR FORM

$$\cong$$

	X	Y	Min
$W_1$	1	0	4
$W_2$	0	1	3
$W_3$	1	2	8
Max	2	5	

$$\cong$$

DUAL PROBLEM (see Table 29). Minimize  $g = 4W_1 + 3W_2 + 8W_3$  subject to

$$(75) \quad \begin{aligned} W_1 + W_3 &\cong 2, \\ W_2 + 2W_3 &\cong 5. \end{aligned}$$

The inequalities,  $\cong$ , are converted to equalities by the subtraction of non-negative slack variables. Then, since  $-1$  cannot be entered into a basis, one may also add artificial variables to provide for the basis. Thus,  $W_1 + W_3 \cong 2$  is first converted into  $W_1 + W_3 - W_4 = 2$ . Then the artificial variable  $W_6$  may be added to provide  $W_1 + W_3 - W_4 + W_6 = 2$ . For a detailed discussion, see Charnes, Cooper, and Henderson, Ref. 36.

If one returns to the simplex solution of the maximization problem of Table 27c, one sees that the following results are given:

$$Z_{\max} = 19,$$

and

$$(76) \quad \begin{array}{ll} X_1 = 2, & Z_1 - C_1 = 0, \\ X_2 = 3, & Z_2 - C_2 = 0, \\ X_3 = 2, & Z_3 - C_3 = 0, \\ X_4 = 0, & Z_4 - C_4 = 1, \\ X_5 = 0, & Z_5 - C_5 = 2. \end{array}$$

Now,  $X_3$ ,  $X_4$ , and  $X_5$  correspond to slack variables. Hence, if one starts with the *first* slack variable and renumbers the  $Z_j - C_j$  in order, and denotes these reordered  $Z_j - C_j$  by  $Z'_j - C'_j$ , one obtains

$$(77) \quad \begin{array}{ll} Z'_1 - C'_1 = 0 & \text{(corresponding to former } Z_3 - C_3), \\ Z'_2 - C'_2 = 1 & \text{(corresponding to former } Z_4 - C_4), \\ Z'_3 - C'_3 = 2 & \text{(corresponding to former } Z_5 - C_5), \\ Z'_4 - C'_4 = 0 & \text{(corresponding to former } Z_1 - C_1), \\ Z'_5 - C'_5 = 0 & \text{(corresponding to former } Z_2 - C_2). \end{array}$$

Setting  $W_j = Z'_j - C'_j$  gives the solution to the dual minimization problem; that is, if the minimization problem were to be solved by the simplex technique, the following results would be obtained:

$$g_{\min} = 19,$$

and

$$(78) \quad \begin{array}{ll} W_1 = 0, & -(g_1 - b_1) = 2, \\ W_2 = 1, & -(g_2 - b_2) = 0, \\ W_3 = 2, & -(g_3 - b_3) = 0, \\ W_4 = 0, & -(g_4 - b_4) = 2, \\ W_5 = 0, & -(g_5 - b_5) = 3, \end{array}$$

where the  $b_j$  are the corresponding coefficients of the  $W_j$  in the minimization function.

Conversely, given the solution to the minimization problem (i.e., given eq. 78), the solution to the dual maximization problem can be determined by starting with the first slack variable  $W_4$  and relabeling the  $-(g_j - b_j)$  in order. Hence, solution eqs. (76) would result.

For dual problems, it can be shown that  $Z_{\max} = g_{\min}$ ; in other words, that the two problems are equivalent (see Ref. 36 and Ref. 40, Chap. XIX). Hence, in solving a linear programming problem, one is free to work with either the stated problem or its dual. Since, as a rule of thumb, *the number of iterations required to solve a linear programming problem is equal to one to one and a half times the number of rows* (i.e., restrictions), one can, by an appropriate choice, facilitate the computation somewhat, especially in such cases where there exists a sizable difference in the number of rows for each of the two problems.

### A Short Cut in Solving Linear Programming Problems

One of the many advantages of both the transportation and simplex techniques is that judgment can be used to good advantage in facilitating the computations required in order to arrive at an optimum solution. In the transportation problem involving  $m$  rows and  $n$  columns, the use of judgment (or a good guess) simply requires designating  $m + n - 1$  cells which are expected to correspond to a solution. After these  $m + n - 1$  cells have been selected, proceed as in the transportation technique, first filling in these cells with circled numbers and then "evaluating" the remaining cells to determine whether or not the solution is an optimum one.

Consider the problem of Fig. 6 and eq. (74). It will be shown that given a "good" guess, the corresponding simplex matrix can be constructed. Then one proceeds to the optimum solution, if the solution guessed is not already optimum. This demonstrates how one may utilize judgment in the general linear programming problem (using the simplex technique).

PROBLEM. To determine  $X, Y \geq 0$  which maximize  $Z = 2X + 5Y$  subject to

$$(79) \quad \begin{aligned} X &\leq 4, \\ Y &\leq 3, \\ X + 2Y &\leq 8. \end{aligned}$$

Converting this system of inequalities to equalities by means of slack variables  $S_3, S_4,$  and  $S_5$  yields

$$(80) \quad \begin{aligned} X + S_3 &= 4, \\ Y + S_4 &= 3, \\ X + 2Y + S_5 &= 8. \end{aligned}$$

Now, suppose that one "guesses" or has reason to believe that the optimum solution is such that it will not involve  $X$ ; i.e., that the final solution will consist of  $Y, S_3,$  and  $S_5$ . This means, accordingly, that  $X = 0$  and  $S_4 = 0$ .

Hence, to obtain the "solution," i.e., the elements of the basis that would appear in the  $P_0$  column of the simplex tableau, one needs only to set  $X = 0$  and  $S_4 = 0$  in eqs. (80), yielding

$$\begin{aligned}
 S_3 &= 4, \\
 Y &= 3, \\
 2Y + S_5 &= 8,
 \end{aligned}
 \tag{81}$$

so that

$$Y = 3, \quad S_3 = 4, \quad S_5 = 2.
 \tag{82}$$

These values are then entered in the simplex tableau (see Table 30) under the column labeled  $P_0$ . Note that  $P_2$  corresponds to  $Y$ .

TABLE 30. FEASIBLE SOLUTION, SHORT-CUT APPROACH

$C_i \backslash C_j$						2	5
	Basis	$P_0$	$P_3$	$P_4$	$P_5$	$P_1$	$P_2$
0	$P_3$	4	1	0	0	1	0
5	$P_2(Y)$	3	0	1	0	0	1
0	$P_5$	2	0	-2	1	1	0
	$Z_j$	15	0	5	0	0	5
	$Z_j - C_j$	15	0	5	0	-2	0

Next, construct the body of the simplex matrix. Since each value of  $Z_j - C_j$  corresponds to the *minimum* cost of deviating from the optimum program by one unit of  $X_j$ , one can determine, for each  $j$ , the corresponding  $Z_j - C_j$  and the  $X_{ij}$  which appear in that column. For *example*, consider that one will deviate from the program of  $Y = 3$ ,  $S_3 = 4$ , and  $S_5 = 2$  by insisting that  $X = 1$ . One then needs to determine the *changes* in  $Y$ ,  $S_3$ , and  $S_5$  which result from the unit change in  $X$ . Therefore, solve

$$\begin{aligned}
 1 + S_3 &= 4, \\
 Y &= 3, \\
 1 + 2Y + S_5 &= 8,
 \end{aligned}
 \tag{83}$$

which result from eqs. (80) by letting  $X = 1$  and  $S_4 = 0$ .

Solving eqs. (83) yields

$$(84) \quad X = 1, \quad Y = 3, \quad S_3 = 3, \quad S_5 = 1.$$

Comparing eqs. (82) with (84) then shows that the following changes in  $Y$ ,  $S_3$ , and  $S_5$  occur because of a unit change in  $X$ :

$$(85) \quad \Delta Y = 0, \quad \Delta S_3 = 1, \quad \Delta S_5 = 1.$$

Therefore, in setting up a simplex tableau (see Table 30), these values would be inserted under the column labeled  $P_1$  which corresponds to the variable  $X$ .

Similarly, for  $S_4$  solve

$$(86) \quad \begin{aligned} S_3 &= 4, \\ Y + 1 &= 3, \\ 2Y + S_5 &= 8. \end{aligned}$$

This yields

$$(87) \quad Y = 2, \quad S_3 = 4, \quad S_5 = 4,$$

so that

$$(88) \quad \Delta Y = 1, \quad \Delta S_3 = 0, \quad \Delta S_5 = -2.$$

Insert these values in column  $P_4$  of Table 30.

Next, since  $P_2$ ,  $P_3$ , and  $P_5$  are in the basis, complete the corresponding columns (as is done in Table 30) by inserting 0's and 1's in the appropriate places.

Finally, compute the  $Z_j - C_j$ 's to determine whether the "solution" is optimum. This is done as at the outset of any simplex solution; i.e., first compute  $Z_j$  by

$$(89) \quad Z_j = \sum_i C_i X_{ij}$$

and then subtract the corresponding  $C_j$ . Since  $P_2$ ,  $P_3$ , and  $P_5$  are in the basis,  $Z_2 - C_2$ ,  $Z_3 - C_3$ , and  $Z_5 - C_5$  are all equal to zero. Additionally, applying eq. (89), yields

$$Z_1 - C_1 = 1(0) + 0(5) + 1(0) - 2 = -2,$$

$$Z_4 - C_4 = 0(0) + 1(5) + (-2)(0) - 0 = 5.$$

Thus Table 30 is completed and, not having an optimum solution (owing to  $Z_1 - C_1$  being negative), one can proceed to obtain the optimum solution as before.

The reader should note that Table 30 is identical with Table 27b and was

generated without a tableau such as is given in Table 27a. The same technique can also be applied to larger size problems so that, with a good estimate of the variables which will make up the solution, a great amount of computation might be eliminated.

### The Assignment Problem (See Ref. 1, Chap. 12)

The assignment problem is a special linear programming problem which may be stated mathematically as follows:

Determine  $X_{ij}$  which minimize

$$T = \sum_{i,j} a_{ij}X_{ij}$$

subject to

$$X_{ij} = X_{ij}^2, \quad i, j = 1, 2, \dots, n$$

$$\sum_{i=1}^n X_{ij} = \sum_{j=1}^n X_{ij} = 1, \quad i = 1, \dots, n; j = 1, \dots, n.$$

In other words, the assignment problem is such that:

(a)  $x_{ij} = 1$ , if the  $i$ th facility is assigned to the  $j$ th job; 0, otherwise.

(b) Each row and column of the solution matrix will have one element unity and all other elements zero.

For both the assignment problem and the transportation problem, so-called "methods of reduced matrices" exist which enable one to obtain the optimum solution with great ease.

## 5. WAITING TIME MODELS

**Problem Statement.** A waiting time problem arises when either units requiring service or the facilities which are available for providing service stand idle, i.e., wait. Problems involving waiting time fall into two different types, depending on their structure.

a. *Waiting line* problems involve arrivals which are randomly spaced and/or service time of random duration. This class of problems includes situations requiring either determination of the optimum number of service facilities or the optimum arrival rate (or times of arrival), or both. The solution of these "facility and scheduling" problems is obtained through what is called *waiting line* theory or (from the British) *queuing* theory.

Queuing theory dates back to the work of A. K. Erlang, who in 1908 published *Use of Waiting-Line Theory in the Danish Telephone System*. In Erlang's and subsequent work up to approximately 1945, applications were restricted in the main to the operation of telephone systems. Since then the theory has been extended and applied to a wide variety of phenomena. See Ref. 42 and Ref. 1, Chap. 14.

Reference 42 also contains an excellent list of activities to which queuing theory has been applied, a description of the use of the Monte Carlo technique in solving queuing problems, and a comprehensive list of references.

b. *Sequencing*. The second type of waiting time problem is not concerned with either controlling the times of arrivals or the number of facilities, but rather is concerned with the order or sequence in which service is provided to available units by a series of service points. This is the so-called *sequencing* problem. See Ref. 1, Chap. 16.

For a discussion of related problems such as the (assembly) *line-balancing* problem and the *traveling salesman* (or routing) problem, see Ref. 1.

### Problem Characteristics of Queuing Models

Every queuing or waiting line problem can be characterized by the following factors:

1. *Input*, the manner in which units arrive and become part of the waiting line.

2. *Stations*, the number of service units (or *channels*) operating on the units requiring service.

3. *Service policy*, limitations on the amount of service that can be rendered or is allowed.

4. *Queue discipline*, the order in which units are served, e.g., first come, first served; random selection for service; priority.

5. *Output*, the service provided and its duration. To specify a queue completely, all five factors must be described.

**Notation** (see Ref. 42).

$\lambda$	mean arrival rate (number of arrivals per unit time)
$\mu$	mean service rate per channel
$c$	the number of service channels
$c_f$	mean number of free service channels
$n$	number of units (customers) in the system
$k$	number of phases in the Erlang service case
$\rho$	utilization factor for service facility: $\rho = \lambda/c\mu$
$P_n(t)$	the probability that there be, at time $t$ , exactly $n$ units in the system, both waiting and in service
$p_n$	the steady-state (time-independent) probability that there be $n$ units in the system, both waiting and in service:

$$\sum_{n=0}^{n=\infty} P_n(t) = \sum_{n=0}^{n=\infty} p_n = 1$$

$c\rho$  traffic intensity in erlangs:

$$c\rho = \frac{\lambda}{\mu} = c - c_f = \sum_{n=0}^{n=c-1} np_n + \sum_{n=c}^{n=\infty} cp_n$$

$P(=0)$	the probability of no waiting
$P(>0)$	the probability of any waiting

$P(>\tau)$  the probability of waiting greater than time  $\tau$   
 $L$  the average number of units in the system, both waiting and in service:

$$L = \sum_{n=0}^{n=\infty} np_n$$

$L_q$  the average number of units waiting in the queue:

$$L_q = \sum_{n=c}^{\infty} (n - c)p_n = L - c + c_f$$

$W$  the average waiting time in the system:

$$W = - \int_0^{\infty} \tau dP(>\tau)$$

$A(t)$  cumulative distribution of times between arrivals with density function  $a(t)$   
 $B(t)$  cumulative distribution of service or holding times with density function  $b(t)$   
 $b_k(t)$  probability density for  $k$ th Erlang distribution

**Input.** Arrivals or inputs into a queuing system may occur at intervals of regular length. For such cases the cumulative distribution of time intervals between arrivals is given by the uniform distribution

$$A(t) = 0 \text{ for } t < t_0; 1 \text{ for } t \geq t_0.$$

If the input distribution is of Poisson type, the time intervals between arrivals are exponentially distributed. The cumulative distribution is then given by

$$A(t) = 1 - e^{-\lambda t}.$$

An intermediate type of input may be described by the Erlangian frequency distribution of times between arrivals

$$b_k(t) = \left[ \frac{(\lambda k)^k}{\Gamma(k)} \right] e^{-\lambda k t} t^{k-1}.$$

This yields the exponential distribution when  $k = 1$  and the uniform distribution when  $k$  becomes infinitely large.

As Saaty points out (Ref. 42), the normal distribution also produces a good fit to arrival data in some practical problems.

**Output (Service or Holding Times).** Distributions of service or holding times are defined as for arrivals or inputs. *In practice, Poisson inputs and exponential service times occur very frequently.*

**Assumptions Leading to a Poisson Input.** (See Ref. 42.) One has a Poisson input when the following assumptions are satisfied:

1. The total number of arrivals during any given time interval is independent of the number of arrivals that have already occurred prior to the beginning of the interval.

2. For any interval  $(t, t + dt)$ , the probability that exactly one arrival will occur is  $\lambda dt + O(dt^2)$ , where  $\lambda$  is a constant, while the probability that more than one arrival will occur is of the order of  $dt^2$  and may be neglected.

For a further discussion of the Poisson input and properties of a Poisson process, see Refs. 42 and 43.

**Assumptions Leading to an Exponential Holding Time Distribution.** If a channel is occupied at time  $t$ , the probability that it will become free during the following time interval  $dt$  is given by  $\mu dt$ , where  $\mu$  is a constant. (See Ref. 42.)

It follows that the frequency function of the service times is  $\mu e^{-\mu t}$ , while the mean duration of service is  $1/\mu$ , since the expected value of  $t$  is

$$E(t) = \mu \int_0^{\infty} t e^{-\mu t} dt = \frac{1}{\mu}.$$

### Queuing Models

To date, there have been essentially two different theoretical approaches to queuing, one through differential difference equations due to Erlang and the other through integral equations as studied by Lindley. The first approach may be illustrated by means of a single channel queuing system with both  $\lambda$  and  $\mu$  constant. A *Poisson input, exponential holding time, first-come, first-served single channel queue* is assumed.

**Differential Difference Equations.** If the operation starts with no items in the queue, then the following equations describe the given system. (See Ref. 1.)

$$P_0(t + dt) = P_0(t)(1 - \lambda dt) + P_1(t)\mu dt \quad (n = 0),$$

$$P_n(t + dt) = P_n(t)[1 - (\lambda + \mu) dt] + P_{n-1}(t)\lambda dt + P_{n+1}(t)\mu dt \quad (n \geq 1).$$

By transposing and passing to the limit with respect to  $dt$ , these equations become

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t) \quad (n = 0),$$

$$\frac{dP_n(t)}{dt} = -(\lambda + \mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t) \quad (n \geq 1).$$

The time-independent steady-state solution is obtained either by solving these time-dependent transient equations and letting  $t \rightarrow \infty$  in the solution, or by setting the derivatives with respect to time equal to zero, and solving the resulting steady-state equations. The latter approach yields, successively:

$$\lambda p_0 = \mu p_1,$$

$$(\lambda + \mu)p_n = \lambda p_{n-1} + \mu p_{n+1}.$$

By mathematical induction, these formulas then reduce to the single equation:

$$p_n = \rho^n(1 - \rho),$$

where  $\rho = \lambda/\mu$ , since  $c = 1$ .

The expected number of units in the system is given by

$$L = \sum_{n=0}^{\infty} np_n = (1 - \rho) \sum_{n=0}^{\infty} n\rho^n = \rho/(1 - \rho).$$

The expected number of units in the line is given by

$$L_q = L - \rho = \rho^2/(1 - \rho).$$

The expected waiting time is given by (see Ref. 42)

$$W = \int_0^{\infty} \tau dP(<\tau) = \frac{\rho}{\mu(1 - \rho)},$$

where  $P(<\tau)$ , the probability that an arrival waits a time less than  $\tau$ , is given by

$$P(<\tau) = 1 - \rho e^{-\mu\tau(1-\rho)}.$$

**Integral Equations.** For a résumé and illustration of the development and application of integral equations to queuing problems, see Ref. 42.

### Queuing Theory Formulas. *Single Channel*

**1. Poisson Input, First-Come, First-Served, Arbitrary Holding Time.** The expected total number of units waiting (queue plus service) is given by

$$L = \lambda + \mu + \frac{\text{Variance}(t) + (1/\mu)^2}{(2/\lambda)(1/\lambda - 1/\mu)},$$

where  $t$  has the holding time distribution.

This may also be expressed as

$$L = \lambda[W + (1/\mu)],$$

where

$$W = \frac{\rho}{2\mu(1 - \rho)} [1 + (s\mu)^2],$$

where  $s$  is the standard deviation of the service time.

(a) *Exponential Holding Time Distribution.* Here

$$p_n = (\lambda/\mu)^n(1 - \lambda/\mu)$$

and

$$L = \rho/(1 - \rho).$$

The expected number waiting in line is given by

$$L_q = \rho^2(1 - \rho).$$

Furthermore,

$$P(> \tau) = \rho e^{\mu(\rho-1)\tau}$$

and

$$W = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu(1 - \rho)}.$$

The expected number waiting, of those delayed, is

$$\frac{1}{(1 - \rho)}.$$

The expected waiting time of those delayed is

$$\frac{W}{P(> 0)} = \frac{1}{\mu(1 - \rho)}.$$

See Refs. 42-44.

(b) *Constant Holding Time Distribution* (Refs. 42 and 45). The steady-state equations are:

$$p_0 = 1 - \rho,$$

$$p_1 = (1 - \rho)(e^\rho - 1),$$

$$p_n = (1 - \rho) \sum_{k=1}^n (-1)^{n-k} e^{k\rho} \left[ \frac{(k\rho)^{n-k}}{(n-k)!} + \frac{(k\rho)^{n-k-1}}{(n-k-1)!} \right] \quad (n \geq 2).$$

Here,

$$P(> \tau) = \rho \sum_{i=0}^k e^{\rho(\mu\tau-i)} [-\rho(\mu\tau - i)]^i / i!,$$

where  $k$  is the largest integer less than or equal to  $\mu\tau$ .

$$W = \lambda \left[ 2\mu^2 \left( 1 - \frac{\lambda}{\mu} \right) \right] = \frac{\rho}{2\mu(1 - \rho)}.$$

Finally, the expected waiting time of those delayed is

$$\frac{1}{2\mu(1 - \rho)}.$$

(c) *Poisson Input, Erlangian Holding Time Distribution*. The probability density function for the  $k$ th Erlang holding time distribution is given by

$$b_k(t) = \left[ \frac{(\mu k)^k}{\Gamma(k)} \right] e^{-\mu k t} t^{k-1}.$$

The steady-state equations are (Ref. 45)

$$\lambda p_0 = \mu p_1 \quad (n = 0),$$

$$(\lambda + \mu)p_n = \mu p_{n+1} + \lambda p_{n-k} \quad (n \geq 1).$$

Here

$$L = \frac{\rho(\rho + 2k - \rho k)}{2k(1 - \rho)},$$

$$W = \frac{\rho(k + 1)}{2\mu k(1 - \rho)}.$$

## 2. Priority Discipline: Arbitrary Holding Time, Nonpreemptive Service (Refs. 42 and 46).

(a) *Finite Number of Priorities, N.* Assume a system with Poisson input for the  $k$ th priority with arrival rate  $\lambda_k$ , arbitrary holding time with service rate  $\mu_k$ , and a priority queue discipline. Items of different types enter the system with assigned priorities for service. Whenever the system is free to service an item, it selects items of highest priority on a first-come, first-served basis. However, if an item of higher priority enters the system while one of lower priority is in service, this service is not preempted, i.e., sent back to the waiting line. For this situation,

$$W_k = \frac{W_0}{(1 - \sigma_{k-1})(1 - \sigma_k)},$$

where

$$\rho_i = \frac{\lambda_i}{\mu_i},$$

$$\lambda = \sum_{i=1}^N \lambda_i,$$

$$\sigma_k = \sum_{i=1}^k \rho_i < 1,$$

$$W_0 = \frac{1}{2} \lambda \int_0^{\infty} t^2 dF(t),$$

$$F(t) = \frac{1}{\lambda} \sum_{i=1}^N \lambda_i F_i(t),$$

and where  $F_k(t)$  is the cumulative holding time distribution function for the  $k$ th priority.

The expected length of the line is given by

$$L = \sum_{i=1}^N \lambda_i W_i.$$

(b) *Two Priorities, Preemptive Service, Exponential Holding Time.* (See Refs. 42 and 47.) Priority 1 and 2 calls arrive at a single channel with arrival rates  $\lambda_1$  and  $\lambda_2$ , respectively. Both priorities have Poisson arrival distribution. Priority 1 calls in the queue enter the channel before all priority 2 calls in queue and replace any priority 2 calls in the channel on their arrival. The priority 2 call in the channel then reenters the queue.

Priority 1 and 2 calls have exponential service time distribution with service rates  $\mu_1$  and  $\mu_2$  respectively. Let  $\rho_1 = \lambda_1/\mu_1$ , and  $\rho_2 = \lambda_2/\mu_2$  where  $\lambda_1/\mu_1 + \lambda_2/\mu_2 < 1$ .

Let  $p_n^m$  be the probability that  $n$  priority 1 calls and  $m$  priority 2 calls are in the queue. The steady-state equations are:

$$\mu_1 p_{n+1}^m - (\mu_1 + \lambda_1 + \lambda_2) p_n^m + \lambda_1 p_{n-1}^m + \lambda_2 p_n^{m-1} = 0 \quad (m, n > 0),$$

$$\mu_1 p_1^m + \mu_2 p_0^{m+1} - (\mu_2 + \lambda_1 + \lambda_2) p_0^m + \lambda_2 p_0^{m-1} = 0 \quad (n = 0, m > 0),$$

$$\mu_1 p_{n+1}^0 - (\mu_1 + \lambda_1 + \lambda_2) p_n^0 + \lambda_1 p_{n-1}^0 = 0 \quad (m = 0, n > 0),$$

$$\mu_1 p_1^0 + \mu_2 p_0^1 - (\lambda_1 + \lambda_2) p_0^0 = 0 \quad (m = n = 0).$$

The expected number waiting, first priority, is given by

$$\frac{\rho_1}{1 - \rho_1}.$$

The expected number waiting, second priority, is given by

$$\rho_2 \left[ \frac{1 - \rho_1 + (\mu_1/\mu_2)\rho_1}{(1 - \rho_1)(1 - \rho_1 - \rho_2)} \right].$$

(c) *Continuous Number of Priorities* (Ref. 42). For an excellent discussion of results for a single channel priority queuing system with application to machine breakdowns, see Ref. 48. The number of available machines is assumed to be infinite. Priorities are assigned according to the length of time needed to service a machine with higher priorities being assigned to shorter jobs. Since the length of service time may correspond to any real number, a continuous number of priorities exists.

**3. Random Selection for Service: Impatient Customers, Exponential Holding Time.** (See Refs. 42 and 49.)

*Assumptions.* Poisson input, exponential holding time, random selection for service, items leave after a wait of time  $T_0$ .

*Steady-State Equations.*

$$-\lambda p_0 + (\mu + C_1)p_1 = 0 \quad (n = 0),$$

$$p_{n-1} - (\lambda + \mu + C_n)p_n + (\mu + C_{n+1})p_{n+1} = 0 \quad (n \geq 0),$$

where  $C_n$  is the average rate at which customers leave when there are  $n$  customers in the system and where  $p_0$  is obtained from

$$\sum_{n=0}^{\infty} p_n = 1.$$

*Solution.*

$$p_n = \lambda^n p_0 \prod_{k=1}^n (\mu + C_k)^{-1} \quad (n = 1, 2, \dots),$$

$$C_n = \frac{\mu \exp(-\mu T_0/n)}{1 - \exp(-\mu T_0/n)}.$$

**4. Limited Source: Exponential Holding Time.** (See Refs. 42 and 43.)

*Assumptions.* Input from a source having only a finite number  $m$  of customers. Exponential service time, single channel (servicing of  $m$  machines).

*Steady-State Equations.*

$$m\lambda p_0 = \mu p_1 \quad (n = 0)$$

$$[(m - n)\lambda + \mu]p_n = (m - n + 1)\lambda p_{n-1} + \mu p_{n+1} \quad (1 \leq n \leq m - 1),$$

$$\mu p_m = \lambda p_{m-1} \quad (n \geq m).$$

*Solution.*

$$p_n = m(m - 1)(m - n + 1)(\lambda/\mu)^n p_0,$$

$$p_0 = 1 - \sum_{n=1}^m p_n,$$

$$L = m - [(\lambda + \mu)/\lambda](1 - p_0).$$

**5. Constant Input: Exponential Holding Time.** (See Refs. 42 and 50.) For constant input at intervals of length  $\delta$ ,

$$p_n = p_0(1 - p_0)^n,$$

where  $p_0$  is given by

$$1 - p_0 = \exp(-\mu p_0 \delta).$$

Furthermore,

$$P(>\tau) = (1 - p_0) \exp(-\mu p_0 \delta),$$

and

$$W = - \int_0^{\infty} \tau dP(>\tau) = (1 - p_0)/\mu p_0.$$

**6. Queue Length-Dependent Parameters and Time-Dependent Parameters.** For an excellent résumé of queuing results for queue length-dependent and time-dependent parameters, see Ref. 42. See also Refs. 51-53.

**Queuing Theory Formulas. Two Channels in Series** (See Ref. 54.)**Exponential Holding Times.** (See Ref. 42.)

*Assumptions.* Poisson input with mean  $\lambda$ , two channels in series with exponential holding times,  $\mu_1$  and  $\mu_2$ , respectively. After finishing service at the first gate, the customer moves on to the second gate.

(a) *Unlimited Input.* The average distribution of customers throughout the system is given in the following table.

	Channel 1	Channel 2	Total System
Average number of customers waiting for service	$\frac{x_1}{1 - x_1}$	$\frac{x_2}{1 - x_2}$	$\frac{x_1^2}{1 - x_1} + \frac{x_2^2}{1 - x_2}$
Average number of customers being served	$x_1$	$x_2$	$x_1 + x_2$
Average total number of customers	$\frac{x_1}{1 - x_1}$	$\frac{x_2}{1 - x_2}$	$\frac{x_1}{1 - x_1} + \frac{x_2}{1 - x_2}$

The steady-state solution giving the probability that there are  $n_1$  customers waiting at the first gate and  $n_2$  at the second is given by

$$p(n_1, n_2) = (x_1)^{n_1}(x_2)^{n_2}(1 - x_1)(1 - x_2),$$

where

$$x_1 = \lambda/\mu_1 < 1 \quad \text{and} \quad x_2 = \lambda/\mu_2 < 1.$$

The probabilities of having  $n$  customers waiting at the first channel and at the second channel are, respectively,

$$p_1(n) = x_1^n(1 - x_1),$$

$$p_2(n) = x_2^n(1 - x_2).$$

(b) *Limited Input.* For a résumé of results for limited input, see Ref. 42.

**Queuing Theory Formulas. Three Channels in Series**

Results for the case of three channels in series can be found in Ref. 54.

**Queuing Theory Formulas. Multiple Channels in Parallel, Poisson Input**

An excellent résumé of results for both a finite and infinite number of channels can be found in Ref. 42. For the case with a *finite* number of channels, this includes: (a) identical exponential holding times, (b) identical constant holding times, (c) (priority discipline) different Poisson inputs, a finite number of priorities with the same exponential holding time (non-preemptive), and (d) (limited source) exponential holding time.

The résumé for an *infinite* number of channels covers: (a) exponential holding time and (b) limited source, exponential holding time.

### Sequencing Models

For a detailed discussion of sequencing models, see Ref. 1, Chap. 16. Only a few results are presented here.

**1. Two-Station and  $n$  Jobs, No Passing.** Consider the case of  $n$  jobs to be processed on two machines,  $A$  and  $B$ , with each job requiring the same sequence of operations and no passing allowed. The order (sequence) in which jobs are processed on machine  $A$  must be retained in processing these same jobs on machine  $B$ . It is assumed that material can be held *between* work stations so that, in the meantime, the preceding work station is left clear to start work on another job. It is further assumed (without loss of generality) that all jobs must first go to machine  $A$  and then machine  $B$ .

Let  $A_i$  = time required by job  $i$  on machine  $A$ ,

$B_i$  = time required by job  $i$  on machine  $B$ ,

$T$  = total elapsed time for jobs 1, 2,  $\dots$ ,  $n$ ,

$X_i$  = idle time on machine  $B$  from end of job  $i - 1$  to start of job  $i$ .

The sequencing problem is to minimize  $T$ , the total elapsed time.

The total elapsed time may be expressed as

$$T = \sum_{i=1}^n B_i + \sum_{i=1}^n X_i.$$

For any given set of items,  $\sum_{i=1}^n B_i$  is constant; therefore, the problem of minimizing  $T$  is equivalent to that of minimizing

$$D_n(S) = \sum_{i=1}^n X_i,$$

where  $D_n(S)$  is a function of the sequence  $S$ .

**Procedure for Finding the Optimum Sequence.** A procedure for finding the optimum sequence for two stations,  $n$  jobs, and no passing is due to Johnson (Ref. 55) and can be described by means of the example represented in Table 31.

TABLE 31. MACHINE TIMES (IN HOURS) FOR FIVE JOBS AND TWO MACHINES

$i$	$A_i$	$B_i$
1	3	6
2	7	2
3	4	7
4	5	3
5	7	4

*Step 1.* Examine the  $A_i$ 's and  $B_i$ 's and find the smallest value [ $\min(A_i, B_i)$ ]. In this illustrative case, this value is  $B_2 = 2$ .

*Step 2.* If the value determined falls in the  $A_i$  column, schedule this job first on machine  $A$ . If the value falls in the  $B_i$  column (as it does in this case), schedule the job last on machine  $A$ . Hence, job 2 goes last on machine  $A$ .

*Step 3.* Cross off the job just assigned and continue by repeating the procedure given in steps 1 and 2. In case of a tie, choose any job among those tied. In this illustrative case, once job 2 is assigned, the minimum value which remains is 3, and it occurs in  $A_1$  and  $B_4$ . There is a choice, so arbitrarily select  $A_1$ . Then job 1 goes on machine  $A$  first. Now  $B_4$  is the minimum remaining value. Hence, job 4 goes on machine  $A$  next to last. The minimum remaining value is 4, and it occurs in  $A_3$  and  $B_5$ . Then job 3 can be put on machine  $A$  second and job 5 on third to the last. The resulting sequence is optimum and is 1, 3, 5, 4, 2.

## 2. Three Stations and $n$ Jobs, No Passing.

Let  $Y_i$  = the idle time on the third machine before it starts work on the  $i$ th job,

$C_i$  = working time of the third machine on the  $i$ th job.

The total elapsed time for three stations and  $n$  jobs (no passing) is given by

$$T = \sum_{i=1}^n C_i + \sum_{i=1}^n Y_i.$$

Since  $\sum_{i=1}^n C_i$  is fixed, the problem is to minimize  $\sum_{i=1}^n Y_i$ . Johnson (Ref. 55) has found an optimum solution to this problem for the special case where either (1)  $\min A_i \geq \max B_i$  or (2)  $\min C_i \geq \max B_i$ . The first of these conditions is satisfied by means of an exact equality in the illustrative data given in Table 32.

TABLE 32. MACHINE TIMES (IN HOURS) FOR FIVE JOBS AND THREE MACHINES

$i$	$A_i$	$B_i$	$C_i$
1	8	5	4
2	10	6	9
3	6	2	8
4	7	3	6
5	11	4	5

To obtain an optimal sequence a new table, such as Table 33, is formed. Then the procedure (described in the preceding) for obtaining an optimum sequence for two stations is applied to Table 33. In this case, the following

TABLE 33. SUMS OF MACHINE TIMES (IN HOURS) FOR FIVE JOBS FOR FIRST AND INTERMEDIATE MACHINES AND FOR INTERMEDIATE AND LAST MACHINES

$i$	$A_i + B_i$	$B_i + C_i$
1	13	9
2	16	15
3	8	10
4	10	9
5	15	9

sequences arise and are optimum for the originally cited three-station problem:

3, 2, 1, 4, 5	3, 2, 1, 5, 4
3, 2, 4, 5, 1	3, 2, 5, 1, 4
3, 2, 4, 1, 5	3, 2, 5, 4, 1

In situations where the conditions  $\min A_i \geq \max B_i$  or  $\min C_i \geq \max B_i$  do not hold, no general procedure is available as yet for obtaining an optimum sequence. It follows that no general solution is yet available for the more general problem of  $n$  jobs and  $m$  machines, each job following an identical route with no passing allowed. However, the following statement holds: *For optimum sequences (the criterion being the total elapsed time), the total idle time of the last machine must be minimized.*

**3. Identical Routing, Passing Permitted.** Although each of  $n$  jobs may have to pass through each of  $m$  stations according to a specific route, the process characteristics do not always require that the order in which  $n$  jobs pass through each of the stations be identical, i.e., passing is permitted. Bellman (Ref. 56) and Johnson (Ref. 55) have shown, that for two or three station processes, the optimum sequence always involves the same ordering of jobs over each station. This result does not necessarily hold where more than three stations are involved.

**4. Different Routing.** In many production operations, particularly in job shops, the various jobs which must be done require different routing through the work stations or centers.

The problem of determining the optimum sequence for two jobs which have to be processed on  $m$  machines using two different routes has been treated by Akers and Friedman (Ref. 57) who, by means of Boolean algebra, have developed a technique for eliminating sequences which are technologically unfeasible. Their technique yields a subset of sequences, one or more of which is optimum.

The Akers-Friedman technique can also be extended to apply to the case of  $n$  jobs and  $m$  stations. See Refs. 56 and 57.

## 6. REPLACEMENT MODELS

**Problem Statement.** The theory of replacement is concerned with the prediction of replacement costs and the determination of the most economic replacement policy. There are two basic types of replacement problems concerned with (a) *items that deteriorate* with age and/or use and, (b) *items with probabilistic life spans* and with efficiencies that do not decline over their life spans.

For type (a) items that deteriorate or degenerate with age and use, *the problem is to determine when to replace equipment* so as to minimize the sum of costs due to loss of efficiency, on the one hand, and cost of new equipment, on the other hand.

For items whose efficiency declines over their life spans (e.g., machine tools, vehicles), prediction of costs involves determining those factors which contribute to increased operating cost, forced idle time, increased scrap, increased repair, etc.

The alternative to the increased cost of operating aging equipment is the cost of replacing old equipment with new. There is some age at which replacement of old equipment is more economical than continuation at the increased operating cost. At that age, the saving from the use of new equipment more than compensates for its initial cost.

For type (b) items that do not essentially deteriorate with age and use, but which have probabilistic life spans (e.g., light bulbs or radio tubes), *the problem is to determine when and how to replace the items* (i.e., individually or in groups) so as to minimize the sum of costs of (1) the items, (2) replacing items after failure, and (3) group replacements.

For a group of items with a probabilistic life span, the prediction of costs involves the estimation of the probability distribution of life spans and calculation from these of the predicted number of failures as a function of the age of the group of items. For several schemes for approximating the number of failures, see Refs. 43 and 58-62.

For a complete discussion of both types of replacement problems, see Ref. 1, Part VII.

### Replacement of Items That Deteriorate

The *measure of efficiency* used as a basis for determining optimum replacement decision rules is the *discounted value of all future costs* associated with any replacement policy. Discounted cost is the amount required at the time of the policy decision to build up a fund at compound interest large enough to pay the pertinent cost when due.

In general, the costs included in the replacement decisions cited here are *all* costs that depend upon the choice or age of the machine. See Ref. 1, Chap. 17, for a discussion of the relevant costs in replacement theory considerations.

**Cost Equation.** Consider a series of time periods 1, 2, 3, 4,  $\dots$ , of equal length, and let the costs incurred in these periods be  $C_1, C_2, C_3, C_4, \dots$ , respectively. *It is assumed throughout that, relevant to items that deteriorate, these costs are monotonically increasing with time.* Assume that each cost is paid at the beginning of the period in which it is incurred, that the *initial cost* of new equipment is  $A$ , and that the cost of investment is  $100r\%$  per period.

The *discounted value*  $K_n$  of *all* future costs associated with a policy of replacing equipment after each  $n$  periods is given by

$$(90) \quad K_n = \left[ A + C_1 + \frac{C_2}{1+r} + \frac{C_3}{(1+r)^2} + \dots + \frac{C_n}{(1+r)^{n-1}} \right] \\ + \left[ \frac{A + C_1}{(1+r)^n} + \frac{C_2}{(1+r)^{n+1}} + \dots + \frac{C_n}{(1+r)^{2n-1}} \right] + \dots$$

Equation (90) may also be written as

$$(91) \quad K_n = \frac{A + \sum_{i=1}^n [C_i / (1+r)^{i-1}]}{1 - [1/(1+r)]^n}$$

or, if

$$(92) \quad X = \frac{1}{1+r},$$

then

$$(93) \quad K_n = \frac{A + \sum_{i=1}^n C_i X^{i-1}}{1 - X^n}.$$

Now, if the best policy is replacement every  $n$  time periods, the two inequalities

$$(94) \quad K_{n+1} - K_n > 0 \quad \text{and} \quad K_{n-1} - K_n > 0$$

must hold. Furthermore, for the case where the  $C_n$  are monotonic increasing, these conditions are sufficient as well as necessary ones for  $K_n$  to be minimum.

From eq. (93),  $K_{n-1} - K_n > 0$  is equivalent to (see Ref. 1)

$$(95) \quad C_n < (1 - X)K_n,$$

and  $K_{n+1} - K_n > 0$  is equivalent to

$$(96) \quad C_{n+1} > (1 - X)K_n.$$

These inequalities, (95) and (96), may also be written as:

$$(97) \quad C_n < \frac{(A + C_1) + C_2X + \cdots + C_{n-1}X^{n-2}}{1 + X + X^2 + \cdots + X^{n-2}}$$

and

$$(98) \quad C_{n+1} > \frac{(A + C_1) + C_2X + \cdots + C_nX^{n-1}}{1 + X + X^2 + \cdots + X^{n-1}},$$

where the right-hand terms are the weighted averages of all costs up to and including the  $(n - 1)$ st and the  $n$ th periods, respectively.

**Decision Rules.** As a result of these two inequalities, the following decision rules for minimizing costs may be stated:

1. *Do not replace* if the next period's cost is less than the weighted average of previous costs.

2. *Replace* if the next period's cost is greater than the weighted average of previous costs.

For further discussion and a geometric interpretation of these decision rules, and also an illustration of their use, see Ref. 1, Chap. 17.

**Replacement of Items that Deteriorate by Different Equipment.** Here, one considers the replacement of equipment by new or alternate pieces of equipment other than those currently in use.

Let  $K'_n = \text{minimum discounted value of all future costs of new equipment,}$

$D_1, D_2, \dots, D_m = \text{costs in each future period incurred with present equipment,}$

$X = 1/(1 + r)$ , the discount factor,

$\pi_m = \text{discount value of all future costs if present equipment is discarded after } m \text{ periods.}$

**Cost Equation.**

$$(99) \quad \pi_m = D_1 + D_2X + \cdots + D_mX^{m-1} + K'_nX^m.$$

Therefore

$$(100) \quad \pi_{m+1} - \pi_m = D_{m+1}X^m + K'_n(X^{m+1} - X^m),$$

and

$$(101) \quad \pi_m - \pi_{m-1} = D_mX^{m-1} + K'_n(X^m - X^{m-1}).$$

The condition  $\pi_{m-1} - \pi_m > 0$  is equivalent to

$$(102) \quad D_m < (1 - X)K'_n$$

whereas the condition  $\pi_{m+1} - \pi_m > 0$  is equivalent to

$$(103) \quad D_{m+1} > (1 - X)K'_n.$$

Conditions (102) and (103) show that *the minimum cost is achieved by continuing the use of the old equipment until the cost for the next period is greater than  $(1 - X)K'_n$* , where  $(1 - X)K'_n$  is the weighted average of the costs of using the equipment for  $n$  periods between replacements.

### Replacement of Items That Fail

The second class of replacement problems is concerned with items that do not deteriorate markedly with service but which ultimately fail after a period of use. The period between installation and failure is not constant for any particular type of equipment but will follow some frequency distribution. This section is concerned only with items that fail with increasing probability as they age. Furthermore, it is assumed hereafter that all failures will be replaced. The problem, therefore, is to plan the replacement of items that have *not* failed.

Replacing a used but still functioning item with a new item is justified only if the cost of replacement is higher after failure than before, and if installing the new item reduces the probability of failure.

The replacement policy will depend upon the probability of failure. It is therefore of considerable importance to estimate the probability distribution of failures. Statistical techniques used in such "life testing" are being developed rapidly and a growing literature on the subject is becoming available. See Refs. 63-65. The costs of replacement before and after failure are the other important factors.

In this section, the cost of the alternatives of replacement or retention is considered and two policies are developed that minimize expected costs as a function of the cost of replacement, cost of failure, and probability of failure.

**Mortality Curves.** The initial information on the life characteristics of a light bulb, for *example*, may be shown in the form of a mortality curve. A group of  $N$  light bulbs is installed, and at the end of  $t$  equal time intervals the number of bulbs surviving equals some function of  $t$ , say  $S(t)$ . The proportion of the initial bulbs remaining is, then,

$$(104) \quad s(t) = \frac{S(t)}{N}.$$

A typical mortality table is shown in Table 34 giving, at regular intervals of time, the number of survivors out of an original group of 100,000 bulbs. Specifically, the mortality curve would result from column 2 in Table 34, namely that given by  $S(t)$ , and is given in Fig. 10.

TABLE 34. LIFE CHARACTERISTICS OF A LIGHT BULB: ORIGINAL POPULATION OF 100,000 UNITS

(1) Time Units Elapsed $t$	(2) Survivors $S(t)$	(3) Reduction in Survivors $S(t-1) - S(t)$	(4) Probability of Failure $p(t)$	(5) Conditional Probability of Failure $v_{t,0}$
0	100,000			
1	100,000	0	0	0
2	99,000	1,000	0.01	0.0100
3	98,000	1,000	0.01	0.0101
4	97,000	1,000	0.01	0.0102
5	96,000	1,000	0.01	0.0103
6	93,000	3,000	0.03	0.0312
7	87,000	6,000	0.06	0.0645
8	77,000	10,000	0.10	0.1149
9	63,000	14,000	0.14	0.1818
10	48,000	15,000	0.15	0.2381
11	32,000	16,000	0.16	0.3333
12	18,000	14,000	0.14	0.4375
13	10,000	8,000	0.08	0.4444
14	6,000	4,000	0.04	0.4000
15	3,000	3,000	0.03	0.5000
16	2,000	1,000	0.01	0.3333
17	1,000	1,000	0.01	0.5000
18	0	1,000	0.01	1.0000

Column (1), number of elapsed periods.

Column (2), survivors at end of period, based on figures supplied by a major light bulb manufacturer.

Column (3), rate of change of column (2).

Column (4), column (3) divided by 100,000.

Column (5), column (3) divided by value in column (2) for previous period.

**Life Span.** Perhaps a more familiar presentation of the life characteristics of a group of items is in the form of a probability distribution of life spans. Such a probability distribution may be derived from the mortality table by taking

$$(105) \quad \frac{S(t-1) - S(t)}{N} = p(t),$$

the proportion of units failing in time period  $t$ . (See Table 34.) This probability function,  $p(t)$ , is plotted against  $t$  in Fig. 11.

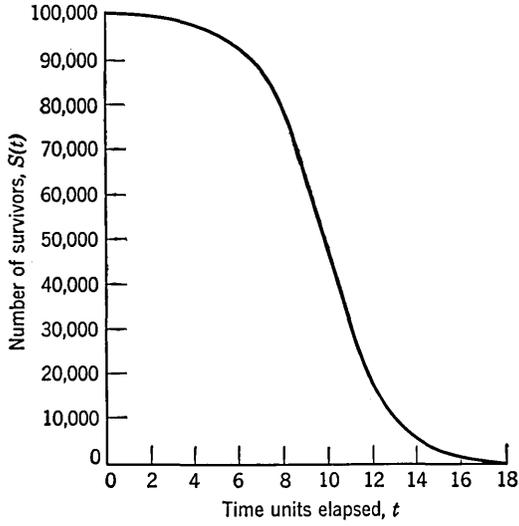


FIG. 10. Number of survivors after  $t$  periods of time. (Data from Table 34.)

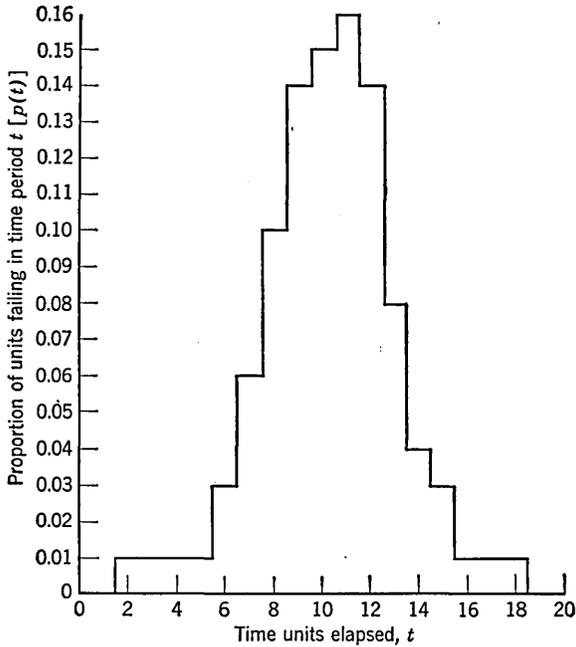


FIG. 11. Probability of failure in  $t$ th period of bulb installed at beginning of first period. (Data from Table 34.)

**Conditional Probability of Failure.** Another descriptive notion of life characteristics is the conditional probability of failure or its complement, the probability that an item at time  $t$  will survive to time  $t + 1$ . This probability is given by

$$(106) \quad v_{t,0} = \frac{S(t-1) - S(t)}{S(t-1)} = 1 - \frac{S(t)}{S(t-1)}$$

and is the proportion of surviving units failing in the subsequent period. (See Table 34.) This conditional probability function is plotted against  $t$  in Fig. 12.

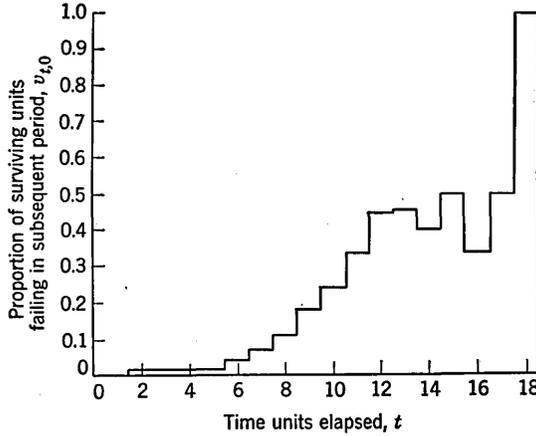


FIG. 12. Conditional probability of failure in  $t$ th period. (Data from Table 34.)

**Replacement Process.** It is assumed here that failures occur only at the end of a unit period of time. During the first  $t - 1$  time intervals, all failures occurring during any given time interval are replaced at the beginning of the next time interval. At the end of the  $t$ th time interval, *all* units are replaced regardless of their ages. The problem is to determine that value of  $t$  which will minimize total cost.

*Rate of Replacement.* The general expression for the number of units failing in time interval  $t$  is

$$(107) \quad f(t) = N \left\{ p(t) + \sum_{x=1}^{t-1} p(x)p(t-x) + \sum_{b=2}^{t-1} \left[ \sum_{x=1}^{b-1} p(x)p(b-x) \right] p(t-b) + \dots \right\},$$

where  $N$  = total units in the installation,  
 $p(x)$  = probability of failure at age  $x$ .

Table 35 illustrates the use of eq. (107) to determine the total number of failures in each time period  $t$ , based upon the data of Table 34.

TABLE 35. TOTAL FAILURES (REPLACEMENTS) IN EACH PERIOD  $t^a$

(1)	(2) Replacements		(1)	(2) Replacements	
Period	Current $f(t)$	Cumulative $\Sigma f(t)$	Period	Current $f(t)$	Cumulative $\Sigma f(t)$
1	0	0	21	12,047	162,167
2	1,000	1,000	22	11,706	173,873
3	1,000	2,000	23	10,820	184,693
4	1,010	3,010	24	9,697	194,390
5	1,020	4,030	25	8,700	203,090
6	3,030	7,060	26	8,288	211,378
7	6,040	13,100	27	8,413	219,791
8	10,090	23,190	28	8,862	228,653
9	14,201	37,391	29	9,523	238,176
10	15,392	52,783	30	10,100	248,276
11	16,665	69,448	31	10,413	258,689
12	15,000	84,448	32	10,507	269,196
13	9,480	93,928	33	10,348	279,544
14	6,175	100,103	34	9,999	289,543
15	6,160	106,263	35	9,636	299,179
16	5,521	111,784	36	9,079	308,258
17	7,309	119,093	37	9,220	317,478
18	9,317	128,410	38	9,271	326,749
19	10,181	138,591	39	9,447	336,196
20	11,529	150,120	40	9,669	345,865

Column (1), periods since original installation.

Column (2), calculated as described in text.

Column (3), cumulative sum of values in column (2).

<sup>a</sup> Data based on Table 34.

A second method for determining the number of failures in each period  $t$ , based upon the conditional probability of failure and using vector algebra is given in Ref. 1, Chap. 17.

**Cost of Replacement.** A second fundamental requirement of a useful replacement policy is that the cost of replacement after failure be greater than the cost of replacement before failure. This difference in cost is the source of savings required to compensate for the expense of reducing the probability of failure by replacing surviving units. Group replacing of units can cost less than replacement of failures by virtue of labor savings, volume discounts on materials, or for other reasons.

*Cost Equation.* Let  $K(t)$  = total cost from time of group installation until the end of  $t$  periods.

Then, if the entire group is replaced at intervals of length  $t$  periods,

$$\frac{K(t)}{t} = \text{Average cost per period of time.}$$

Let  $C_1$  = unit cost of replacement in a group,

$C_2$  = unit cost of individual replacement after failure,

$f(X)$  = number of failures in the  $X$ th period,

$N$  = number of units in the group.

Then the total cost,  $K(t)$ , will be given by

$$(108) \quad K(t) = NC_1 + C_2 \sum_{X=1}^{t-1} f(X).$$

Therefore, the cost per period is given by

$$(109) \quad \frac{K(t)}{t} = \frac{NC_1}{t} + \frac{C_2}{t} \sum_{X=1}^{t-1} f(X).$$

*Minimization of Costs.* Costs are minimized for a policy of group replacing after  $t$  periods if

$$(110) \quad \frac{K(t)}{t} < \frac{K(t-1)}{t-1},$$

and if

$$(111) \quad \frac{K(t)}{t} < \frac{K(t+1)}{t+1}.$$

By using eq. (109), conditions (110) and (111) may be rewritten respectively as

$$(112) \quad C_2 f(t-1) < \frac{NC_1 + C_2 \sum_{X=1}^{t-2} f(X)}{t-1},$$

and

$$(113) \quad C_2 f(t) > \frac{NC_1 + C_2 \sum_{X=1}^{t-1} f(X)}{t}.$$

Conditions (110) and (111) and, in turn, conditions (112) and (113), are necessary conditions for optimum group replacement. They are not sufficient as illustrated by the function  $F(t) = t \sin t$ ,  $0 \leq t \leq 4\pi$ , which satisfies these conditions for not one but two values of  $t$ , although the function has but one true (as opposed to relative) minimum point.

TABLE 36. AVERAGE COSTS FOR ALTERNATIVE GROUP REPLACEMENT POLICIES:  $C_1/C_2 = 0.25$

(Data from Table 35)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
$t$	$f(t)$ (Current)	$\sum_{X=1}^t f(X)$ (Cumulative)	$(t-1)f(t-1) - \sum_1^{t-2} f(X)$	$tf(t) - \sum_1^{t-1} f(X)$	Total Cost $K(t) = NC_1 + C_2 \sum_1^{t-1} f(X)$	Average Cost per Period $K(t)/t$
1	0	0	0	0	$25,000C_2$	$25,000C_2$
2	1,000	1,000	0	2,000	$25,000C_2$	$12,500C_2$
3	1,000	2,000	2,000	2,000	$26,000C_2$	$8,667C_2$
4	1,010	3,010	2,000	2,040	$27,000C_2$	$6,750C_2$
5	1,020	4,030	2,040	2,090	$28,010C_2$	$5,602C_2$
6	3,030	7,060	2,090	14,150	$29,030C_2$	$4,838C_2$
7	6,040	13,100	14,150	35,220	$32,060C_2$	$4,580C_2^a$
8	10,090	23,190	35,220	67,620	$38,109C_2$	$4,762C_2$
9	14,201	37,391	67,620	104,619	$48,190C_2$	$5,354C_2$
10	15,392	52,783	104,619	116,529	$62,391C_2$	$6,239C_2$

Column (1), number of periods between group replacements.

Column (2), number of replacements from Table 35.

Columns (3), (4), (5), calculated as indicated in column headings

Column (6), calculated as indicated with  $C_1 = 0.25C_2$ .

Column (7), column (6) divided by column (1).

<sup>a</sup> Therefore  $\hat{t} = 7$ .

Conditions (112) and (113) may be interpreted as follows:

1. *One should not group replace* at the end of the  $t$ th period if the cost of individual replacements at the end of the  $t$ th period is less than the average cost per period through the end of  $t$  periods.

2. *One should group replace* at the end of the  $t$ th period if the cost of individual replacements for the  $t$ th period is greater than the average cost per period through the end of  $t$  periods.

The use of these decision rules for the light bulb example (see Table 34) is illustrated in Table 36. For a full discussion of this replacement model and the solution of the light bulb example, see Ref. 1.

### Solution of Replacement Problems by Monte Carlo Technique

In the determination of optimum group replacement policies for items that fail, one needs to determine that value of  $t$  such that

$$(114) \quad K(t) = NC_1 + C_2\phi(t)$$

will be a minimum, when

$N$  = number of units in the group,

$C_1$  = unit cost of replacement in a group,

$C_2$  = unit cost of individual replacement after failure,

$\phi(t)$  = number of failures in time  $t$ .

**Failure Equations.** If  $f(t)$  is the probability density function of failure, then the expected number of first generation failures in time  $t$  is given by

$$(115) \quad \phi(t) = N \int_0^t f(t) dt = NF(t),$$

where

$$(116) \quad F(t) = \int_0^t f(t) dt.$$

Similarly, the number of second generation failures is given by

$$\phi_2(t) = N \int_0^t \left( \int_0^\alpha f(t) dt \right) \left( \int_0^{t-\alpha} f(t) dt \right) d\alpha.$$

That is,

$$(117) \quad \phi_2(t) = N \int_0^t F(\alpha)F(t - \alpha) d\alpha.$$

The number of third generation failures is given by

$$\phi_3(t) = N \int_{\alpha=0}^t \int_{\beta=\alpha}^t \left( \int_0^\alpha f(t) dt \right) \left( \int_0^{\beta-\alpha} f(t) dt \right) \left( \int_0^{t-\beta} f(t) dt \right) d\beta d\alpha.$$

That is,

$$(118) \quad \phi_3(t) = N \int_{\alpha=0}^t \int_{\beta=\alpha}^t F(\alpha)F(\beta - \alpha)F(t - \beta) d\beta d\alpha.$$

Therefore, the total expected number of failures in time  $t$  is given by

$$\phi(t) = \phi_1(t) + \phi_2(t) + \phi_3(t) + \dots.$$

That is,

$$(119) \quad \phi(t) = NF(t) + N \int_0^t F(\alpha)F(t - \alpha) d\alpha \\ + N \int_{\alpha=0}^t \int_{\beta=\alpha}^t F(\alpha)F(\beta - \alpha)F(t - \beta) d\beta d\alpha + \dots.$$

Unless simplifying assumptions are made relative to second and higher generation failures, it is almost impossible to obtain an analytic solution for  $\phi(t)$  as given by eq. (119). However, by the use of the Monte Carlo technique one can solve for values of  $\phi(t)$  without making any simplifying assumptions. That is, one can determine  $\phi(t)$  for many values of  $t$  and then construct  $K(t)$  as a function of  $t$  in order to determine the optimum group replacement policy.

**EXAMPLE.** *The Use of the Monte Carlo Technique in Solving Replacement Problems.* Assume that one wishes to determine the optimum group replacement policy for a group of light bulbs whose life pattern follows a normal distribution, the mean and standard deviations of which are 30 and 10 days, respectively. (That is,  $\mu = 30$  days and  $\sigma = 10$  days.) Furthermore, assume that

$$C_1 = \$0.50, \\ C_2 = \$1.00, \\ N = 10, \\ T = 360 \text{ days,}$$

where  $T$  is the total time period under consideration.

For purposes of illustration only, further assume that if group replacement is used, it can be done only at the end of 10, 20, 30, or 40 days.

A chart can then be set up and, by use of a table of random normal numbers, the total expected number of failures can be determined for each value of  $t$  ( $t = 10, 20, 30,$  or  $40$  days). Tables of random normal numbers are based on a mean of 0 and a standard deviation of 1. Hence, any number selected from the table of random normal numbers must first be multiplied by 10 and then added to (if positive) or subtracted from (if negative)

30. Thus, if the first random normal number selected from the table is 0.464, the adjusted random normal number will be

$$(0.464)(10) + 30 = 34.64.$$

That is, in the simulation of the light bulb system, the first bulb will last 34.64 (or 35) days before failing.

The next number from the table of random normal numbers is, say, 0.137, which is adjusted to  $(0.137)(10) + 30 = 31.37$ . Therefore, the replacement to the first bulb will last 31 days, that is,  $35 + 31 = 66$  days after the start of the analysis. Therefore one can expect the first bulb to burn out after 35 days and its replacement to last through the balance of the 40-day period under discussion.

This procedure is carried out for all ten lighting fixtures in the installation, and the expected number of failures for each of the intervals 10, 20, 30, and 40 days is determined as in Table 37.

TABLE 37. FAILURE TABLE FOR  $N = 10$

Fixture $t$	1	2	3	4	5	6	7	8	9	10	Total Failures
10	35	31	55	27	29	33	27	43	32	20	0
20	35	31	55	27	29	33	27	43	32	20	0
30	35	31	55	27, 38	29, 64	33	27, 59	43	32	20, 55	4
40	35, 66	31, 36, 62	55	27, 38, 75	29, 64	33, 47	27, 59	43	32, 36	20, 55	10

The entire procedure was repeated nine more times and the ten samples (each of sample size  $N = 10$ ) gave the results shown in Table 38.

TABLE 38. SUMMARY TABLE FOR TEN SAMPLES WITH  $N = 10$

$t$	Total Number of Failures	Average Number of Failures, $\phi(t)$
10	1	0.1
20	15	1.5
30	51	5.1
40	96	9.6

From Table 38 and eq. (114), one can determine and compare the cost of group replacement for the 10-, 20-, 30-, and 40-day periods. These total expected costs over time period  $T$  (360 days) are:

$$K_{10} = \frac{360}{10} [(10)(0.50) + (0.1)(1.00)] = \$183.60,$$

$$K_{20} = \frac{3.60}{20}[(10)(0.50) + (1.5)(1.00)] = \$117.00,$$

$$K_{30} = \frac{3.60}{30}[(10)(0.50) + (5.1)(1.00)] = \$121.20,$$

$$K_{40} = \frac{3.60}{40}[(10)(0.50) + (9.6)(1.00)] = \$131.40.$$

Thus, *if* one is to group replace at 10-, 20-, 30-, or 40-day intervals, one would do so every 20 days. (This assumes, of course, that, in practice, one has taken a sufficiently large sample of random normal numbers.)

If one did not group replace, one could expect to replace each bulb, on the average, every 30 days. Accordingly, the total expected cost over time period  $T$  of *not* group replacing, call it  $K_\infty$ , is

$$K_\infty = 10\left(\frac{3.60}{30}\right)(1.00) = \$120.$$

Therefore, under the assumptions of this illustration, one *should* group replace every 20 days (since  $K_\infty > K_{20}$ ).

### Other Models

Although the solutions presented apply only to the particular model described earlier, models of other characteristics may be approached in the same way. For *example*, a model could be concerned with group replacement in which new bulbs are used for group replacement only, and used bulbs replace failures in between group replacements. A different model is needed when surviving bulbs are replaced at a fixed age, rather than at fixed intervals of time. The considerations of this chapter have been limited to demonstrating an approach to two basic replacement problems, one involving deterioration, and the other involving probabilistic life spans of equipment.

For a discussion of the models which have been developed and solutions obtained for various sets of assumptions about the conditions of the problem, see Ref. 1, Chap. 17. For a useful review of equipment replacement rules from an industrial point of view, see Ref. 66.

## 7. COMPETITIVE PROBLEMS

### Introduction

A *competitive problem* is one in which the efficiency of one's decision is affected by the decisions of one's competitors. Such problems include, for example, competitive advertising for a relatively fixed market or bidding for a given set of contracts.

**Game Problems.** The most publicized competitive problem in O.R. is the "game" as developed by the late John von Neumann and discussed in his *Zur Theorie der Gesellschaftsspiele* (Ref. 67) in 1928 and, jointly with

Oskar Morgenstern, in their *Theory of Games and Economic Behavior* (Ref. 68) in 1944.

For many decades, economists tended to take as their standard model for their science, the situation of Robinson Crusoe, marooned on an uninhabited island and concerned with behaving in such a manner as to maximize the goods he could obtain from nature. It was generally felt that it would be possible to gain insight into the behavior of groups of individuals by starting with a detailed analysis of the behavior in this simplest possible case: the case of a single individual all alone and struggling against nature.

This line of attack on economic problems, however, suffers from the defect that in going from a one-man society to even a two-man society, qualitatively different situations arise which could not have been foreseen from the one-man case. Von Neumann was led to believe that group economics could more profitably be viewed as analogous to parlor *games of strategy*.

*Von Neumann's game* is characterized by a fixed set of rules and a known number of competitors whose possible choices are also known. Furthermore, the payoff for each combination of choices is also assumed to be known. The solution to von Neumann's game is obtained by a principle of conservatism called the *minimax principle*, namely one which will *maximize the minimum expected gain or minimize the maximum expected loss*.

Very little has been accomplished by way of applying the von Neumann theory of games. Military applications have been referred to but have not been made public. Several authors have explored the possibility of applying game theory to industrial problems, but they have not dealt with actual applications. What then is the significance and value of game theory? This can best be answered by quoting Williams' (Ref. 69, p. 217) succinct appraisal:

While there are specific applications today, despite the current limitations of the theory, perhaps its greatest contribution so far has been an intangible one; the general orientation given to people who are faced with over-complex problems. Even though these problems are not strictly solvable—certainly at the moment and probably for the indefinite future—it helps to have a framework in which to work on them. The concept of a strategy, the distinction among players, the role of chance events, the notion of matrix representations of the payoffs, the concepts of pure and mixed strategies, and so on, give valuable orientation to persons who must think about complicated conflict situations.\*

**Bidding Problems.** A second type of competitive problem is one in which bidding takes place. Bidding problems differ from game problems in that: (a) the number of competitors is not usually known, (b) the number of choices is not known (since one can bid over a large range), and (c) the payoffs are not usually known but, rather, are subject to estimation

\* Reprinted by permission from *The Compleat Strategyst* by J. D. Williams, copyright 1954. McGraw-Hill Book Co.

(e.g., in bidding for mineral rights). Furthermore, in some bidding situations (e.g., those in which one bids a dollar amount *plus* a percentage of the royalties), one may not be able to determine readily whether or not a given bid would have won or lost.

Only a limited theory of bidding exists to date, although the concepts and techniques of statistical decision theory hold great promise in this area. A major research contribution has been made by Friedman (Ref. 70 and Ref. 1, Chap. 19). The number of applications of bidding theory has been very limited; however, in at least one instance, the results obtained have been spectacularly successful. Bidding models will not be discussed here. See Refs. 1 and 70.

### The Theory of Games

#### Definitions.

*Game*, a set of rules and conventions for playing.

*Play*, a particular possible realization of the rules.

*Move*, a point in a game at which one of the players selects an alternative from some set of alternatives.

*Choice*, that particular alternative selected.

*Strategy*, a player's predetermined method for making his choices during the play.

#### Classification of Games.

1. *Players*, the number of sets of opposing interests: (a) one-person, (b) two-person, (3)  $n$ -person ( $n > 2$ ).

2. *Payment*, (a) zero-sum game, a game in which the sum of the payoffs, counting winnings as positive and losses as negative, to all players is zero; (b) nonzero-sum game, a game in which the sum of the payoffs to all players is not zero.

3. *Number of moves*: (a) finite, (b) infinite.

4. *Number of choices*: (a) finite, (b) infinite.

5. *Amount of information regarding opponent's choices*: (a) all, (b) part, (c) none.

**One-Person Games.** One-person zero-sum games are trivial games which say "do nothing" since there is no gain to be made by the one participant in the game. One-person nonzero-sum games are the ordinary maximization and minimization problems solvable by calculus and other optimization techniques. Thus, in order to study the characteristic properties of games of strategy, it is necessary to go to games which involve more than one player. The discussion here will center mainly on two-person zero-sum games.

**Two-Person, Zero-Sum Games.** Analysis of the very simplest of games shows that there are two general kinds, which may be illustrated by two kinds of coin matching.

*Single Strategy Games.* Assume that one is matching dimes and quarters where, if both coins are the same, it is a standoff, but if the coins differ, the quarter takes the dime. In this game it is safest always to play a quarter, for then one can never lose, whereas one may lose by playing a dime. Such games in which each opponent will find it safest to stick to one strategy are called single strategy games.

*Mixed Strategy Games.* The second general type of game may be illustrated by the usual penny-matching situation in which each player chooses either heads or tails. If the coins match, the matcher wins; if they do not match, the matchee wins. In this case, if either player sticks to one strategy, he may consistently lose. The only safe way to play the game is to play heads or tails in a completely random manner, as, for instance, by flipping the penny in the air just before one plays it. Such games are called games of mixed strategy.

*Payoff Matrix.* Games can have any number of strategies. In principle, once each player has chosen one of the sets of strategies available to him, it is possible to calculate the probable outcome of the game. The net payoffs can then be arranged in a two-dimensional matrix, the payoff matrix. From the payoff matrix, one can then find whether the game is a single or a mixed strategy game and, if mixed, in what proportions to mix the playing.

For further discussion of the definitions, classification of games, and examples of the construction of payoff matrices, see Ref. 71, Chap. 1.

### Single Strategy, Two-Person, Zero-Sum Games

**Minimax Principle.** Consider the game whose payoff matrix with respect to player  $P_1$  is

$$A = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{vmatrix}$$

If player  $P_1$  chooses the number  $i$  (i.e., adopts the  $i$ th strategy,  $i = 1, 2, \dots, m$ ), he is certain to receive *at least*

$$\min_j a_{ij}, \quad j = 1, 2, \dots, n.$$

Since he can choose  $i$  as he pleases, he can, in particular, choose  $i$  so as to make  $\min_j a_{ij}$  as large as possible. Thus player  $P_1$  can choose  $i$  so as to receive *at least*

$$\max_i \min_j a_{ij}.$$

Similarly, player  $P_2$  can choose  $j$  so as to make certain that he will receive *at least*

$$\max_j \min_i (-a_{ij}),$$

since for two-person, zero-sum games, the payoff matrix with respect to player  $P_2$  will consist of elements (payments) which are the negative of the elements of matrix  $A$ . That is, player  $P_2$  can choose  $j$  so as to make certain that he will receive *at least*

$$-\min_j \max_i a_{ij}$$

or, equivalently, that player  $P_1$  will get *at most*

$$\min_j \max_i a_{ij}.$$

*Saddle Point.* In summary,  $P_1$  can guarantee that he will receive at least

$$\max_i \min_j a_{ij}$$

and  $P_2$  can prevent  $P_1$  from receiving more than

$$\min_j \max_i a_{ij}.$$

If

$$\max_i \min_j a_{ij} = \min_i \max_j a_{ij} = a_{i_0 j_0} = v,$$

$P_1$  will settle for  $v$  and  $P_2$  will settle for  $-v$ . Games for which the equation above holds are called *games with a saddle point*. More specifically,  $(i_0, j_0)$  is called a *saddle point* and  $a_{i_0 j_0}$  is called the *value of the game* for player  $P_1$ . Furthermore, the best strategy for player  $P_1$  is  $i_0$ , and the best strategy for player  $P_2$  is  $j_0$ . (See Ref. 71, Chap. 1.)

It should be noted that a saddle point of a matrix is a pair of integers  $(i_0, j_0)$  such that  $a_{i_0 j_0}$  is at the same time the minimum element of its row and the maximum element of its column.

*Every single strategy two-person, zero-sum game has a saddle point.* This saddle point provides the solution of the game by designating the best strategies for each player and the value of the game. *Example.* The game represented by

$$\begin{vmatrix} 4 & 3 & 7 & 6 \\ 5 & 2 & 1 & 0 \\ 0 & 1 & 3 & 4 \\ 2 & 2 & 1 & 5 \end{vmatrix}$$

has a saddle point at  $(1, 2)$ . Its solution consists of the strategies 1 and 2, respectively, and the value of the game is 3.

Stated in another manner, every game which contains a saddle point is a single strategy game (see Ref. 71). Games without saddle points, such as the game represented by

$$\begin{vmatrix} 1 & -1 \\ -1 & 1 \end{vmatrix}$$

are mixed strategy games.

**Mixed Strategy, Two-Person, Zero-Sum Games**

Consider the (penny-matching) game whose payoff matrix is

		$P_2$	
		Heads	Tails
$P_1$	Heads	+1	-1
	Tails	-1	+1

Such a matrix has no saddle point and, hence, is not a single strategy game. Furthermore, one can readily see that it makes little difference to player  $P_1$  whether he chooses strategy 1 (heads) or strategy 2 (tails), for, in either case, he will receive 1 or  $-1$  according as  $P_2$  makes the same or opposite choice. Player  $P_1$  must play the game by making his selections by means of some chance device. The procedures for determining optimum mixed strategies are discussed below.

*Dominance.* If  $a = (a_1, a_2, \dots, a_n)$  and  $b = (b_1, b_2, \dots, b_n)$  are vectors (or rows or columns of a matrix), and if  $a_i \geq b_i$  (for  $i = 1, 2, \dots, n$ ), one says that  $a$  dominates  $b$ . If  $a_i > b_i$  (for  $i = 1, 2, \dots, n$ ), one says that  $a$  strictly dominates  $b$ .

*Convex Linear Combination.*

Let  $x^{(1)} = (x_1^{(1)}, \dots, x_n^{(1)})$ ,

$x^{(2)} = (x_1^{(2)}, \dots, x_n^{(2)})$ ,

.....

$x^{(r)} = (x_1^{(r)}, \dots, x_n^{(r)})$ ,

$x = (x_1, \dots, x_n)$ .

Let

$$a = (a_1, \dots, a_r)$$

such that  $a_i \geq 0$  ( $i = 1, 2, \dots, r$ ) and  $a_1 + a_2 + \dots + a_r = 1$ . Then  $x$  is a convex linear combination of  $x^{(1)}, \dots, x^{(r)}$  with weights  $a_1, \dots, a_r$  if

$$x_j = a_1x_j^{(1)} + a_2x_j^{(2)} + \dots + a_rx_j^{(r)}, \text{ for } j = 1, 2, \dots, n.$$

Thus, the point  $(0, 15)$  is a convex linear combination (with weights  $\frac{1}{6}$ ,  $\frac{1}{3}$ , and  $\frac{1}{2}$ ) of the points  $(6, 12)$ ,  $(-9, 15)$ , and  $(4, 16)$ .

**THEOREM.** *Let  $\Gamma$  be a rectangular game whose matrix is  $A$ ; suppose that, for some  $i$ , the  $i$ -th row of  $A$  is dominated by some convex linear combination of the other rows of  $A$ ; let  $A'$  be the matrix obtained from  $A$  by omitting the  $i$ -th row; and let  $\Gamma'$  be the rectangular game whose matrix is  $A'$ . Then the value of  $\Gamma'$  is the same as the value of  $\Gamma$ ; every optimum strategy for  $P_2$  in  $\Gamma'$  is also an optimum strategy for  $P_2$  in  $\Gamma$ ; and if  $w$  is any optimum strategy for  $P_1$  in  $\Gamma'$  and  $x$  is the  $i$ -place extension of  $w$ , then  $x$  is an optimum strategy for  $P_1$  in  $\Gamma$ . Moreover, if the  $i$ -th row of  $A$  is strictly dominated by the convex linear combination of the other rows of  $A$ , then every solution of  $\Gamma$  can be obtained in this way from a solution of  $\Gamma'$ . (See Ref. 71, Chap. 2.)*

*Note.* A similar theorem applies to dominating columns. (See Ref. 71, Chap. 2.)

**EXAMPLE.** The following example of the application of this theorem is cited in Ref. 71 (p. 50):

$$\begin{vmatrix} 3 & 2 & 4 & 0 \\ 3 & 4 & 2 & 4 \\ 4 & 2 & 4 & 0 \\ 0 & 4 & 0 & 8 \end{vmatrix}$$

Row 1 is dominated by row 3, yielding

$$\begin{vmatrix} 3 & 4 & 2 & 4 \\ 4 & 2 & 4 & 0 \\ 0 & 4 & 0 & 8 \end{vmatrix}$$

Column 1 dominates column 3, resulting in

$$\begin{vmatrix} 4 & 2 & 4 \\ 2 & 4 & 0 \\ 4 & 0 & 8 \end{vmatrix}$$

Column 1 dominates a convex linear combination of columns 2 and 3, namely:

$$4 > \frac{1}{2}(2) + \frac{1}{2}(4),$$

$$2 = \frac{1}{2}(4) + \frac{1}{2}(0),$$

$$4 = \frac{1}{2}(0) + \frac{1}{2}(8).$$

Thus, the first column can be omitted, yielding

$$\begin{vmatrix} 2 & 4 \\ 4 & 0 \\ 0 & 8 \end{vmatrix}$$

Row 1 is now dominated by a convex linear combination of rows 2 and 3, since

$$2 = \frac{1}{2}(4) + \frac{1}{2}(0),$$

$$4 = \frac{1}{2}(0) + \frac{1}{2}(8).$$

Therefore, the matrix reduces to

$$\begin{vmatrix} 4 & 0 \\ 0 & 8 \end{vmatrix}$$

As will be seen later, the solution to this latter matrix consists of the mixed strategy  $(\frac{2}{3}, \frac{1}{3})$  for each player and a game value of  $\frac{8}{3}$ . Therefore, the value of the original game is  $\frac{8}{3}$ , and the optimum strategy for the original game is  $(0, 0, \frac{2}{3}, \frac{1}{3})$  for each player.

### General Theorems for Rectangular Games (Refs. 1 and 71)

**THEOREM 1.** *Every rectangular game has a specific value  $g$ . This value is unique. Furthermore, there exists for player  $P_1$  a best strategy, i.e., there exist non-negative frequencies  $x_1, x_2, \dots, x_m$  such that  $x_1 + x_2 + \dots + x_m = 1$  and such that if he plays plan I with frequency  $x_1$ , plan II with frequency  $x_2, \dots$ , plan M with frequency  $x_m$ , then he can assure himself at least an expected gain of  $g$ , which is the value of the game.*

*Similarly, for player  $P_2$ , there exists a best strategy  $Y = (y_1, y_2, \dots, y_n)$ ,  $y_1 + y_2 + \dots + y_n = 1$ , such that if  $P_2$  played plans I, II,  $\dots$ , N with the above frequencies, respectively, he ( $P_2$ ) can assure himself at most a loss of  $g$ .*

**THEOREM 2.** *The unknowns,  $x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_n$  and  $g$  (for the solution of a game) can be determined from the following relations:*

$$x_1 + x_2 + \dots + x_m \equiv \sum_{i=1}^m x_i = 1, \quad x_i \geq 0 \quad (i = 1, 2, \dots, m);$$

$$y_1 + y_2 + \dots + y_n \equiv \sum_{j=1}^n y_j = 1, \quad y_j \geq 0 \quad (j = 1, 2, \dots, n);$$

$$\sum_{i=1}^m x_i a_{ij} \geq g \quad (j = 1, 2, \dots, n);$$

$$\sum_{j=1}^n a_{ij} y_j \leq g \quad (i = 1, 2, \dots, m).$$

**THEOREM 3.** Let  $X^* = (x_1^*, x_2^*, \dots, x_n^*)$  and  $Y^* = (y_1^*, y_2^*, \dots, y_n^*)$  be any optimal strategies for  $P_1$  and  $P_2$ , respectively, for a game whose value is  $g$ . If, for any  $i$ ,

$$\sum_{j=1}^n a_{ij}y_j < g,$$

then

$$x_i^* \equiv 0.$$

Similarly, if for any  $j$ ,

$$\sum_{i=1}^m x_i a_{ij} > g,$$

then

$$y_j^* \equiv 0.$$

**Solutions of Rectangular Games**

**Two-by-Two Games.** To solve two-by-two rectangular games, first look for a saddle point. If one exists, the game is a single strategy game and the solution is immediately given as discussed above. If no saddle point exists, the game is a mixed strategy game and is solvable by either of the following methods.

**Algebraic Solution.**

Given:

$$A = \begin{array}{|c|c|} \hline a & b \\ \hline c & d \\ \hline \end{array}$$

Let  $x$  and  $1 - x$  be the frequencies with which  $P_1$  plays plans I and II respectively. Then, if player  $P_2$  plays plan I,  $P_1$  can expect

$$a(x) + c(1 - x) = c + (a - c)x.$$

On the other hand, if player  $P_2$  plays plan II,  $P_1$  can expect

$$b(x) + d(1 - x) = d + (b - d)x.$$

The solution of any two-by-two game is given by the minimax principle, namely, by solving

$$c + (a - c)x = d + (b - d)x.$$

**EXAMPLE.** Given:

-3	7
6	1

Then  
yields

$$-3(x) + 6(1 - x) = 7(x) + 1(1 - x)$$

$$x = \frac{1}{3},$$

$$(1 - x) = \frac{2}{3}.$$

Similarly, one determines that

$$y = \frac{2}{5},$$

and

$$(1 - y) = \frac{3}{5}.$$

Finally, the value of the game is given by (since  $x = \frac{1}{3}$ )

$$g = -3(\frac{1}{3}) + 6(\frac{2}{3}) = 3.$$

For this and other algebraic procedures, see Ref. 1, Chap. 18.

**Method of Oddments** (*Two-by-Two Game*).

The method of oddments for two-by-two games is given by Williams (Ref. 69).

EXAMPLE. The method may be stated by means of the game whose payoff matrix is

		$P_2$	
		I	II
$P_1$	I	-3	7
	II	6	1

To determine the optimum frequencies for  $P_1$ , subtract the numbers in the second column from those in the first column. This gives:

-10
5

One of the two numbers will always be negative. Ignore the minus sign for the purpose of computing oddments.

Then, the oddment for  $P_1(I)$  is given by

I	5
---	---

whereas the oddment for  $P_1(II)$  is given by

II	10
----	----

Therefore, the oddments for  $P_1$  are 5 and 10, respectively, or, equivalently, the optimum frequencies are

$$\frac{5}{5 + 10} = \frac{1}{3} \quad \text{and} \quad \frac{10}{5 + 10} = \frac{2}{3}$$

Similarly, by subtracting rows, one can determine that the optimum frequencies for player  $P_2$  are  $\frac{2}{5}$  and  $\frac{3}{5}$ .

**Two-by- $n$  Games.** To find the solution of a two-by- $n$  game:

1. Look for a saddle point. If one exists, the game is a single strategy game and the solution is given by the saddle point.
2. If no saddle point exists, examine the payoff matrix for dominance and, eliminate all *dominated* strategies (if any) for  $P_1$  and all *dominant* strategies (if any) for  $P_2$ .
3. The matrix which remains will then contain a two-by-two submatrix with the property that its solution is also a solution to the two-by- $n$  game. The pertinent two-by-two submatrix can be found in one of several ways, probably the easiest of which is the *graphical method*.

*Graphical Solution of Two-by- $n$  Games.*

Given the game whose payoff matrix is:

		( $P_2$ )						
		1	2	3	4	5	6	7
( $P_1$ )	1	-6	1	3	5	0	-4	-1
	2	7	3	-2	4	-3	0	1

Plot the payoffs for each strategy of  $P_2$  on two parallel axes, as shown in Fig. 13.

Then, join the line segments which bound the figure from *below* and mark the *highest point* on this boundary. The lines which intersect at this point identify the strategies that player  $P_2$  should use.

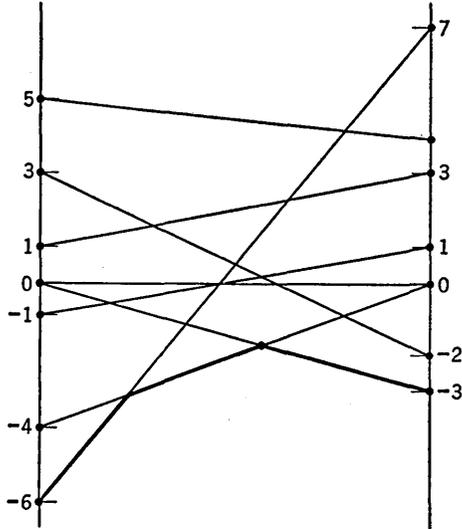


FIG. 13. Graphical solution of two-by- $n$  game.

In the given example, these are strategies 5 and 6. (Note that strategies 2, 4, and 7 dominate strategy 6 and could have been eliminated immediately. Similarly strategy 3 dominates 5 and could be eliminated.) Therefore, the appropriate two-by-two subgame is

		( $P_2$ )	
		5	6
( $P_1$ )	1	0	-4
	2	-3	0

which, by the method of oddments, gives

$$x^* = \left(\frac{3}{7}, \frac{4}{7}\right) \quad \text{and} \quad y^* = \left(\frac{4}{7}, \frac{3}{7}\right)$$

and a game value of  $g = -1\frac{2}{7}$ .

Hence, the solution to the original game is

$$X^* = (\frac{3}{7}, \frac{4}{7}) \quad \text{and} \quad Y^* = (0, 0, 0, 0, \frac{4}{7}, \frac{3}{7}, 0),$$

and, again,  $g = -1\frac{2}{7}$ .

It should be noted that, for  $m$ -by-two games, one proceeds as above, marking, however, the line segments which bounds the graph from above and then identifying the lowest point on this boundary. *Note.* This is merely a graphical application of the minimax principle. See Refs. 1, 69, and 71.

**Three-by-Three Games.** To find the solution of three-by-three games:

1. Look for saddle point.
2. If none exists, examine the payoff matrix for dominance and reduce it accordingly.
3. If a three-by-three matrix remains, solve by method of oddments to see if a three-by-three solution exists.
4. If oddments method fails, try the two-by-two subgames for a solution.

**EXAMPLE.** *Method of Oddments (Three-by-Three Game).* Consider the game whose payoff matrix is

( $P_2$ )

		1	2	3
1	6	0	6	
2	8	-2	0	
3	4	6	5	

( $P_1$ )

To determine the optimum frequencies for player  $P_1$ , subtract each column from the preceding column, yielding

1	6	-6
2	10	-2
3	-2	1

( $P_1$ )

The oddment for  $P_1(1)$  is given by

1		
	10	-2
	-2	1

the numerical value of which is the difference between the diagonal products:

$$10(1) - (-2)(-2) = 6.$$

Similarly,  $P_1(2)$  is given by

	6	-6
2		
	-2	1

namely,

$$6(1) - (-6)(-2) = -6,$$

and  $P_1(3)$  is given by

	6	-6
	10	-2
3		

or

$$(6)(-2) - (-6)(10) = 48.$$

Therefore, the oddments for  $P_1$  are

$$6:6:48$$

so that the optimum frequencies are

$$X^* = \left(\frac{1}{10}, \frac{1}{10}, \frac{4}{5}\right).$$

Similarly, by subtracting rows, one determines the oddments for  $P_2$ , namely,

$$38:14:8$$

and optimum frequencies

$$Y^* = \left(\frac{19}{30}, \frac{7}{30}, \frac{4}{30}\right).$$

Furthermore, the value of the game is given by

$$g = \frac{(1)(6) + 1(8) + 8(4)}{10} = \frac{23}{5}.$$

*Note.* Every solution obtained by the method of oddments must be tested. It may well be that the three-by-three game does not have a three-by-three solution, but, rather, a two-by-two solution. See Ref. 69.

**Three-by- $n$  Games.** To find the solution of three-by- $n$  games:

1. Look for a saddle point.
2. If none exists, examine the payoff matrix for dominance and reduce it accordingly.
3. If a three-by- $n$  matrix remains, the problem is then to find the solution by the earlier methods, since every three-by- $n$  matrix has solutions which are either three-by-three or two-by-two (or a saddle point).

Solve the two-by- $n$  subgames by the graphical method. If no two-by-two solutions exist, the solution must then be a three-by-three solution which can be obtained by successively trying each three-by-three subgame.

See Refs. 1, 69, and 71.

**Four-by-Four Games.** For games which do not have a saddle point (i.e., mixed strategy games) and which, after removing rows and/or columns due to dominance, reduce to a four-by-four game, there is a method of oddments for obtaining the desired solution. For this method, see Williams (Ref. 69).

### Other Solutions of Rectangular Games

There are a variety of other methods for solving rectangular games, a few of which are cited here.

#### Matrix Solution of Games.

Let  $A = (a_{ij})$  be the  $m \times n$  matrix of a game,

$B = (b_{ij})$  be any square submatrix of  $A$  of order  $r > 1$ ,

$J_r = (1, 1, \dots, 1)$ , a  $1 \times r$  matrix,

$C^T$  = transpose of  $C$ , where  $C$  is any matrix,

adj  $B$  = adjoint of  $B$ ,

$X = (x_1, x_2, \dots, x_m)$ ,  $x_i \geq 0$ ,  $\sum x_i = 1$ ,

$Y = (y_1, y_2, \dots, y_n)$ ,  $y_j \geq 0$ ,  $\sum y_j = 1$ ,

$\bar{X}$  = a  $1 \times r$  matrix obtained from  $X$  by deleting those elements corresponding to the rows deleted from  $A$  to obtain  $B$ ,  
 $\bar{Y}$  = a  $1 \times r$  matrix obtained from  $Y$  by deleting those elements corresponding to the columns deleted from  $A$  to obtain  $B$ .

*Solution.*

1. Choose a square submatrix  $B$  of  $A$  of order  $r$  ( $\geq 2$ ) and calculate

$$\bar{X} = \frac{J_r \text{adj } B}{J_r(\text{adj } B)J_r^T} = (x_1, x_2, \dots, x_r),$$

and

$$\bar{Y} = \frac{J_r(\text{adj } B)^T}{J_r(\text{adj } B)J_r^T} = (y_1, y_2, \dots, y_r).$$

2. If some  $x_i < 0$  or some  $y_j < 0$ , reject the chosen  $B$  and try another.
3. If  $x_i \geq 0$  and  $y_j \geq 0$  for all  $i, j = 1, 2, \dots, r$ , calculate

$$g = \frac{|B|}{J_r(\text{adj } B)J_r^T}$$

and construct  $X$  and  $Y$  from  $\bar{X}$  and  $\bar{Y}$  by adding zeros in the appropriate places.

Check whether

$$\sum_{i=1}^m x_i a_{ij} \geq g, \quad \text{for all } j,$$

and whether

$$\sum_{j=1}^n y_j a_{ij} \leq g, \quad \text{for all } i.$$

If one of the relations does not hold, try another  $B$ . If *all* relations hold, then  $X$ ,  $Y$ , and  $g$  are the required solutions. See Refs. 1 and 71.

**Iterative Method for Solving a Game.** There is an approximate method of solving rectangular games which enables one to find the value of such games to any desired degree of accuracy and also to approximate to optimal strategies. See Ref. 71, Chap. 4, and Ref. 1, Chap. 18.

**Solution of Rectangular Games by Linear Programming.** It can be shown that the problem of solving an arbitrary rectangular game can be regarded as a special linear programming problem and, conversely, that many linear programming problems can be reduced to problems in game theory. Thus, *the techniques for solving linear programming problems (e.g., the simplex technique), especially through the use of high-speed electronic computers, can be applied to the solution of game theory problems.* See Ref. 71, Chap. 14.

### Zero-Sum, $n$ -Person Games

The theory of  $n$ -person games,  $n > 2$ , is not in an altogether satisfactory state. For an excellent exposition on the elements of zero-sum  $n$ -person games, see Ref. 71 and the original text on the subject, namely that of von Neumann and Morgenstern (Ref. 68). A very brief discussion is to be found in Ref. 1.

## 8. DATA FOR MODEL TESTING

**Introduction.** The type of evidence one uses to test a model depends very much on the kind of test one has in mind. In testing a model one asks, "What are the possible ways in which a model can fail to represent reality adequately and hence lose some of its potential usefulness?" Following are four ways in which one may question the adequacy of a model.

1. The model may assert a dependence of the effectiveness of the system (the dependent variable) on one or more (independent) variables which, as a matter of fact, do *not* affect the system's effectiveness. That is, the model may fail by including variables which are not pertinent.

2. The model may *fail to include* a variable which does have a significant effect on the system's effectiveness.

3. The model may inaccurately express the actual relationship which exists between the measure of effectiveness and one or more of the *pertinent* independent variables.

4. Finally, even if the model is an accurate picture of reality in the sense of conforming to the foregoing three conditions, it may still fail to yield good results if the parameters contained in it are not evaluated properly.

In testing the model, begin by testing it as a whole, i.e., by determining the accuracy of its prospective or retrospective predictions of the system's effectiveness. If this procedure shows that the model is not adequate, further testing will be required to find out which of the four types of deficiencies mentioned here is present.

The design of the process of collecting data consists of the following parts: (1) definition (including measurement), (2) sampling (including experimental designs), (3) data reduction, (4) use of the data in the test, (5) examination of the result, and (6) possible redesign of the evidence.

**Scientific Definitions.** Scientific defining consists of specifying the best *conceivable* (not necessarily obtainable) conditions under which, and procedures by which, values of the variables can be obtained.

Concern with *ideal* (or *optimum*) observational conditions and procedures is quite important if one wants to know how good are the results one eventually obtains. Further, and more important, the ideal conditions and procedures act as a *standard* by means of which one can evaluate the attainable

observational conditions and operations, determine their shortcomings, and make any necessary adjustments in the resultant data. For a detailed discussion of scientific defining, see Ref. 72.

The two most common types of quantitative variables are the *enumerative* and the *metric*. The enumerative variable requires counting for its evaluation whereas the metric variable requires measurement.

**Scientific Definitions of Enumerative Variables.** Two types of errors can arise in the counting operation, *overenumeration* and *underenumeration*. Overenumeration results either from counting the same unit more than once or from counting units which should not be counted at all. Underenumeration, on the other hand, results from the failure to count a unit which should be counted. Furthermore, these errors can occur because of a failure to match elements with *consecutive* integers (e.g., overenumeration because of skipping numbers and underenumeration through duplication of numbers).

It is desirable to design the best conceivable counting procedure, even if the design cannot be carried out in practice. This involves specifying the standard environment in which, and the standard operations by which, the count can ideally be made, as well as providing an explicit definition of the elements to be counted. Once this standard is specified, it will be possible to use it to evaluate alternative practically realizable counting procedures and to select the best of these. The standard also provides a basis for estimating the error that is likely to occur in the practical counting procedure which is eventually used.

**Scientific Definitions of Properties (Metric Variables).** The idealized design of a procedure for measuring properties depends primarily on the type of property involved. Scientific definitions of properties involve specifying the following characteristics of the idealized measuring procedure:

1. Identification of the thing, event, or class of things or events which should be observed.
2. Specification of the environment in which the observations should be made.
3. Specification of the changes in the environment which should be made, if any, during the observation period.
4. Specification of the operations to be performed and the instruments and measure to be used by the observer.
5. Specification of the readings (data) to be made.
6. Specification of the analysis of the data.

The formal description of the measure to be used states what logical and mathematical operations one wants to be able to perform on the data to be obtained in evaluating a variable. For a complete discussion of the theory

of measurement see Refs. 73-79. The scientific definition (observational standard) states how, ideally, one would go about collecting pertinent data. The operational specification of the data collection process states how one actually intends to collect and adjust the data. Errors can arise in each of these three stages of planning relative to testing the model.

**Sampling.** In evaluating variables, one is either involved in measuring the property of a single unit or in counting the members of, or measuring the properties of, a class of units (a population). The definition of a property of a single unit specifies the conditions under which the observation should be made. If these conditions can be met and observations can be made without error, only one observation is required. But if the conditions are not met, observations are subject to error which can only be estimated if two or more observations are made. How many observations to make, and where, are sampling questions. Since the standard conditions specified in the definition can seldom be met in practice, one must choose one of two courses: (1) An experimental design must be chosen which, by techniques such as the analyses of variance and covariance, makes it possible to assess the magnitude of the deviations and ascribe them to specific environmental factors, or (2) observations must be made on a subset (of the population) which make it possible to draw inferences that are valid for the whole population with the least possible bias. The subject of sampling is concerned with the selection of appropriate subsets.

In the main, *sampling* can be described as the selection of items from a population. The "population" of objects, events, environments, and stimuli to be sampled should be specified in the definition of the variable being evaluated. The population represents all the possible data of the relevant kind that can be collected.

**Evaluation of Samples and Sample Estimates.** The decision which must be made in designing a sampling procedure is concerned with the method of drawing the sample and the method of making estimates about the population from the sample. If a prescribed method is carried out correctly, there are two opposing considerations: (1) the probability that the estimate made on the basis of the sample will actually deviate from the true population value by an amount greater than some amount  $x$ ; (2) the cost of taking the sample.

In the main, the probability of deviations will decrease with an increase in the sample size, but the cost of taking the sample will increase with an increase in sample size.

**Types of Sampling Designs.** In *unrestricted random sampling* every possible sample has the same chance of being chosen. *Restricted random sampling* represents methods by which each possible sample does not have an equal probability of being drawn. But in each case where random

sampling is used scientifically, the probability of selecting any sample is known.

All the various schemes for sampling are based on very simple, practical considerations. These are:

1. Items of the population may fall into recognizable groups (e.g., in terms of location or dollar amounts on an invoice). If this is the case, it is reasonable to think in terms of sampling from these groups, because in general one reduces the variance of the estimates and (more important) one can be selective in the amount of sampling that is done in each group. Invoices with large dollar amounts are more important than ones with small dollar amounts; hence a larger sample of the more important items should be taken.

2. Items of a population often fall into clusters (e.g., a shipment shown on an invoice; people in a house, block, or town; items in a warehouse). If one looks at some item in a cluster, one might just as well look at the rest of the items. Hence the cluster becomes the basis of sampling, not the original items. The use of clusters may increase the variance of the estimates but greatly decrease the costs of gathering the sample—the usual economic balancing problem.

3. One does not have to plan completely in advance. One can let the sample information that comes in dictate how the next steps are to be taken.

The following is a general classification of the principal types of sampling designs:

I. *Fixed sampling design.* The sampling design is fixed and not subject to change in terms of sample data.

A. *Unrestricted random sampling.* A random sample is selected from the whole population by either

1. *Simple random sampling.* Assigning a different number to each element in the population and using random numbers to select the sample, or

2. *Systematic random sampling.* Where a population is ordered, selecting a starting place at random and then selecting subsequent elements at a fixed interval from the first and subsequent selections.

Tables of random numbers can be found in Refs. 10 and 80–82.

Details on the generation of such numbers can be found in Ref. 83.

B. *Restricted random sampling.* The population is divided into subgroups (and possible subsubgroups, etc.) and either some of these are selected and/or random samples from some or all of these are selected.

1. *Multistage random sampling.* Random samples are drawn from subgroups which have themselves been selected (*a*) with equal

probability, or (b) with probability proportionate to the relative size of the subgroup, or some other criterion.

2. *Stratified random sampling.* A random sample is drawn from every subgroup of the population. The size of the sample from the subgroups may be (a) independent of the size of the subgroups (i.e., samples of equal size), (b) proportionate to the relative size of the subgroup, or (c) proportionate to the relative size of the subgroup and the dispersion of the elements within it (optimum allocation).
  3. *Cluster sampling.* A random sample of subgroups is selected, all elements of which are included in the final sample.
  4. *Stratified cluster sampling.* A combination of B2 and B3, where more than two stages of sampling are involved.
- II. *Sequential sampling.* A small random sample is selected and analyzed, on the basis of which a decision is made as to whether or not to continue sampling and if so, how. The samples may be either
- A. In groups, as in *double* or *multiple sampling*, or
  - B. Single items taken one at a time.

For details on sequential sampling see Refs. 84–87.

The aspect of sampling called *experimental design* usually refers to a sampling plan based on the variables in the model which is to be tested. Instead of keeping everything fixed except one variable, it is possible to design data collection systems at *optimum* locations of some of the variables of the model. This sampling method assumes that the variables of the model can be manipulated in reality—or at least in a realistic model. For information on various types of experimental designs, see Refs. 88–91. For comprehensive surveys of contemporary sampling theory see Refs. 72 and 92–95.

**Reduction of Data.** The observations made on the sampled items or in a sample of situations provide the raw data on the basis of which variables are assigned values and hence, provide the basis for testing all or part of the model. In many cases the data require collation, editing, coding, punching, etc. Discussion of these phases of data processing can be found in Ref. 72, Chap. X.

In general, the ultimate form to which the data must be transformed to be useful in the testing process will be either an estimate of the value of a parameter or an inferential “statistic” which describes a relationship between two or more variables. For *example*, in testing a lot-size model the cost variables must first be evaluated in order to compute the total cost “predicted” by the model. Once these predictions are obtained they are compared with observed values in order to derive a “statistic” which can be used to determine whether or not the model predicts well.

For a discussion of data reduction and the problems of estimation and obtaining estimates of the variability of estimates, see Ref. 1, Chap. 20.

For statistical tests of the significance or nonsignificance of a variable, see Refs. 96-105. These tests may make it possible to determine such matters as: (a) whether a variable should or should not be included, (b) whether the form of an analytic function is linear or some other type, (c) whether the form of a probability function is normal or some other type, (d) whether the model has failed to include a variable that ought to have been included.

For further discussion of procedures for testing the adequacy of models and the solutions derived from them, see Ref. 74, Chap. 20.

## 9. CONTROLLING THE SOLUTION

**Introduction.** Many, if not most, O.R. projects deal with management decisions that are recurrent. Hence the solution must be used over and over again. But the systems which are dealt with in O.R. are seldom stable. Their structure is subject to change. Relationships between the variables, or system *parameters*, which define the system and the value of the parameters themselves are usually subject to change.

In such situations the relationships and parameters used in the decision rule must be adjusted for changes in the system as they occur. Costs may change, the distribution of demand may change in some or all of its characteristics, and the relationships between variables may change over time. Hence the values of the relationships and parameters should be periodically reevaluated and the assumptions involved in the model (from which the decision rule is derived) should be reexamined periodically. That is, the solution must be *controlled* lest it lose some of its effectiveness because of changes in the system.

Complete methodologies have not yet been developed for optimizing control procedures. Enough is known, however, to design procedures which are more likely to lead to success than either leaving this phase of the project to chance or relying on others (management or operating personnel) to take care of them. For a complete discussion, see Ref. 1, Chap. 21.

**Controlling the Solution.** The effectiveness of a solution in an O.R. problem may be reduced by changes in either values of the parameters of the system or the relationships between them, or both. A previously insignificant parameter may become significant, or, conversely, a previously significant parameter may become insignificant. Changes in the values and functional relations of the parameters which remain significant may also affect the effectiveness of the solution.

Not every change in a parameter or relationship is *significant*. In general terms, a change is significant if (1) adjustment of the solution for the

change results in an improvement in effectiveness and (2) the cost of making the adjustment and carrying it out does not offset the improvement in effectiveness.

Design of a control system, then, consists of three steps: (1) listing the variables, parameters, and relationships that either are included in the solution or should be if their values were to change; (2) development of a procedure for detecting significant changes in each of the parameters and relationships listed; (3) specification of action to be taken or adjustments to be made in the solution when a significant change occurs.

The last two steps are interrelated because determination of the significance of a change (step 2) depends on the cost of making the adjustment specified in step (3).

**Control of Parameters.** The first step in designing a control procedure involves listing all the variables and relationships which, if they were to change in value, might affect the effectiveness of the solution.

The parameters which are listed should be classified into two types:

1. Variables whose values during the period covered by a decision can be known in advance, such as the number of models in a line, the number of work days in an accounting period, and the price for which an item is to be sold. Control of such measures consists either (a) of establishing communication lines between those who know these values and those who use the decision rules or (b) of providing the latter with source material (such as a calendar in the case of the number of work days per accounting period).

2. Measures whose values cannot be known in advance, such as number of units sold, number of hours worked, and arrival rate of trucks. These values must be estimated in advance.

Essential to the control of any measure, is the determination of whether its true value or one or more of the characteristics of its estimate have changed. This determination consists of testing the hypothesis that no change has occurred in the variable or the characteristics of its estimate (which are themselves variables).

**Errors in Detecting Changes.** Determination of whether or not such a change has occurred is subject to two types of error: (I) asserting that a change has occurred when it has not and (II) asserting that a change has not occurred when it has. An understanding of the two types of error is essential to comprehend what is involved in controlling a variable. See Ref. 1, Chap. 21.

**Detection of and Adjustment for Significant Changes.** Ideally, in the design of a control system for a variable, six interdependent decisions should be made if possible. (In some situations there may be no choice with regard to one or more of these decisions.) The decisions are as follows:

1. The frequency of (i.e., period between) control checks.
2. The number of observations per control check, if more than one is possible.
3. The way items should be selected for observation (i.e., the sampling design), if more than one observation is specified.
4. The statistical testing procedure to be used to determine whether or not a value has changed.
5. The specific decision rule based on the test.
6. The action to be taken, if the test indicates that a parameter's value has changed.

*Costs.* Again, ideally, these decisions should be made in such a way as to minimize the sum of the following costs:

1. The cost of taking the observations.
2. The cost of performing the test.
3. The expected cost of a type I error (i.e., the cost of changing a value when it is not warranted).
4. The expected cost of type II errors (i.e., the cost of not changing a value when it is warranted).

Unfortunately, at the present time the six decisions listed cannot be made in such a way as to assure minimization of the sum of the four costs. The design of an optimizing procedure can be specified in general terms; i.e., a model can be constructed which expresses the total expected cost as an abstract (but not as a concrete) function of the six decisions. In addition, some of the expressions which would appear in the model cannot be evaluated. In most situations, for example, the expected costs associated with type I and type II errors cannot be determined. For further reading, see Refs. 106-109.

Further details on methods of controlling parameters are given in Refs. 110-118.

**Controlling Relationships.** Every probability distribution asserts a relationship between the probability of an event and the values of other variables. In the case of distributions which appear in the model and solution, as, for *example*, the distribution of demand, the parameters which define the distribution must be controlled (e.g., the mean and variance) as well as the form of the distribution (e.g., normal or Poisson). Both aspects of the distribution should be subjected to control.

There are no "standard" procedures for controlling the form of a distribution. Such control may be obtained by periodically testing the "goodness of fit" in the manner given by standard statistics texts. The frequency with which such tests should be conducted depends on the rate at which data are generated. The visual plotting of data, as they become available, can frequently indicate when a check should be made. Examina-

tion of these charts can provide clues to changes in the parameters of the distribution as well as to the form.

The control of relationships which do not take the form of probability distributions also involves control over the form of the function which relates the variables and the values of the variables.

Every O.R. project has unique characteristics which create unique control problems but which also offer challenging opportunities for the development of unusual control procedures. There is a good deal of room for scientific creativity in this phase of the research. For a full discussion and illustration of the development of control procedures, see Ref. 1, Chap. 21.

## 10. IMPLEMENTATION

**Concern of the O.R. Team.** Once the solution has been derived and tested, it is ready to be put to work. Conversion of the solution into operation should be of direct concern to the research team for two reasons:

1. No matter how much care has been taken in deriving a decision rule and testing it, shortcomings may still appear when it is put into operation or ways of improving the solution may become apparent. If adjustment of the decision rule to take care of unforeseen operating problems is left in the hands of those who do not understand how it was derived, the adjustment may seriously reduce its effectiveness. Operating personnel may, for *example*, see no harm in making what appears to them to be a slight change, but such a change may be critical.

2. Carrying out the solution may not be as obvious a procedure in the context of complex operations as it initially appears to be to the researchers. The solution must be translated into a procedure that is workable if its potential is to be fully realized. The procedure must be as accurate a translation of the solution as is practically feasible and only the researchers can minimize the loss in the solution's effectiveness that is incurred in this translation.

The *nature of the implementation problem* depends on whether the solution pertains to a one-time or repetitive decision. In the case of one-time decisions the problem is simpler but by no means disappears.

Translation of the solution into the operating procedure involves answering three questions and proceeding accordingly. The three questions are: (1) Who should do what? (2) When? (3) What information and facilities are required to do it? On the basis of the answers to these questions the operating procedure can be designed and any necessary training and transition can be planned and executed.

Implementation of a solution involves people taking action. These people must be identified, and the required action must be specified. The details cannot be enumerated without a thorough knowledge of the opera-

tions and the division of responsibility in the organization under study. The analysis of the organization provides much of the needed information and the rest should be provided by management and operating people working with the research team. This and the other phases of implementation require continuous cooperation and communication among management, operators, and researchers.

Each person who is given responsibility for initiating action in carrying out the solution or using the decision rule should be instructed as to when they are to take action. The tools required to do the job should be made available to those who need them, and these people should be trained in their use. The tools should not be too complex for the operating personnel to use. It may be necessary, for example, to convert even a simple equation into a nomograph or tables. In some cases the tools may require simplification even if such simplification results in a loss of some of the original solution's power. The solution or decision rule is generally used by personnel whose mathematical sophistication is less than desired. Consequently, if one wants to assure use of the recommended decision rules, one must frequently simplify them before handing them over to executives and operating personnel. In many cases this means that one must either translate elegant solutions into approximations that are easy to use or sidestep the elegance and move directly to a "quick-and-dirty" decision rule.

It should be realized that in one sense almost every solution in O.R. is an approximation and is "quick and dirty" to some degree. This follows from the fact that in constructing every model some simplifying assumptions are made. Reality is too difficult to represent in all its complexity. These simplifying assumptions reduce the generality of the model and solutions derived from it. But this is only a polite way of saying that quickness and dirtiness are involved. It is well for the operations researcher to realize that an approximate solution which is used may be a great deal better than a more exact solution which is not.

For further discussion of the problems of implementing the solution, see Ref. 1, Chap. 21.

## REFERENCES

1. C. W. Churchman, R. L. Ackoff, and E. L. Arnoff, *Introduction to Operations Research*, Wiley, New York, 1957.
2. F. N. Trefethen, "A History of Operations Research," in *Operations Research for Management*, J. F. McCloskey and F. N. Trefethen (Editors), The Johns Hopkins Press, Baltimore, Md., 1954.
3. M. L. Hurni, Observations on Operations Research, *J. Opns. Research Soc. Am.*, 2 [3], 234-248 (1954).

4. H. F. Smiddy and L. Naum, Evolution of a "Science of Managing" in America, *Mgmt. Sci.*, 1 [1], 1-31 (1954).
5. J. H. Curtiss, "Sampling Methods Applied to Differential and Difference Equations," in *Seminar on Scientific Computation*, International Business Machines Corporation, New York, 87-109, Nov. 1949.
6. I. S. Sokolnikoff and E. S. Sokolnikoff, *Higher Mathematics for Engineers and Physicists*, McGraw-Hill, New York, 1941.
7. R. E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, N. J., 1957.
8. C. C. Holt, F. Modigliani, and H. A. Simon, A linear decision rule for production and employment scheduling, *Mgmt. Sci.*, 2 [1], 1-30 (1955).
9. C. C. Holt and H. A. Simon, Optimal decision rules for production and inventory control, *Proceedings of the Conference on Operations Research in Production and Inventory Control*, Case Institute of Technology, Cleveland, O., 1954.
10. The RAND Corporation, *A Million Random Digits*, The Free Press, Glencoe, Ill., 1955.
11. H. Kahn, *Applications of Monte Carlo*, Project RAND, RM-1237-AEC, RAND Corporation, Santa Monica, Calif., April 19, 1954.
12. G. W. King, The Monte Carlo method as a natural mode of expression in Operations Research, *J. Opns. Research Soc. Am.*, 1 [2], 46-51 (1953).
13. U. S. Department of Commerce, National Bureau of Standards, *Monte Carlo Method*, Applied Mathematics Seminar 12, June 11, 1951.
14. J. B. Crockett and H. Chernoff, Gradient methods of maximization, *Pacific J. Math.*, 5 (1955).
15. B. Klein, Direct use of extremal principles in solving certain optimizing problems involving inequalities, *J. Opns. Research Soc. Am.*, 3 [2], 168-175 (1955).
16. H. W. Kuhn and A. W. Tucker, "Nonlinear Programming," in *Second Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman (Editor), University of California Press, Berkeley, Calif., 481-492, 1951.
17. K. Arrow, T. Harris, and J. Marschak, Optimal inventory policy, *Econometrica*, 19 [3], 250-272 (1951).
18. C. Eisenhart, *Some Inventory Problems*, National Bureau of Standards, Techniques of Statistical Inference, A2-2C, Lecture 1, Jan. 6, 1948 (hctographed notes).
19. C. B. Tompkins, Lead time and optimal allowances—an extreme example, *Conference on Mathematical Problems in Logistics*, George Washington University, Appendix I to Quarterly Progress Rept. No. 1, Dec. 1949-Feb. 1950.
20. T. M. Whitin, *The Theory of Inventory Management*, 2nd edition, Princeton University Press, Princeton, N. J., 1957.
21. A. Dvoretzky, J. Kiefer, and J. Wolfowitz, On the optimal character of the  $(s, S)$  policy in inventory theory, *Econometrica*, 21 [4], 586-596 (1953).
22. A. Dvoretzky, J. Kiefer, and J. Wolfowitz, The inventory problem, *Econometrica*, 20 [2], 187-222 (1952); [3], 450-466 (1952).
23. E. B. Berman and A. J. Clark, An optimal inventory policy for a military organization, RAND Rept. D-647, RAND Corporation, Santa Monica, Calif., March 30, 1955.
24. H. A. Simon, On the application of servomechanism theory in the study of production control, *Econometrica*, 20 [2], 247-268 (1952).
25. R. Bellman, I. Glicksberg, and O. Gross, On the optimal inventory equation, *Mgmt. Sci.*, 2 [1], 83-104 (1955).
26. H. J. Vassian, Application of discrete variable servo theory to inventory control, *J. Opns. Research Soc. Am.* 3 [3], 272-282 (1955).

27. A. Charnes, W. W. Cooper, and D. Farr, Linear programming and profit preference scheduling for a manufacturing firm, *J. Opns. Research Soc. Am.*, **1** [3], 114-129 (1953).
28. G. Dannerstedt, Production scheduling for an arbitrary number of periods given the sales forecast in the form of a probability distribution, *J. Opns. Research Soc. Am.*, **3** [3], 300-318 (1955).
29. R. Bellman, Some applications of the theory of dynamic programming, *J. Opns. Research Soc. Am.*, **2** [4], 275-288 (1954).
30. R. Bellman, Some problems in the theory of dynamic programming, *Econometrica*, **22** [1], 37-48 (1954).
31. R. Bellman, The theory of dynamic programming, *Bull. Am. Math. Soc.* [6], 503-516 (1954).
32. R. Bellman, I. Glicksberg, and O. Gross, The theory of dynamic programming as applied to a smoothing problem, *J. Soc. Ind. Appl. Math.*, **2** [2], 82-88 (1954).
33. T. M. Whitin, Inventory control and price theory, *Mgmt. Sci.*, **2**, 61-68 (1955).
34. T. M. Whitin, Inventory control research: A survey, *Mgmt. Sci.*, **1**, 32-40 (1954).
35. H. A. Simon and C. C. Holt, The control of inventory and production rates—A survey, *J. Opns. Research Soc. Am.*, **2** [3], 289-301 (1954).
36. A. Charnes, W. W. Cooper, and A. Henderson, *An Introduction to Linear Programming*, Wiley, New York, 1953.
37. G. H. Symonds, *Linear Programming: The Solution of Refinery Problems*, Esso Standard Oil Co., New York, 1955.
- 37a. A. Orden, Survey of research on mathematical solutions of programming problems, *Mgmt. Sci.*, **1**, 170-172 (1955).
38. G. B. Dantzig, Ref. 40, Chaps. I, II, XX, XXI, and XXIII.
39. A. Charnes and W. W. Cooper, The stepping stone method of explaining linear programming calculations in transportation problems, *Mgmt. Sci.*, **1** [1], 49-69 (1954).
40. T. C. Koopmans (Editor), *Activity Analysis of Production and Allocation*, Cowles Commission Monograph No. 13, Wiley, New York, 1951.
41. A. Henderson and R. Schlaifer, Mathematical programming, *Harvard Business Rev.*, **32**, 73-100 (May-June 1954).
42. T. L. Saaty, Résumé of queuing theory, *Opns. Research*, **5**, 161-200 (1957).
43. W. Feller, *An Introduction to Probability Theory and Its Applications*, Wiley, New York, 1950.
44. E. Brockmeyer, H. L. Holstrom, and Arne Jensen, The life and works of A. K. Erlang, *Trans. Danish Acad. Tech. Sci.*, **2**, Copenhagen, 1948.
45. Raymond, Haller and Brown, Inc., *Queuing Theory Applied to Military Communication Systems*, State College, Pa., 1956.
46. A. Cobham, Priority assignment in waiting line problems, *J. Opns. Research Soc. Am.*, **2**, 70-76 (1954); also **3**, 547 (1955).
47. J. Y. Barry, A priority queuing problem, *Opns. Research*, **4**, 385-386 (1956).
48. E. Koenigsberg, Queuing with special service, *Opns. Research*, **4**, 213-220 (1956).
49. D. Y. Barrer, A waiting line problem characterized by impatient customers and indifferent clerks, *J. Opns. Research Soc. Am.*, **3**, 360-361 (1955).
50. R. Kronig, On time losses in machinery undergoing interruptions, *Physica*, **10**, 215-224 (1943).
51. A. B. Clarke, The time-dependent waiting line problem, Univ. Mich. Engr. Research Inst., Rept. No. M720-1 R 39, 1953.
52. A. B. Clarke, A waiting line process of Markov type, *Ann. Math. Stat.*, **27** [2], 452-459 (1956).

53. T. Homma, On a certain queuing process, *Rept. Stat. Appl. Research*, **4** [1] (1955).
54. R. R. P. Jackson, Queuing systems with phase type service, *Operat. Research Quart.*, **5**, 109-120 (1954).
55. S. M. Johnson, Optimal two- and three-stage production schedules with setup times included, *Nav. Research Log. Quart.*, **1** [1], 61-68 (1954).
56. R. Bellman, Mathematical aspects of scheduling theory, RAND Rept. P-651, RAND Corporation, Santa Monica, Calif., April 11, 1955.
57. S. B. Akers, Jr., and J. Friedman, A non-numerical approach to production scheduling problems, *J. Opns. Research Soc. Am.*, **3**, 429-442 (1955).
58. A. W. Brown, A note on the use of a Pearson type III function in renewal theory, *Ann. Math. Stat.*, **11**, 448-453 (1940).
59. N. R. Campbell, The replacement of perishable members of an operating system, *J. Roy. Stat. Soc.*, **B7**, 110-130 (1941).
60. W. Feller, On the integral equation of renewal theory, *Ann. Math. Stat.*, **13**, 243-267 (1941).
61. A. J. Lotka, A contribution to the theory of self-renewing aggregates, with special reference to industrial replacement, *Ann. Math. Stat.*, **10**, 1-25 (1939).
62. A. J. Lotka, *The Present Status of Renewal Theory*, Waverly Press, Baltimore, Md., 1940.
63. B. Epstein and M. Sobel, Life Testing. I, *J. Am. Stat. Assoc.*, **48**, 486-502 (1953).
64. B. Epstein and M. Sobel, Some theorems relevant to life testing from an exponential distribution, *Ann. Math. Stat.*, **25**, 373-381 (1954).
65. L. Goodman, Methods of measuring useful life of equipment under operational conditions, *J. Am. Stat. Assoc.*, **48**, 503-530 (1953).
66. *Tested Approaches to Capital Equipment Replacement*, Special Rept. No. 1, American Management Association, New York, 1954.
67. J. von Neumann, Zur Theorie der Gesellschaftsspiele, *Math. Ann.*, **100**, 295-320 (1928).
68. J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, 3rd edition, Princeton University Press, Princeton, N. J., 1953.
69. J. D. Williams, *The Compleat Strategyst*, McGraw-Hill, New York, 1954.
70. L. Friedman, *Competitive Bidding Strategies*, Ph.D. Dissertation, Case Institute of Technology, Cleveland, O., 1957.
71. J. C. C. McKinsey, *Introduction to the Theory of Games*, McGraw-Hill, New York, 1952.
72. R. L. Ackoff, *The Design of Social Research*, University of Chicago Press, Chicago, Ill., 1953.
73. N. R. Campbell, *An Account of the Principles of Measurements and Calculations*, Longmans, Green and Co., New York, 1928.
74. C. W. Churchman, "A Materialist Theory of Measurement," in *Philosophy for the Future*, R. W. Sellars, V. J. McGill, and M. Farber (Editors), Macmillan, New York, 1949.
75. E. Nagel, Measurement, *Erkenntnis*, **2**, 313-333 (1931).
76. E. Nagel, *On the Logic of Measurement*, Thesis, Columbia University, New York, 1930.
77. F. F. Stephan, "Mathematics, Measurement, and Psychophysics," in *Handbook of Experimental Psychology*, S. S. Stevens (Editor), Wiley, New York, 1951.
78. S. S. Stevens, On the problem of scales for the measurement of psychological magnitudes, *J. Univ. Sci.*, **9**, 94-99 (1939).

79. S. S. Stevens, On the theory of scales of measurement, *Science*, **103**, 677-680 (1946).
80. H. B. Horton, *Random Decimal Digits*, Interstate Commerce Commission, Washington, D. C., 1949.
81. M. G. Kendall, "Tables of Random Sampling Numbers," in *Tracts for Computers*, No. 24, Cambridge University Press, Cambridge, England, 1940.
82. L. H. C. Tippett, "Tables of Random Sampling Numbers," in *Tracts for Computers*, No. 15, Cambridge University Press, Cambridge, England, 1927.
83. M. G. Kendall and B. B. Smith, Randomness and random sampling of numbers, *J. Roy. Stat. Soc.*, **101**, 147-166 (1938).
84. C. W. Churchman, *Statistical Manual: Methods of Making Experimental Inferences*, Pittman-Dunn Laboratory, Frankford Arsenal, Philadelphia, Pa., 1951.
85. Statistical Research Group, *Sequential Analysis of Statistical Data: Application*, Columbia University Press, New York, 1946.
86. A. Wald, Foundations of a general theory of sequential decision functions, *Econometrica*, **15**, 279-313 (1947).
87. A. Wald, *Sequential Analysis*, Wiley, New York, 1947.
88. W. C. Cochran and G. M. Cox, *Experimental Designs*, Wiley, New York, 1950.
89. W. T. Federer, *Experimental Design*, Macmillan, New York, 1955.
90. R. A. Fisher, *The Design of Experiments*, Oliver and Boyd, London, 1949.
91. H. B. Mann, *Analysis and Design of Experiments*, Dover, New York, 1949.
92. W. E. Deming, *Some Theory of Sampling*, Wiley, New York, 1950.
93. M. H. Hansen, W. N. Hurwitz, and W. G. Madow, *Sampling Survey Methods and Theory*, Wiley, New York, 1953.
94. F. F. Stephan, History of the uses of modern sampling, *J. Am. Stat. Assoc.*, **43**, 12-39 (1948).
95. F. Yates, *Sampling Methods for Censuses and Surveys*, Griffin, London, 1949.
96. W. J. Dixon and F. J. Massey, Jr., *Introduction to Statistical Analysis*, McGraw-Hill, New York, 1951.
97. R. A. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd, London, 1948.
98. A. Hald, *Statistical Theory with Engineering Applications*, Wiley, New York, 1952.
99. P. G. Hoel, *Introduction to Mathematical Statistics*, 2nd edition, Wiley, New York, 1954.
100. P. O. Johnson, *Statistical Methods in Research*, Prentice-Hall, New York, 1949.
101. F. J. Massey, Jr., The Kolmogorov-Smirnov test for goodness of fit, *J. Am. Stat. Assoc.*, **46**, 68-78 (1951).
102. E. B. Mode, *The Elements of Statistics*, Prentice-Hall, New York, 1941.
103. G. W. Snedecor, *Statistical Methods*, 4th edition, Iowa State College Press, Ames, Ia., 1946.
104. H. M. Walker, *Elementary Statistical Methods*, Holt, New York, 1943.
105. S. S. Wilks, *Elementary Statistical Analysis*, Princeton University Press, Princeton, N. J., 1949.
106. C. W. Churchman, *Theory of Experimental Inference*, Macmillan, New York, 1948.
107. J. Neyman and E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses, *Phil. Trans.*, **A231**, 289-337 (1933).
108. A. Wald, "On the Principles of Statistical Inference," in *Notre Dame Mathematical Lectures*, No. 1, Notre Dame University, Notre Dame, Ind., 1942.
109. A. Wald, Statistical decision functions, *Ann. Math. Stat.*, **20**, 165-205 (1949).

110. A. J. Duncan, *Quality Control and Industrial Statistics*, Irwin, Chicago, Ill., 1952.
111. E. L. Grant, *Statistical Quality Control*, 2nd edition, McGraw-Hill, New York, 1952.
112. J. M. Juran (Editor), *Quality Control Handbook*, McGraw-Hill, New York, 1946.
113. C. W. Kennedy, *Quality Control Methods*, Prentice-Hall, New York, 1948.
114. S. B. Littauer, Social aspects of scientific method in industrial production, *Phil. Sci.*, **21**, 93-100 (1954).
115. S. B. Littauer, Technological stability in industrial operations, *Trans. N. Y. Acad. Sci.*, ser. II, **13** [2], 66-72 (1950).
116. P. Peach, *An Introduction to Industrial Statistics and Quality Control*, 2nd edition, Edwards and Broughton, Raleigh, N. C., 1947.
117. J. G. Rutherford, *Quality Control in Industry—Methods and Systems*, Pitman, New York, 1948.
118. W. A. Shewhart, *Statistical Methods from the Viewpoint of Quality Control*, U. S. Department of Agriculture, Washington, D. C., 1939.



# INFORMATION THEORY AND TRANSMISSION

## D. INFORMATION THEORY AND TRANSMISSION

16. *Information Theory, by Peter Elias*
17. *Smoothing and Filtering, by Pierre Mertz*
18. *Data Transmission, by Pierre Mertz*



## Information Theory

Peter Elias

1. Introduction	16-01
2. General Definitions	16-02
3. Simple Discrete Sources	16-08
4. More Complicated Discrete Sources	16-19
5. Discrete Noiseless Channels	16-24
6. Discrete Noisy Channels I. Distribution of Information	16-26
7. Discrete Noisy Channels II. Channel Capacity and Interpretations	16-32
8. The Continuous Case	16-39
References	16-46

### 1. INTRODUCTION

**Basis of Information Theory.** As used here, *information theory* is a body of results based on a particular quantitative definition of *amount of information*. This definition has a firm claim to unique importance in connection with the engineering questions which arise in systems which transmit and store information. It has proved interesting and sometimes useful in other fields (Refs. 1-4). However, other definitions have also been proposed (Ref. 5) and one of them has a long and useful history in statistics (Ref. 6). Caution is therefore needed in applying this definition to a situation in which the theorems which are its main justification in transmission and storage problems do not apply.

**Communication Theory.** Information theory is a subdivision of a broader field, the *statistical theory of communication*, which includes all the

probabilistic analysis of communications problems. This broad field includes in addition to information theory the analysis of random noise (Ref. 7), work on optimum linear filtering and prediction (Ref. 8, see also Chap. 17), statistical analysis of signal detection (Refs. 9, 10), and many other applications of probabilistic ideas which make no use of an information measure.

**Note on Terminology.** Some authors, particularly in England, use *information theory* in a very broad sense, to include theories of scientific method and of statistical inference along with communications problems (Ref. 11). They then use "communication theory," or "mathematical theory of communication," or "theory of selective information," to denote what is here defined as information theory.

**Mathematical Character.** Information theory is essentially a branch of mathematics. Although the language has a physical ring, the words *information source, channel, coder, etc.*, are mathematical concepts physically inspired. The theory can be presented as a formal series of definitions, theorems, and comments. However, its relevance to a given problem is then not very clear. Section 2 provides contextual definitions and qualitative results: the later sections are more formal.

## 2. GENERAL DEFINITIONS

### A Communications System

Figure 1 shows the model of a communications system which is used in information theory. At the transmitter the source produces an output that is coded and fed into the channel. The channel output may be identi-

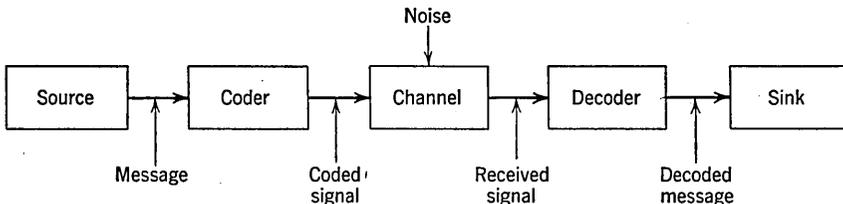


FIG. 1. The model of a communications system which is used in information theory.

cal to the channel input, or it may be altered by noise or distortion. At the receiver the channel output is decoded and used. The model derives from communications, but it is applicable to the communications aspects of other problems. *Examples.* The storage of a digital computer is a channel, possibly noisy, with input and output separated by time. A control system may be a channel with electrical input and mechanical output.

## The Information Source

For purposes of the theory, the source in a given analysis is the point at which information enters the part of the system under consideration. The source may not actually generate information, but may merely store or relay it. *Examples:* a stack of telegrams waiting to be transmitted; a reel of recorded magnetic tape. Whether the source under consideration is a true generator of information or merely a storage point is not of concern.

**Controlled and Uncontrolled Sources.** A *controlled source* is one which generates information at a rate controllable by the transmitter. *Examples.* The stack of telegrams being read by a telegraph operator is a controlled source; so is a speaker who may be slowed down by his audience when he speaks too rapidly for them to take notes. An *uncontrolled source* is one which produces information at a rate determined internally, which cannot be adjusted to the coding and transmission facilities available.

**Segmentation.** The *output* of a source is a sequence of symbols. It is convenient to break this sequence into segments at a number of different *levels*. This process is called *segmentation* by the linguists. *Example.* The output of a teletype system is a sequence of binary selections, "Mark" and "Space." If these are denoted by "0" and "1," then 0 and 1 are called the *elements* of the representation.

The group of consecutive elements which represents a single letter, number or mark is called a *character* or *letter* of the representation, and the set of all possible characters or letters is called the *alphabet*. The alphabet in a teletype system may be thirty-two characters in number, with each character a group of five elements.

**Words and Messages.** In many alphabets one of the characters is called the *space*, and given special significance. Sequences of characters occurring between successive spaces are called *words*. A sequence of words which is more or less independent of the preceding and succeeding source output is called a *message*. *Example.* A single telegram might be called a message. If successive source symbols are highly correlated the whole (possibly infinite) source output is the message.

**Other Levels.** The above list of levels is not exhaustive, nor are all these levels relevant to the description of a particular source. *Examples.* In written English the additional levels of *syllable*, *phrase*, *clause*, *sentence*, and *paragraph* are recognized, but the element level is not used.

In the analysis of spoken language a set of elements, the *distinctive features*, have been introduced (Ref. 12). In a binary digital computer, *element* and *character* coincide as being the symbols 0 and 1 used to represent the binary digits. In a binary-coded-decimal machine the elements are the same, but the *characters* are groups of four, five, or six elements representing one decimal digit or one alphanumeric symbol. The group of

digits which fits into one storage register is called a *word*, and one line of coding, consisting of several related words, which is called an *instruction* in computer terminology, might be called a message.

**Choice of Levels.** The choice of levels of segmentation is not standardized outside linguistics, nor is there any agreement on terminology. When no particular type or level of segmentation is implied, the output of a source will be a sequence of *symbols*, selected from some finite *alphabet*. When two levels are needed at once, as in discussion of word-by-word translation of a sequence of letters, *word* will be used for the higher level and symbol or letter for the lower level. In mathematical discussion, a segment of a sequence will be called a *message* only if it is strictly statistically independent of preceding segments.

### Representations and Codes

If a source makes a series of binary selections, its output may be represented, for example, by a sequence of A's and B's or by a sequence of 0's and 1's. These are two *representations* of the source output. If the first is taken as *primary*, then the second is called a *coded version* of the first.

**Codebooks.** To get from one representation to the other requires a dictionary, called an *encoding codebook*. This has two entries:

$$(1) \quad \begin{array}{l} A \rightarrow 0 \\ B \rightarrow 1 \end{array}$$

to get back from the second representation to the first requires a *decoding codebook* with the entries

$$(2) \quad \begin{array}{l} 0 \rightarrow A \\ 1 \rightarrow B \end{array}$$

**Codes.** A *code* is a transformation, which is defined by an encoding codebook or an equivalent set of rules. If a coded message is to be decoded, the inverse transformation as defined by the decoding codebook is also required. In the example of eqs. (1) and (2), the transformation is one-to-one on each symbol and defines its own inverse, so that only one codebook is required and it may be written with double-headed arrows:

$$(3) \quad \begin{array}{l} A \leftrightarrow 0 \\ B \leftrightarrow 1 \end{array}$$

In representations with large alphabets the two codebooks may still be useful even if only one is necessary. *Example.* It is convenient to find a telephone number in a standard directory by looking up a name, but the inverse operation is tedious, though unambiguous.

**Transliteration.** A code is called a *transliteration* if each input symbol is transformed directly into one output symbol, so that symbol-by-symbol coding is possible and the number of entries in the codebooks is equal to the size of the alphabet. The code of eq. (3) has this property, but not all one-to-one codes do. *Example.* The representation of the binary source output as a *sequence* of A's and B's may be coded into a representation as a *sequence* of 0's and 1's by the codebook

$$(4) \quad \begin{aligned} AA &\rightarrow 0 \\ AB &\rightarrow 10 \\ BA &\rightarrow 110 \\ BB &\rightarrow 111 \end{aligned}$$

To each input sequence corresponds one output sequence, which may be decoded into its original form by reversing the arrows in the codebook if the time origin of the coded version is known. The coded output is a one-to-one transformation on the input sequences, but it is not a transliteration of the symbols A and B. Choose a different level of segmentation (word, rather than character) and let the first representation consist of sequences of the four *words* AA, AB, BA, BB. Then if the four *words* 0, 10, 110, and 111 are taken as the dictionary for the second representation, the code becomes a transliteration.

**Significant Codes.** In assigning code numbers to objects (*Example:* the items in a catalog) a distinction is made between *significant* and *non-significant* codes. Each code number may be considered as a coded version of a description of the item in English. A *significant* code is one in which transliteration is possible at some level of segmentation below that of the entire code number. *Example.* The code number assigned to a garment in a catalog may consist of a sequence of groups of decimal digits, the first group denoting type of garment, the second size, the third color. Each group may be independently decoded into English words, so that transliteration is possible if each group is considered as a word in the coded version of the message.

A code that cannot be decoded piece by piece is called *nonsignificant*. *Example:* a code assigning simple serial numbers to items in a catalog.

**Coding and Decoding Delay.** Note that *coding and decoding delays arise when the coding is not a transliteration*. *Example.* In the codebook of eq. (4), after the source has selected its first symbol, the coder must wait until the second symbol has also been selected before it can encode the pair. In decoding, after a 1 has been received, it is necessary to wait for one or two more input symbols before the appropriate output pair can be selected from the set AB, BA, BB.

**Representation and Selection.** The output of a coder is called a *representation* of its input if it is obtained from the input by a one-to-one transformation with at most a finite encoding delay. Note that this definition agrees with the colloquial meaning of representation for a significant code, in which each segment of the output *represents* a corresponding segment of the input.

A nonsignificant code requires a different interpretation. The code number corresponding to a telegraphic greeting is not a modified version of the message, but an instruction as to where in the decoding codebook the message will be found. The coded version does not *represent* the message but *selects* it. *This concept is basic for information theory* (Ref. 13).

**Coders.** The coder in Fig. 1 *matches the source to the channel*. The first requirement on the coder is that it match alphabets. It must transform sequences of symbols from the source alphabet into sequences of symbols in the alphabet which the channel will accept. This requirement does not specify the coder completely. The two codebooks of eqs. (3) and (4) both transform A's and B's into 0's and 1's, but they describe different coders.

**Statistical Matching.** For economy of transmission facilities the coder may be designed to minimize the number of channel symbols required, on the average, per source symbol. This requires knowledge of the statistics of the source. *Example.* The codebook of eq. (4) is more complicated than the codebook of eq. (3) and introduces delay. However, if A's occur 99 per cent of the time and B's only 1 per cent, the output coded via eq. (4) will require only 0.5015 channel symbol per source symbol, whereas the output coded via eq. (3) will require one channel symbol per source symbol, and so will take nearly twice as long to transmit. Economical coding will be discussed in Sect. 3.

## Channels

A *discrete channel*, like a coder, accepts a sequence of symbols selected from its input alphabet and produces a related sequence of symbols selected from its output alphabet. The precise boundaries of the channel in a given system are a matter of choice. *Example.* A teletype system may be analyzed by using as a channel the medium of transmission of the electric pulses. A second analysis of the same system might treat the channel as running from input keyboard to output printer.

As essential difference between a channel and a coder is that the channel output may not be an accurate representation of its input: some information about the message may get lost in transit. This may occur in two ways.

**a. Loss.** A channel is *lossy* if it is possible to make finer distinctions at its input than are preserved in its output. *Example.* Pulses of fixed duration and of any amplitude greater than 0 may be used successfully to trigger a circuit which provides an output pulse of fixed duration and amplitude, no output being produced if the input pulse is smaller than 0. A channel is made of this circuit by using as an input alphabet pulses of the ten amplitude levels  $-5, -4, \dots, -1, +1, \dots, +5$ . This channel accepts ten input symbols and produces two output symbols: it loses the additional

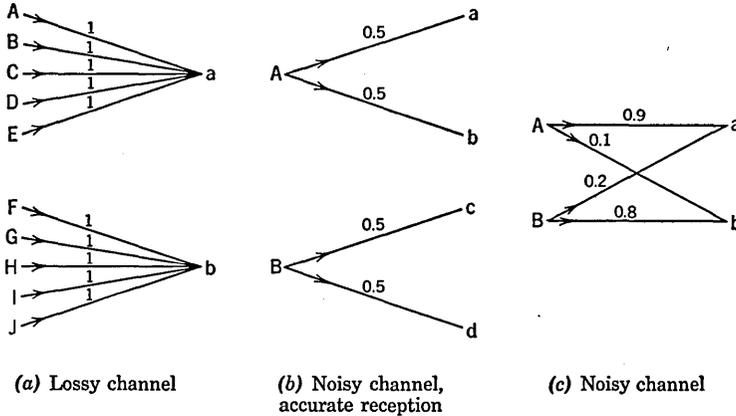


FIG. 2. Some examples of discrete channels.

amplitude information present in the input. Such a channel is shown schematically in Fig. 2a.

**b. Noise.** A channel is *noisy* if a given input sequence may be received as any one of a number of possible output sequences, depending on random action of the channel. In a lossy channel the received sequence is determined by the transmitted sequence; in a noisy channel this is not true. *Examples.* Figure 2b shows a noisy channel in which the noise does not bother the receiver, who is still able to tell what has been transmitted, although the transmitter does not know exactly what has been received. Figure 2c shows a noisy channel in which the channel noise prevents either transmitter or receiver from knowing with certainty what happens at the other end of the channel.

**Decoders.** Any of the channels of Fig. 2 can transmit information at a definite rate with an arbitrarily small probability of error. This can be done simply for the first two channels by merely lumping together some of the input symbols and output symbols, respectively. It can also be done for the channel of Fig. 2c, by making use of the proper coder and

decoder. The coder is still like the codebooks of eqs. (3) and (4), although the matching job performed is more sophisticated. But the decoder is different. Since the channel performs a transformation on the input sequences which is many-to-many rather than one-to-one, the decoder must perform a many-to-one transformation. In the channel, each input sequence may produce many output sequences. The decoder must decode all of these (or at least all of them which occur with appreciable probability) into the same output sequence, if it is to avoid making errors. This will be discussed in Sect. 7.

### 3. SIMPLE DISCRETE SOURCES

#### Self-Information Measure

Let  $x$  denote a particular event, and let  $\text{Prob } \{x\}$  be its probability. The *amount of information* associated with the occurrence of the event  $x$  is defined to be

$$(5) \quad I(x) = -\log \text{Prob } \{x\},$$

where the choice of logarithmic base corresponds to the choice of a unit of information. The quantity  $I(x)$  is sometimes called *self-information* (Ref. 14) to distinguish it from the *mutual* information relating two events, discussed in Sect. 6.

**Units.** Logarithms to the base 2 are chosen for eq. (5). The resulting unit is the *bit*, which is the amount of information associated with the occurrence of an event of a priori probability one-half. Other information units include the *Hartley*, which is the information given by an event of probability  $1/10$ , and the *nat*, or natural unit, which is the information given by an event of probability  $1/e$ , where  $e$  is the base of Napierian logarithms,  $e = 2.71828 \dots$ .

**Bits and Binits.** In computer terminology *bit* is often used as a contraction of *binary digit*. This practice cannot be followed in information theory, since the occurrence of a binary symbol with a priori probability other than  $1/2$  does not provide a bit of information. The word *binits* will therefore be used as an abbreviation for *binary digit* (Ref. 15).

**Properties.** The information measure has the following two important properties.

1. Since  $\text{Prob } \{x\} \leq 1$  for any event  $x$ ,

$$(6) \quad I(x) \geq 0.$$

2. Let  $x$  and  $y$  be two statistically independent events, and let  $x, y$  denote the event which is their joint occurrence. Then

$$(7) \quad I(x, y) = I(x) + I(y),$$

since the probability of the event  $x, y$  is, by hypothesis of independence, the product of the probabilities of  $x$  and  $y$ .

### Distribution of Information

**Message Source.** Consider a set  $M$  of  $n$  different messages,  $M = \{m_i\}$ ,  $1 \leq i \leq n$ , and a random process that generates sequences by selecting messages from this set. (The word *message* implies that successive selections are statistically independent. See Sect. 2.) Let  $x_k$  be the  $k$ th message selected in time sequence,  $-\infty < k < \infty$ . Then  $x_k$  is a random variable, taking values from the set  $M = \{m_i\}$ , with

$$(8) \quad \text{Prob} \{x_k = m_i\} = p(m_i) = p_i$$

as the probability of selecting the  $i$ th message as the  $k$ th choice. As eq. (8) implies, it is assumed that the process is stationary, so that  $p_i$  is independent not only of the earlier selections but also of the time index  $k$ .

**Bar Plot of Distribution of Information.** The *amount of information* associated with the selection of message  $m_i$  is then also independent of  $k$  and of prior selections: it is given by

$$(9) \quad I(m_i) = I_i = -\log p_i = -\log p(m_i).$$

The random variable  $I(x_k)$  takes its values from the set

$$I(m_i) = I_i,$$

with probabilities

$$(10) \quad \text{Prob} \{I(x_k) = I_i\} = p_i.$$

Because of eq. (9), if all the probabilities  $p_i$  are different from one another, then on a bar plot of the distribution of information, the bars which give the probabilities of the different possible information values all terminate on the single exponential given by

$$(11) \quad p_i = 2^{-I_i}.$$

**Mean and Variance.** The information distribution is completely determined by the probabilities  $p_i$ , via eq. (9). The most important parameters of the distribution are its mean value,

$$(12) \quad \bar{I} = -\sum_{i=1}^n p_i I_i = -\sum_{i=1}^n p_i \log p_i,$$

and its variance,

$$(13) \quad \sigma_I^2 = \overline{(I - \bar{I})^2} = \bar{I}^2 - \bar{I}^2.$$

EXAMPLE 1. The source illustrated in Fig. 3 selects messages from the set  $M = \{A, B, C\}$  with probabilities  $\{0.755, 0.185, 0.060\}$  and information values  $\{0.405, 2.434, 4.059\}$ . The information distribution has mean value 1.00 bit/symbol, and standard deviation  $\sigma_I = 1.10$  bits/symbol.

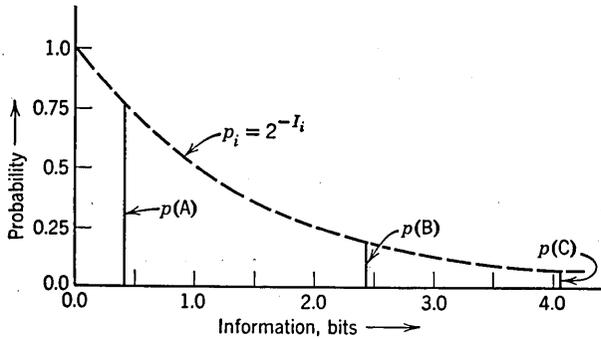


FIG. 3. Bar plot of information distribution.

The three message probabilities terminate on the exponential of eq. (11), shown dotted.

EXAMPLE 2. The source illustrated by Fig. 4 has an alphabet  $M = \{0, 1\}$ , with probabilities  $\{\frac{1}{2}, \frac{1}{2}\}$ . This distribution also has a one-bit

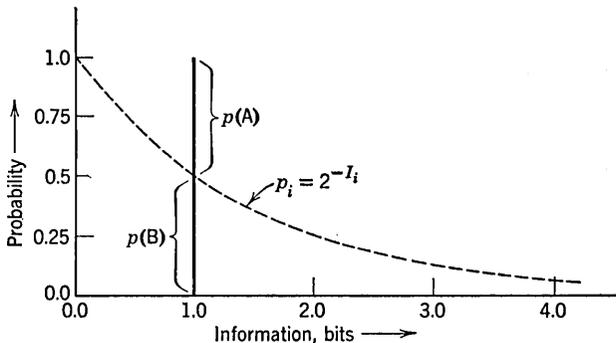


FIG. 4. Bar plot of degenerate information distribution.

mean value, but has zero variance, since the selection of either symbol requires the same amount of information. Note that in this case the total probability line occurring at  $I = 1$  does not terminate on the exponential of eq. (9), because of the fact that the two symbols have the same information value. Such a distribution is called *degenerate*. The degeneracy

may be removed by shifting the probabilities slightly, to  $\{1/2 - \delta, 1/2 + \delta\}$ , so that the two bars wind up side by side rather than end on end.

**Rate and Entropy.** For sources like those illustrated in Figs. 3 and 4, which generate sequences of *messages* rather than sequences of statistically related symbols, the *average rate*  $R$  at which the source generates information, in bits per source symbol, is just the mean of the information distribution. From eq. (12), this mean is a function of the probabilities  $\{p_i\}$  of the message set only. When considered as a function of the probabilities, rather than as a mean of the information distribution, this quantity is called the *entropy*  $H$  of the probabilities  $\{p_i\}$ . Thus for these sources,

$$(14) \quad R = \bar{I} = H \text{ bits/source symbol.}$$

For the more complicated sources to be discussed later the average rate  $R$  is still the average value of the self-information, but the averaging

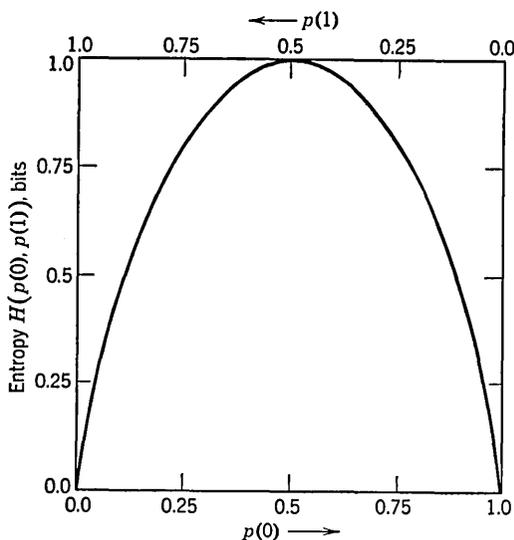


FIG. 5. The entropy function  $H(p(0), p(1)) = -p(0) \log p(0) - p(1) \log p(1)$ .

operation is more complicated, and  $\bar{I}$  no longer has the form of the entropy function of a probability distribution.

The entropy function  $H(p_1, p_2)$  is illustrated in Fig. 5. Since  $p_1 + p_2 = 1$ , this is actually a function of a single variable only.

### Binary Coding

The rate of a source is a significant parameter because it determines the communications facilities required to transmit the source output after

proper coding. The source in Fig. 4 generates information at a rate  $R = 1$  bit/symbol, and each output symbol is just one binit. The curve of Fig. 5 shows that one bit is the maximum average amount of information that one binit can convey. In this case, the rate  $R$  has the interpretation that the source output may be represented in binites so as to require  $R = 1$  output binites per source symbol. This interpretation can be generalized to other sources.

**FIRST BINARY CODING THEOREM (Controlled Sources).** *Given a discrete message source which generates information at an average rate  $R$  bits per message and given any  $\delta > 0$ , it is possible to construct a representation of sequences of messages as sequences of binary symbols so that, on the average, less than  $R + \delta$  output binary symbols are required per input symbol from the source. It is not possible to find a representation using fewer than  $R$  output binary symbols per source symbol (Ref. 16).*

A code which satisfies the requirements of this theorem does the job of statistical matching referred to in Sect. 2.

**Shannon-Fano Coding.** The general strategy in constructing efficient binary codes is to divide the message set into two subsets of nearly equal probability and to use the first digit of the coded output sequence to indicate in which half the selected message lies. Each half is divided into two subsets again by the next digit, and the process terminates on subsets which contain only one message (Ref. 17).

This procedure is not quite explicit, however. It will not be possible to make all dichotomies equiprobable unless all the message probabilities are powers of  $1/2$ . If not, then there are many possible not-quite-perfect codes, and it is difficult to choose among them. The following procedure, called *Huffman coding*, is explicit, and it gives a "best possible" code (Ref. 18).

#### **Huffman Coding.**

1. List all possible messages in order of decreasing probability, and assign as the last digit in the coded output a 0 to the next-to-last message and a 1 to the last message. These two messages will agree in all the (as yet unknown) digits preceding the last one.

2. Merge the last two messages, adding their probabilities, and insert the sum in its proper position in the list of message probabilities. Now repeat step 1. Continue until all messages are merged.

**EXAMPLE.** The process is illustrated for the message set of Fig. 3 in Fig. 6, by a kind of a graph which is called a tree for obvious reasons. The code for each message is read off starting at the left node and reading the 0's and 1's which label the branches along the (unique) path terminating in the selected message. For Fig. 6 this leads to the codebook

(15)

$$\begin{aligned} A &\rightarrow 0 \\ B &\rightarrow 10 \\ C &\rightarrow 11. \end{aligned}$$

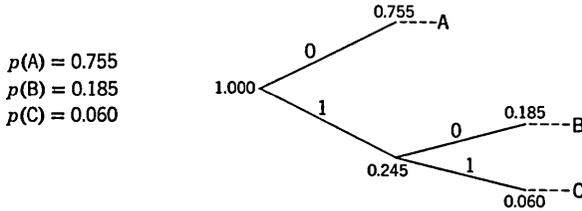


FIG. 6. A coding tree.

**Codewords. The Prefix Property.** The output sequences of a codebook are called *codewords*. The code of eq. (15) and of Fig. 6 illustrates a characteristic feature of Huffman codes called the prefix property: no codeword is a prefix of any other longer codeword.

The prefix property is a *sufficient* condition to guarantee that a sequence of codewords written down in order without spacing can be uniquely decoded into a sequence of source symbols. Decodability is required in order that the output be a representation of the input, and spaces between words are not permitted since, if they were used, the output alphabet would be ternary and not binary. The prefix property is not *necessary*, however. It is possible to construct codes, which can be decoded after some delay, that do not satisfy this condition.

EXAMPLE. The codebook

$$\begin{aligned} A &\rightarrow 0 \\ B &\rightarrow 01 \\ C &\rightarrow 11 \end{aligned}$$

is decodable, but the codewords do not satisfy the prefix condition since 0 is a prefix of 01. There is no advantage to such codes in the binary coding case, and there seems to be none in general (Refs. 19, 20, 21).

**The Szilard-Kraft Inequality.** Let  $w_i$  be the number of bits in the codeword for the  $i$ th message  $m_i$ . Thus for the codebook of eq. (15) one has

$$\begin{aligned} w_1 &= 1 \\ w_2 &= 2 \\ w_3 &= 2. \end{aligned}$$

The smaller the  $w_i$  are, the fewer output bits are required per input symbol. However, if they are too small, the codewords cannot all be different. Thus if  $w_i = 1$  for all  $i$ , the only distinct codewords are 0 and 1, which cannot distinguish three different messages. A condition on the lengths of code words is given by:

**THE SZILARD-KRAFT INEQUALITY.** *Given a set of  $n$  messages, it is possible to assign a codeword of length  $w_i$  to the  $i$ -th message, and to satisfy the prefix condition, if and only if the  $w_i$  satisfy the inequality*

$$(18) \quad \sum_{i=1}^n 2^{-w_i} \leq 1.$$

If the codeword lengths do not satisfy this condition, no decodable code can be constructed (Refs. 22, 23).

**Coding Implications.** Suppose all  $p_i$  are powers of  $1/2$ , so that all information values are integers. Then let  $w_i = I_i = -\log p_i$ . This gives

$$(19) \quad \sum_{i=1}^n 2^{-w_i} = \sum_{i=1}^n 2 \log p_i = \sum_{i=1}^n p_i = 1,$$

which satisfies the constraint of eq. (17). Thus a decodable code can be constructed in which each message has a codeword length in bits equal to its information content in bits. Then the average codeword length, which is the average number of output bits per input message, is

$$(20) \quad \bar{w} = \sum_{i=1}^n p_i w_i = \sum_{i=1}^n p_i I_i = \bar{I} = R.$$

It can be shown that no smaller value of  $\bar{w}$  can be obtained from codeword lengths satisfying eq. (18). This proves the First Binary Coding Theorem for these special cases, with  $\delta$  in the theorem = 0.

**EXAMPLE.** Consider the following set of five messages and their codes.

Message	Probability	Codeword	$I_i$	$w_i$
$m_1$	$\frac{1}{2}$	0	1	1
$m_2$	$\frac{1}{4}$	10	2	2
$m_3$	$\frac{1}{8}$	110	3	3
$m_4$	$\frac{1}{16}$	1110	4	4
$m_5$	$\frac{1}{16}$	1111	5	5

**General Case.** In general the  $I_i$  are not integers, since the  $p_i$  are not powers of  $1/2$ . However, a decodable code can always be constructed in which  $w_i$  is the smallest integer which is greater than or equal to  $I_i$ . Then

$$(21) \quad \begin{aligned} I_i &\leq w_i < I_i + 1, \\ p_i I_i &\leq p_i w_i < p_i I_i + p_i, \\ R = \bar{I} &\leq \bar{w} \leq \bar{I} + 1 = R + 1. \end{aligned}$$

so that the average number of output bits per message is never more than one in excess of the average number of bits per message. This means that if the number of bits per message is large, the percentage excess is small. One can always make the number of bits per message large by coding sequences of input messages, taking all possible sequences of length  $L$  messages as a new message set, containing  $n^L$  different messages.

EXAMPLE. For the codebook of eq. (15) and the source of Fig. 6, the codeword lengths  $w_i$  are given in eq. (17). One can compute  $\bar{w}$ , the average number of bits per source symbol:

$$(22) \quad \begin{aligned} \bar{w} &= \sum_{i=1}^n p_i w_i \\ &= 1 \times 0.755 + 2 \times 0.185 + 2 \times 0.060 = 1.245 \text{ bits/bit.} \end{aligned}$$

Now form all  $3^2 = 9$  possible *pairs* of messages selected by the source of Fig. 6. Using the Huffman coding procedure, as illustrated by the tree

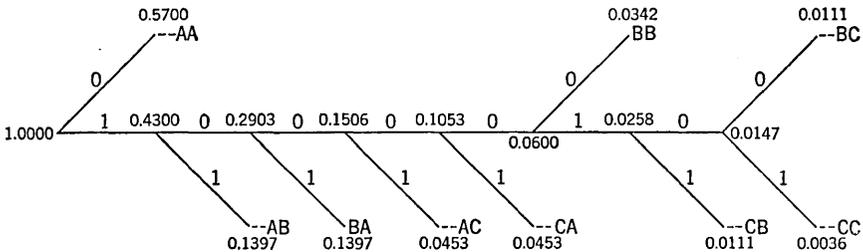


FIG. 7. A coding tree for message pairs.

in Fig. 7, gives the following set of messages, probabilities and codes.

Message	Probability	Code
AA	0.5700	0
AB	0.1397	11
AC	0.1397	101
BA	0.0453	1001
BB	0.0453	10001
BC	0.0342	100000
CA	0.0111	1000011
CB	0.0111	10000100
CC	0.0036	10000101

Evaluating eq. (22) in this case gives  $\bar{w} = 2.0767$ . The average information per message is two bits: for since successive messages are statistically independent, the information in a sequence of two messages is the sum of the informations in each of the two, and the average of a sum is the sum of the averages. The efficiency of this coding in bits per binit is about 0.96. In terms of the inequality of eq. (21) this might be as low as  $\frac{2}{3}$  or as high as 1.00. The fact that the coding here has an efficiency well above its lower bound is typical of coding results. It is due to the fact that the entropy curve in Fig. 5 has a very broad maximum, so that the message set must be quite skewed in probabilities before the efficiency of coding drops very low.

**General Case Continued.** Define  $\bar{w}_L$  as the average number of binites required to code a block of  $L$  source symbols. Rewriting inequality (21) for the new message set gives

$$LR \leq \bar{w}_L \leq LR + 1, \quad (23)$$

$$R \leq \frac{\bar{w}_L}{L} \leq \frac{R + 1}{L}.$$

Now  $\bar{w}_L/L$  is the average number of output binites per original input message, so that the last line of eq. (23) satisfies the requirements of the First Binary Coding Theorem for any  $L > 1/\delta$ . This proves the theorem in the general case of a message source with arbitrary information distribution. It also justifies a definition which assigns the same rate to the two very different sources of Figs. 3 and 4, if these sources can be controlled.

**Controlled Source Coding.** The source of Fig. 4 may be controlled to read out one binit per second. The source of Fig. 3 may be controlled so that its coded output produces one binit per second. This will require that the source be speeded up when it generates A's, and slowed down when it generates B's or C's, but *on the average* it will be generating very nearly a symbol per second. The average rate of the source, then, determines the communications facilities required to transmit its encoded output. The differences in the distributions of Figs. 3 and 4 affect only the amount of delay and the size of the codebook required for efficient coding into binites. This result holds for any controlled source, whose *symbol rate* may be varied in order to keep its *information rate* constant.

**Uncontrolled Sources.** Here the source generates messages at a constant rate. If there is any variance in its information distribution, the rate at which it generates information will then fluctuate. In an efficient code,  $w_i$ , the number of output binites in the codeword for the message  $m_i$ , is still close to the message self-information  $I(m_i)$ . The average number

of bits per message  $\bar{w}_L$  is still near to the source rate  $R$ . But it is *possible* (though highly improbable) that all  $L$  of the messages in a block will be those of maximum self-information. Therefore it is *not* possible to transmit *all* message sequences as they come along unless a channel is used which can transmit bits at a rate equal to  $w_{\max}$  times the (fixed) rate at which the uncontrolled source generates messages. Here  $w_{\max}$  is the largest of the  $w_i$ , i.e., the length of the longest codeword.

**Minimax Coding.** If all the message sequences generated by an uncontrolled source are to be transmitted unambiguously, the best code to use is *not* the Huffman code, which minimizes  $\bar{w}$  but will have a large  $w_{\max}$  if the information distribution has appreciable variance. Rather it is better to use a code that minimizes the value of  $w_{\max}$ , a minimax problem with a simple solution (Ref. 24). The codewords are taken of uniform length  $w_m$ , where  $w_m$  is the integer such that

$$(24) \quad 2^{w_m-1} < n_m \leq 2^{w_m},$$

or

$$w_m - 1 < \log_2 n_m \leq w_m.$$

Here  $n_m$  is the number of messages. Since there are  $2^{w_m}$  codewords of length  $w_m$ , there are at least enough to label all the messages. *Note* that this coding procedure is quite independent of the message probability or information distribution and depends only on the number of messages in the set.

**Efficient Coding: Uncontrolled Source.** The average rate  $R$  at which a source generates information still has significance when the source is uncontrolled. Not *all* sequences of  $L$  messages can be coded into about  $LR$  bits, but *almost* all of them can. More precisely we have:

**SECOND BINARY CODING THEOREM (Uncontrolled Sources).** *Given a discrete message source which generates information at an average rate  $R$  bits per message, and given any  $\delta > 0$  and any  $\epsilon > 0$ , it is possible to construct a representation of sequences of messages as sequences of binary symbols so that, for each message sequence, less than  $R + \delta$  output binary symbols are required per input symbol from the source, except for a set of message sequences whose total probability is less than  $\epsilon$ .*

The procedure is to code the messages in blocks of length  $L$ , coding each block into a codeword of length  $w_i \leq I_i + 1$ . The theorem follows because the self-information of a sequence of messages is the sum of the self-informations of the component messages (since statistical independence is assumed), and because the sum of a large number  $L$  of identically distributed, statistically independent random variables is very likely to be very near, percentagewise, to  $L$  times the mean of the distribution.

**Sums of Random Variables.** The Second Binary Coding Theorem follows from the weak law of large numbers. Stronger results derive from the Tehebysheff inequality and the central limit theorem. These three results applied to self-information are:

1. **WEAK LAW OF LARGE NUMBERS.** For any  $\epsilon > 0$  and any  $\delta > 0$ , an integer  $L_0$  can be found so large that the probability that a sequence of  $L > L_0$  messages will have an amount of self-information greater than  $L(\bar{I} + \delta)$  is  $< \epsilon$ .

2. **TCHEBYSHEFF INEQUALITY.** For any  $\delta > 0$ , the probability that a sequence of  $L$  messages has an amount of self-information greater than  $L(\bar{I} + \delta)$  is  $< \epsilon = \sigma_I^2 / L\delta^2$ .

3. **CENTRAL LIMIT THEOREM.** For any  $\delta > 0$ , the probability that a sequence of  $L$  messages has an amount of self-information greater than  $L[\bar{I} + (\delta/\sqrt{L})]$  is asymptotically given by the expression

$$(25) \quad \epsilon = \frac{1}{\sqrt{2\pi L\sigma_I^2}} \int_{x=\delta/\sqrt{L}}^{\infty} e^{-x^2/2L\sigma_I^2} dx = \frac{1}{\sqrt{2\pi\sigma_I^2}} \int_{y=\delta}^{\infty} e^{-y^2/2\sigma_I^2} dy.$$

Here  $\bar{I}$  is the mean and  $\sigma_I^2$  is the variance of the self-information distribution of the messages.

**Coding Interpretations.** Each of these results translates into a result for efficient coding, since by eq. (21) it is possible to assign distinct binary codewords to all possible message sequences of length  $L$  so that the difference between codeword length in bits and information in bits is less than unity for each sequence. Thus to every sequence in a set of sequences of total probability  $\geq 1 - \epsilon$ , we can certainly assign codes of length  $< L(\bar{I} + \delta) + 1$ . The remaining sequences, of total probability  $\leq \epsilon$ , may all be assigned the same codeword, and will cause ambiguity or error a fraction  $\epsilon$  of the time.

**Storage and Delay.** The central limit theorem shows that for fixed error probability  $\epsilon$ , the difference between  $\bar{w}_L/L$ , the bits per message, and  $\bar{I}$ , the bits per message, decreases with blocklength  $L$  only like  $1/\sqrt{L}$ . This implies that it may be necessary to use much longer blocks to get efficient coding for an uncontrolled source than would be required for a controlled source, for which the difference between  $\bar{w}_L/L$  and  $\bar{I} = R$  decreases like  $1/L$ , as in eq. (23).

**Effective Number of Messages.** From the weak law of large numbers, there is a set of message sequences of total probability  $> 1 - \epsilon$ , each sequence of which has self-information within  $\pm L\delta$  of  $L\bar{I}$ . Each sequence in this set then has probability in the range

$$(26) \quad 2^{L(\bar{I} \pm \delta)},$$

and the total number of sequences in the probable set, for large  $L$ , lies

in the range

$$(27) \quad (1 \pm \epsilon) 2^{L(\bar{I} \pm \delta)},$$

no matter how small  $\epsilon$  and  $\delta$ . Adding unity to  $L$  ultimately multiplies the number of messages in the probable set by  $2^{\bar{I}}$ . This is what would happen if there were just

$$(28) \quad n_{\text{eff}} = 2^{\bar{I}}$$

different messages in the set, and  $n_{\text{eff}}$  is therefore called the *effective number of messages*, or the *effective alphabet size*. Since  $\bar{I} \leq \log n$ , we have

$$(29) \quad n_{\text{eff}} \leq n.$$

That is, the fact that all messages are not equiprobable produces a growth in the probable sequence set *as if* a smaller equiprobable message set were being used.

#### 4. MORE COMPLICATED DISCRETE SOURCES

Most natural sources are more complicated than those discussed in Sect. 3. A more general source is a random process which generates sequences of symbols like the letters of English text, in which each symbol is selected with a probability which depends on the values of the preceding symbols.

**Joint Probabilities.** Consider a set  $S$  of  $n$  different symbols,

$$S = \{s_i\}, \quad 1 \leq i \leq n.$$

Let  $x_k$  be the symbol selected at (integer) time  $k$ ,  $-\infty < k < +\infty$ . Then  $x_k$  is a *random variable*, taking values from the set  $S = \{s_i\}$ . The random process is well defined if the joint probabilities

$$(30) \quad \begin{aligned} & p_1(x_k) \\ & p_2(x_k, x_{k-1}) \\ & \vdots \\ & p_{j+1}(x_k, x_{k-1}, \dots, x_{k-j}), \end{aligned}$$

are known for all combinations of  $x$  values and all values of  $j$ . It will be assumed that the process is *stationary* (and indeed ergodic), so that the probabilities are independent of the time index  $k$ .

**Conditional Probabilities.** Knowledge of the joint probabilities of eq. (30) is equivalent to knowledge of the conditional probabilities

$$(31) \quad \begin{aligned} & q_0(s_i) \\ & q_1(s_i | x_{k-1}) \\ & \vdots \\ & q_j(s_i | x_{k-1}, x_{k-2}, \dots, x_{k-j}). \end{aligned}$$

The two sets of probabilities of eqs. (30) and (31) are related by

$$(32) \quad \begin{aligned} q_0(s_i) &= p_1(s_i) \\ q_j(s_i | x_{k-1}, x_{k-2}, \dots, x_{k-j}) p_j(x_{k-1}, x_{k-2}, \dots, x_{k-j}) \\ &= p_{j+1}(s_i, x_{k-1}, x_{k-2}, \dots, x_{k-j}). \end{aligned}$$

### Markov Sources

If for some integer  $N$  and all integers  $j > 0$

$$(33) \quad q_{N+j}(s_i | x_{k-1}, x_{k-2}, \dots, x_{k-N-j}) = q_N(s_i | x_{k-1}, x_{k-2}, \dots, x_{k-N}),$$

so that a knowledge of  $N$  preceding symbols gives all the probabilistic information available about the next symbol value, then the process is a *multiple Markov process* of order  $N$  (Ref. 52). When  $N = 1$ , the process is called a *simple Markov process*.

**Self-Information.** If the process is Markov of order  $N$ , then the self-information provided when the symbol  $s_i$  occurs is the negative logarithm of its probability, but this is now a conditional probability depending on the values of the preceding  $N$  symbols. The self-information is thus a random function of the  $N$  random variables  $x_{k-1}, x_{k-2}, \dots, x_{k-N}$ .

$$(34) \quad I_N(s_i | x_{k-1}, x_{k-2}, \dots, x_{k-N}) = -\log q_N(s_i | x_{k-1}, x_{k-2}, \dots, x_{k-N}).$$

**Average Self-Information.** A self-information of the symbol  $s_i$  which is not a random function may be obtained by averaging eq. (34) over the conditional probability  $r_N(x_{k-1}, x_{k-2}, \dots, x_{k-N} | s_i)$  that when  $s_i$  occurs at time  $k$ , the preceding  $N$  symbols will have the values  $x_{k-1}, x_{k-2}, \dots, x_{k-N}$ . By *Bayes's theorem*,

$$(35) \quad r_N(x_{k-1}, x_{k-2}, \dots, x_{k-N} | s_i) = \frac{p_{N+1}(s_i, x_{k-1}, x_{k-2}, \dots, x_{k-N})}{p_1(s_i)}$$

and

$$(36) \quad \begin{aligned} I_N(s_i) &= - \sum_{x_{k-1} \in S} \cdots \sum_{x_{k-N} \in S} r_N(x_{k-1}, x_{k-2}, \dots, x_{k-N} | s_i) \\ &\quad \times \log q_N(s_i | x_{k-1}, x_{k-2}, \dots, x_{k-N}) \end{aligned}$$

is defined as the *average self-information of the symbol  $s_i$*  for an  $N$ th order Markov process.

**Source Rate.** The *average rate  $R$*  at which a Markov source generates information is equal to the average  $\bar{I}_N$  of the  $\bar{I}_N(s_i)$  over the probabilities of the symbols  $p_1(s_i)$ . This gives the *average self-information per symbol of the process*:

$$(37) \quad R = \bar{I}_N = \sum_{i=1}^n p_1(s_i) I_N(s_i).$$

**Other Sources**

If the source is not a Markov process of finite order, the self-information of a symbol may not be well defined, since it may then be a function of an infinite number of random variables. However, the quantity  $I_N(s_i|x_{k-1}, x_{k-2}, \dots, x_{k-N})$  given by eq. (34) is still defined for each  $N$ , and it is called the  $N$ th order conditional self-information of  $s_i$ : the quantity  $I_N(s_i)$  defined by eq. (36) is called the  *$N$ -th order average self-information of the symbol  $s_i$* . It can be shown that for any process, for all  $s_i$  and any  $N$ ,

$$(38) \quad \begin{aligned} I_N(s_i) &\geq 0 \\ I_N(s_i) &\geq I_{N+1}(s_i); \end{aligned}$$

the average information provided by the occurrence of  $s_i$  when the  $N$  preceding symbols are known is a monotone decreasing function of  $N$ . Further knowledge of the past, on the average, makes whatever symbol happens next more probable, and therefore less informative.

**Upper Bounds on Source Rate.** If an information source has a measured first-order probability distribution  $p_1(s_i)$ , the average self-information of each of its symbols is at most equal to  $-\log p_1(s_i)$ , the value it would have if successive symbols were statistically independent. If the source has a given conditional distribution of order  $N$ , the average self-information of each symbol is bounded above by the average self-information  $I_N(s_i)$  of a Markov process of order  $N$  with the same  $N$ th order conditional distribution. Again, statistical dependence beyond what is contained in the given distribution can only reduce the average self-information of each symbol.

**Self-Information of a Symbol.** From eq. (38) it follows that one can define a limit,

$$(39) \quad I(s_i) = \lim_{N \rightarrow \infty} I_N(s_i),$$

which converges for each  $s_i$  to a non-negative number. The limit,  $I(s_i)$ ,

is the average self-information added by the occurrence of the symbol  $s_i$  when all the preceding symbols are known.

Note that no general lower bound better than 0 can be obtained. A process which looks random on the basis of  $N$ th order statistics may always be deterministic on the basis of statistics of order  $N + 1$  and have an average rate of zero.

**Source Rate.** The average self-information  $\bar{I}$  of the process is again equal to its rate  $R$ , and is given by either of the two expressions

$$\begin{aligned} \bar{I} &= \sum_{s_i \in S} p_1(s_i) I(s_i) \\ (40) \qquad &= \lim_{N \rightarrow \infty} \bar{I}_N, \end{aligned}$$

where  $\bar{I}_N$ , the average  $N$ th order self-information of the process, is given by eq. (37).

**More General Sources.** The only type of source of greater generality than the multiple Markov process of finite order which has been studied in detail is called a *finite-state* source (Ref. 25). *Note.* A finite state source (a) includes the Markov processes of finite order but it is not included among them and (b) is still not general enough to generate all and only the grammatical sentences in English word by word (Refs. 26, 27). Because of the complexity of natural sources like written language, indirect methods must be used to estimate their rate. A straightforward application of the definitions involves the measurement of probability distributions of order so high that in all the written English there would be too small a sample for an accurate estimate.

### Coding and Delay

The first binary coding theorem still applies to more complicated sources. It carries over unaltered to any discrete ergodic source, whether it be multiple Markov, finite state, or still more general. This may be shown by two coding methods.

**1. Block Coding.** Segment the source output into blocks of length  $L$ , and code each block from a fixed codebook containing a binary sequence for each of the  $n^L$  possible sequences of source symbols. Each source sequence may be coded as before into a number of binitis at most one greater than its total information content in bits.

The average self-information in a block of  $L$  source symbols is equal to  $\bar{I}_0$ , the average self-information of the first symbol in the block when no past history is known, plus  $\bar{I}_1$ , the average self-information of the second symbol when the first is known, etc., the last term being  $\bar{I}_{L-1}$ , the average

self-information of the  $L$ th symbol when all preceding  $L - 1$  are known. Averaging over all sequences gives  $w_L$ , the average number of bits per block of  $L$  input symbols, as bounded by

$$(41) \quad \sum_{j=0}^{L-1} \bar{I}_j \leq w_L \leq \sum_{j=0}^{L-1} \bar{I}_j + 1.$$

Dividing eq. (41) by  $L$  gives the average number of bits per source symbol:

$$(42) \quad \frac{1}{L} \sum_{j=0}^{L-1} \bar{I}_j \leq \frac{w_L}{L} \leq \frac{1}{L} \sum_{j=0}^{L-1} \bar{I}_j + \frac{1}{L}.$$

The summation in eq. (42) is the average of the first  $L$  average self-informations of the process. Since  $\bar{I}_j$  cannot increase with  $j$ , the average will in general be greater than the limit  $I = R$ , but it will approach the limit as  $L \rightarrow \infty$ . Given any  $\delta > 0$ , it is always possible to find an  $L$  so large that the first coding theorem is satisfied, but the required  $L$  may be very large even if the process is Markov of small order. *Example.* A simple Markov process (of order one) has  $\bar{I}_0 = 10$  bits per symbol, and  $\bar{I}_1 = R = \frac{1}{10}$  bit per symbol. It will take  $L = 100$  to code the output of this source at 50 per cent efficiency, i.e., at two output bits per input bit.

**2. Conditional Coding.** A more complicated but more efficient procedure is conditional coding. Here blocks of  $L_2$  symbols are encoded by using one of  $n^{L_2}$  codebooks. The codebook to be used is determined by the preceding  $L_1$  symbols. Such a process has an encoding delay of  $L = L_1 + L_2$  input symbols, and an average number of bits per source symbol given by

$$(43) \quad \frac{1}{L_2} \sum_{j=L_1}^{L-1} \bar{I}_j \leq \frac{w_{L_2}}{L_2} \leq \frac{1}{L_2} \sum_{j=L_1}^{L-1} \bar{I}_j + \frac{1}{L_2}.$$

**EXAMPLE.** In the simple Markov example given for block coding, conditional coding will give better than 50 per cent efficiency for  $L_1 = 1$ ,  $L_2 = 10$ ,  $L = 11$ , which is much less delay and a much smaller codebook than is required by the  $L = 100$  above for block coding.

**Correlated Information Values.** In extending the second binary coding theorem to more complicated uncontrolled sources, an additional kind of delay arises. The occurrence of a symbol with self-information value above the mean may make it more probable that the succeeding symbol will also have self-information above the mean. Then it will be necessary to encode much larger blocks of symbols in order to make it highly probable that the percentage deviation of the self-information of the block from its mean value will be small. Mathematically the problem

becomes one of adding *correlated* rather than statistically independent random variables, and convergence to the mean may be much slower. The second binary coding theorem itself extends to a very broad class of sources, but the Tchebysheff inequality and the central limit theorem do not apply in the form given above.

## 5. DISCRETE NOISELESS CHANNELS

**Type I. Channels.** The *simplest noiseless channel* is one in which each of the set  $S = \{s_i\}$  of symbols which may be applied as an input to the channel is received unaltered at its output. Since there is a one-to-one correspondence between input and output symbols, they may be identified and called *channel symbols*. If all channel symbols are of equal duration, their number  $n$  and their common duration  $t$  completely specify the channel. Such a channel will be called a channel of type I.

**Channel Capacity.** The capacity of a noiseless channel is the maximum average rate at which information can be received over it. The capacity may be measured in bits per symbol, denoted by  $C$ , or in bits per second, denoted by  $C_t$ . If the common symbol duration is  $t$  seconds, then

$$(44) \quad tC_t = C.$$

For a type I channel, inequality (38) and the following discussion show that statistical dependence between successive symbols cannot increase the rate of transmission. Therefore the capacity can be computed by assuming statistical independence and maximizing the average rate  $R$  in bits per symbol.

$$(45) \quad R = - \sum_{s_i \in S} p_1(s_i) \log p_1(s_i)$$

with respect to variations in the probability distribution  $p_1(s_i)$ . But this rate is just the entropy of the  $p_i(s_i)$  distribution, which has the maximum value  $\log n$ , attained when all  $p_1(s_i)$  are equal to  $1/n$ . Thus for a type I channel,

$$(46) \quad \begin{aligned} C &= \log n \text{ bits per symbol,} \\ C_t &= (1/t) \log n \text{ bits per second.} \end{aligned}$$

**Redundancy.** If a source is connected to a type I channel and selects channel symbols with unequal probabilities or with statistical dependence, the rate  $R = I$  at which it generates information will be less than the capacity  $C$  of the channel. The difference  $C - R$  is defined as the (*absolute redundancy*), in bits per symbol, of the source *with respect to the channel*. The ratio of absolute redundancy to channel capacity, a number between

0 and 1, is defined as the *relative* redundancy of the source with respect to the channel. In terms of the interpretation at the end of Sect. 3 of  $R = \bar{I}$  as the logarithm of the effective alphabet size of the source, the redundancy is a measure of the reduction in logarithm of the size of the effective alphabet, due to nonoptimum utilization of the channel.

**Type II Channels.** A *more complicated noiseless channel*, which will be called type II, has a different duration  $t_i$  for each channel symbol  $s_i$ . It is again true that capacity is attained by using symbols chosen with statistical independence and maximizing the average rate  $R_t$  (now in bits per *second*) with respect to the symbol probabilities  $p_1(s_i)$ . But the rate is now given by

$$(47) \quad R_t = \frac{- \sum_{s_i \in S} p_1(s_i) \log p_1(s_i)}{\sum_{s_i \in S} p_1(s_i) t_i},$$

and the maximization leads to the condition that the instantaneous rates at which each symbol transmits information all be equal to the channel capacity  $C$ . Thus

$$(48) \quad C_t = \frac{I(s_i)}{t_i} = - \left( \frac{1}{t_i} \right) \log p_1(s_i),$$

or

$$p_1(s_i) = 2^{-C_t t_i}.$$

The capacity  $C_t$  is determined, for given durations  $t_i$ , by the normalization requirement that the probabilities of the symbols sum to one; this gives

$$(49) \quad \sum_{s_i \in S} p_1(s_i) = \sum_{s_i \in S} 2^{-C_t t_i} = 1,$$

and  $C_t$  is the (unique) real root of this equation. Redundancy with respect to a channel of type II is defined as it was for a channel of type I, using  $R_t$  and  $C_t$  rather than  $R$  and  $C$ . Notice that a source is redundant with respect to a type II channel unless the probabilities with which it chooses symbols are *unequal*, and are given by eq. (48).

**Type III Channels.** Shannon (Ref. 28) discusses a *finite-state channel*, which will be called type III. Here the symbols are of different durations, and the alphabet of symbols available at each instant depends on the preceding symbols which have been sent over the channel. An expression for the capacity of such a channel has been given (*op. cit.*). Type III channels have storage. They will not be discussed here, except to note

that for each such channel there is a corresponding finite-state source, which has no redundancy with respect to the given channel. This optimizing source no longer selects successive symbols with statistical independence.

**Noiseless Channel Coding Theorems.** The capacity  $C$  of a noiseless channel is a rate at which some particular optimizing source can transmit information over the channel. Since the source whose output may need transmission will not usually be an optimal source, this is not a justification for considering  $C$  to be an important channel parameter. The justification is that given a channel of capacity  $C$ , any source of rate  $R < C$ , and no source of rate  $R > C$  may be so encoded as to permit reliable transmission over the channel.

Binary coding theorems one and two can be interpreted as showing how any source can be coded into a binary noiseless channel of type I. These theorems can be generalized.

#### NOISELESS CHANNEL CODING THEOREMS

*I. Controlled Source.* Given a discrete controlled source and a discrete noiseless channel which has capacity  $C_t$  bits per second, it is possible to control the source to any average rate  $R_t < C_t$  and to encode its output for unambiguous reception over the channel. This is not possible for any  $R_t > C_t$ .

*II. Uncontrolled Source.* Given: a discrete uncontrolled source of type I, II, or III with average rate  $R_t$  bits per second; a discrete noiseless channel of capacity  $C_t$  bits per second; and any  $\delta > 0$ . If  $R_t < C_t$ , it is possible to encode sequences of source symbols for transmission over the channel so that the probability that such a sequence will be incorrectly decoded is  $< \delta$ . This is not possible if  $R_t > C_t$ .

### 6. DISCRETE NOISY CHANNELS. I. DISTRIBUTION OF INFORMATION

#### Mutual Information

Let  $x$  and  $y$  denote two related events, and let  $x, y$  denote the event which is their joint occurrence. Let  $\text{Prob } \{x\}$ ,  $\text{Prob } \{y\}$ , and  $\text{Prob } \{x, y\}$  be the associated probabilities.

The *self-information* given by the occurrence of  $x$  is defined in eq. (5) as

$$(50) \quad I(x) = -\log \text{Prob } \{x\}.$$

If  $y$  is now observed, and  $x$  and  $y$  are not statistically independent, the probability of  $x$  a priori will be changed a posteriori to

$$(51) \quad \text{Prob } \{x|y\} = \frac{\text{Prob } \{x, y\}}{\text{Prob } \{y\}}.$$

This change in the probability of  $x$  changes the amount of information required to select it to

$$(52) \quad I(x|y) = -\log \text{Prob} \{x|y\} = -\log \frac{\text{Prob} \{x, y\}}{\text{Prob} \{y\}}.$$

The difference between eqs. (50) and (52) measures how the amount of information required to select  $x$  is changed by the knowledge of  $y$ . This difference is denoted by  $I(x; y)$ , the amount of *mutual information* between  $x$  and  $y$ . Then

$$(53) \quad \begin{aligned} I(x; y) &= -\log \text{Prob} \{x\} + \log (\text{Prob} \{x, y\}/\text{Prob} \{y\}) \\ &= -\log \text{Prob} \{y\} + \log (\text{Prob} \{x, y\}/\text{Prob} \{x\}) \\ &= \log \frac{\text{Prob} \{x, y\}}{\text{Prob} \{x\} \text{Prob} \{y\}}. \end{aligned}$$

Mutual information is measured in the same units as self-information (see Sect. 3).

**Properties.**

1.  $I(x; y)$  is symmetric:

$$(54) \quad I(x; y) = I(y; x).$$

This follows from the last line of eq. (53), and justifies the name “mutual information.”

2.  $I(x; y)$  vanishes if  $x$  and  $y$  are statistically independent. If not, there is a decomposition generalizing eq. (7):

$$(55) \quad I(x, y) = -\log \text{Prob} \{x, y\} = I(x) + I(y) - I(x; y),$$

showing that  $I(x; y)$  plays the role of a correlation (Ref. 2). If  $\text{Prob} \{x, y\}$  is greater than  $\text{Prob} \{x\} \text{Prob} \{y\}$  then  $I(x; y)$  is positive.

3.  $I(x; y)$  may be positive or negative, but cannot be greater than the self-information of  $x$  or  $y$ :

$$(56) \quad \begin{aligned} I(x; y) &\leq I(x) \\ &\leq I(y). \end{aligned}$$

This follows from eq. (53), since the conditional probabilities are at most unity, and have nonpositive logarithms (Refs. 14, 29).

**Notation.** Any  $I$  function whose argument contains no semicolons is interpreted as the negative logarithm of the probability of its argument: thus  $I(x|y) = -\log \text{Prob} \{x|y\}$ . Any  $I$  function whose argument contains a semicolon between two sets of variables is interpreted as in eq. (53), where  $x$  and  $y$  stand for the expressions to the left and right of the semicolon, and  $x, y$  stands for their conjunction.

### Distribution of Mutual Information

Mutual information measures how much information one symbol provides *about another*. It can be used in the discussion of Sect. 5 on sources which generate sequences of related symbols. Here it will be applied only to the discussion of noisy channels.

**Discrete Noisy Channel.** Consider a simple noisy channel, as illustrated in Fig. 2. There is a set  $U = \{u_i\}$ ,  $1 \leq i \leq n_u$ , of symbols which may be transmitted, and a set  $V = \{v_j\}$ ,  $1 \leq j \leq n_v$ , of symbols which may be received. It will be assumed throughout that the channel is without storage and that it and the source are stationary: the probability of transmitting the symbol  $u_i$  and receiving the symbol  $v_j$  is independent of time and of prior transmissions and receptions. Let  $x_k$  and  $y_k$  be the transmitted and received symbols at (integer) time  $k$ ,  $x_k \in U$  and  $y_k \in V$ . Then the pair  $(x_k, y_k)$  is a stationary random variable, taking values from the set  $U \times V = \{u_i, v_j\}$  of ordered pairs of transmitted and received symbols, with probabilities

$$(57) \quad \text{Prob} \{x_k = u_i, y_k = v_j\} = p(u_i, v_j).$$

Denote the first order probabilities by

$$(58) \quad p(u_i) = \sum_{j=1}^{n_v} p(u_i, v_j), \quad q(v_j) = \sum_{i=1}^{n_u} p(u_i, v_j)$$

and the conditionals by

$$(59) \quad p(u_i|v_j) = \frac{p(u_i, v_j)}{q(v_j)}, \quad q(v_j|u_i) = \frac{p(u_i, v_j)}{p(u_i)}.$$

**Mutual Information of a Noisy Channel.** The amount of information given by  $y_k$  about  $x_k$  is also a stationary random variable, which takes values equal to the numbers  $I(u_i; v_j)$  with probabilities  $p(u_i, v_j)$ .

The distribution of this random variable is completely determined by  $p(u_i, v_j)$ , through eqs. (53) and (58). Its most important parameter is its mean value, the average rate  $R$  at which the received symbols give information *about the transmitted symbols*:

$$(60) \quad \begin{aligned} R &= \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} p(u_i, v_j) I(u_i; v_j), \\ &= \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} p(u_i, v_j) \log \frac{p(u_i, v_j)}{p(u_i)q(v_j)}. \end{aligned}$$

Although for particular  $u_i, v_j$  the mutual information may be negative, the average  $R$  of eq. (60) is always positive.

EXAMPLE 1. *The Binary Erasure Channel.* As illustrated in Fig. 8, this channel accepts two input symbols, 0 and 1, and produces three output symbols, 0, 1, and  $X$ . With probability  $p$  its output reproduces its input;

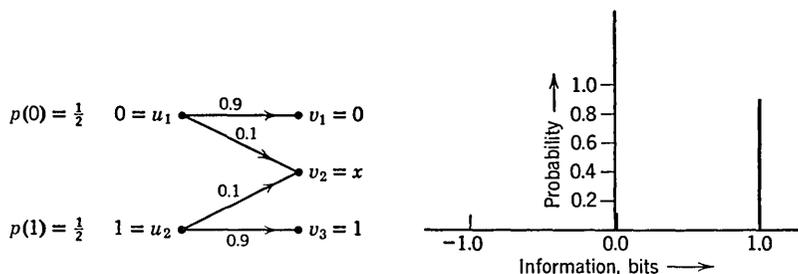


FIG. 8. Binary erasure channel and mutual information distribution.

with probability  $q$ , its input is erased and an output  $X$  indicates the erasure; 0 and 1 are transmitted with equal probability.

When a 0 or a 1 is received, the mutual information is

$$\begin{aligned}
 (61) \quad I(0; 0) &= \log \frac{\text{Prob} \{u_1, v_1\}}{\text{Prob} \{u_1\} \text{Prob} \{v_1\}} = \frac{\text{Prob} \{v_1|u_1\}}{\text{Prob} \{v_1\}} \\
 &= \frac{\log p}{p/2} = \log 2 = 1 \text{ bit} = I(1; 1).
 \end{aligned}$$

When an  $X$  is received,

$$\begin{aligned}
 (62) \quad I(0; X) &= \log \frac{\text{Prob} \{v_2|u_1\}}{\text{Prob} \{v_2\}} = \log \frac{q}{q} = \log 1 = 0 \\
 &= I(1; X).
 \end{aligned}$$

This gives the information distribution shown in Fig. 8, with the average value

$$\begin{aligned}
 (63) \quad R &= \sum_{i=1}^2 \sum_{j=1}^3 p(u_i, v_j) I(u_i; v_j) \\
 &= (p/2)I(0; 0) + (q/2)I(0; X) + (p/2)I(1; 1) + (q/2)I(1; X) \\
 &= 2(p/2) + 2(q/2) \times 0 = p.
 \end{aligned}$$

EXAMPLE 2. *The Binary Symmetric Channel.* As illustrated in Fig. 9, this channel also accepts the two input symbols 0 and 1, but it only produces the same two output symbols. With probability  $p$  its output reproduces

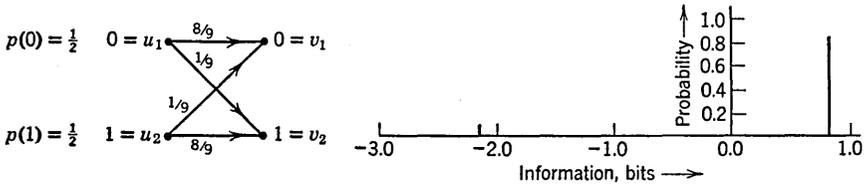


FIG. 9. Binary symmetric channel and information distribution.

its input: with probability  $q = 1 - p$  its output is the incorrect symbol. 0 and 1 are transmitted with equal probability, and  $q$  is  $< 1/2$ .

When the correct symbol is received, the mutual information is

$$(64) \quad I(0; 0) = \log \frac{\text{Prob} \{v_1 | u_1\}}{\text{Prob} \{v_1\}} = \log p / (\frac{1}{2}) = \log 2p > 0$$

$$= I(1; 1).$$

When an error is made,

$$(65) \quad I(0; 1) = \log \frac{\text{Prob} \{v_2 | u_1\}}{\text{Prob} \{v_2\}} = \log q / (\frac{1}{2}) = \log 2q < 0.$$

This gives the mutual information distribution illustrated for  $q = 1/9$  in Fig. 8, with the average rate

$$(66) \quad R = p \log 2p + q \log 2q = p + q + p \log p + q \log q$$

$$= 1 - H(p, q),$$

where  $H(p, q)$  is the entropy function illustrated in Fig. 5. For  $q = 1/9$ ,

$$H(p, q) = H(\frac{8}{9}, \frac{1}{9}) = 0.5032,$$

and

$$R = 0.4968 \text{ bit per symbol.}$$

**Averages of Information Measures.** In addition to the average rate  $R$ , other averages of information measures must be considered.

**Notation.** An average of an information function  $I(u_i; v_j)$  over the joint distribution  $p(u_i, v_j)$  is denoted by replacing the names " $u_i$ " and " $v_j$ " of the symbols by the names " $U$ " and " $V$ " of the sets from which the symbols are selected. A single capital denotes an average over a

univariate distribution. Thus

$$\begin{aligned}
 I(U; V) &= R = \sum_{i=1}^{n_p} \sum_{j=1}^{n_u} p(u_i, v_j) I(u_i; v_j) \\
 I(U) &= H(\{p(u_i)\}) = \sum_{i=1}^{n_u} p(u_i) I(u_i) \\
 (67) \quad I(U|V) &= \sum_{i=1}^{n_u} \sum_{j=1}^{n_p} p(u_i, v_j) I(u_i|v_j) \\
 &= - \sum_{i=1}^{n_u} \sum_{j=1}^{n_p} p(u_i, v_j) \log p(u_i|v_j).
 \end{aligned}$$

**Equivocation.** The average rate at which information about the transmitted symbols is supplied to the channel is the average self-information of the transmitted symbols. This by eq. (56) is greater than the average rate at which such information is received. The difference is

$$\begin{aligned}
 I(U) - I(U; V) &= - \sum_{i=1}^{n_u} p(u_i) \log p(u_i) - \sum_{i=1}^{n_u} \sum_{j=1}^{n_p} p(u_i, v_j) \log \frac{p(u_i|v_j)}{p(u_i)} \\
 &= - \sum_{i=1}^{n_u} \sum_{j=1}^{n_p} p(u_i, v_j) \log p(u_i) \\
 (68) \quad &= - \sum_{i=1}^{n_u} \sum_{j=1}^{n_p} p(u_i, v_j) \log \frac{p(u_i|v_j)}{p(u_i)} \\
 &= - \sum_{i=1}^{n_u} \sum_{j=1}^{n_p} p(u_i, v_j) \log p(u_i|v_j) \\
 &= I(U|V) \geq 0.
 \end{aligned}$$

This quantity is the *conditional entropy* of the set  $U$  given  $V$ . It measures the average amount of information about the transmitted symbol which the receiver still lacks after noisy reception, and thus the average rate at which it would be necessary to transmit additional information over an extra channel in order to make the receiver certain of each transmitted symbol. This quantity is also called the average *equivocation* of the received symbols. Equivocation is present in the channels of Figs. 2a and 2c, but not in Fig. 2b.

**Irrelevance.** The average rate at which the received symbols give information (subject matter unspecified) is the average self-information of the received symbols. From eq. (56), this is also greater than the average rate at which information is received about the transmitted symbols.

The difference may be shown, as in eq. (67), to be

$$\begin{aligned}
 I(V) - I(U; V) &= - \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} p(u_i, v_j) \log q(v_j | u_i) \\
 (69) \qquad \qquad &= I(V | U) \geq 0.
 \end{aligned}$$

This quantity is the conditional entropy of  $V$  given  $U$ . It measures the amount of received information *not relevant* to the transmitted information, but relevant only to the channel noise.

The names "Spread," "Dispersion," and "Prevarication" have all been used for this quantity. "Irrelevance" seems more appropriate in the case, for example, of the channel of Fig. 2b, in which the receiver receives information which is irrelevant but not misleading. The channel of Fig. 2a has no irrelevance, but that of Fig. 2c does.

## 7. DISCRETE NOISY CHANNELS. II. CHANNEL CAPACITY AND INTERPRETATIONS

### The Noisy Channel Specification

Formally a noisy channel without storage is a graph, like those in Figs. 2, 8, and 9, in which each branch connects a transmitted symbol with a received symbol and has a number on it. The numbers are the conditional probabilities  $q(v_j | u_i)$  which define statistically what the channel does to given input symbols. These numbers are fixed for a given channel, but the transmitter is free to decide how to use the input symbols. Only the channel with no storage will be considered.

**Transmitter Strategy.** The transmitter strategy formally is a random process, selected by the transmitter to generate sequences of transmitter symbols for transmission over the noisy channel. If an input message sequence is coded into a sequence of transmitter symbols, the random process which generates the messages and the operation of the coder may be combined to obtain the new random process which is the transmitter strategy.

In Sect. 4, eq. (38) *et seq.*, it was pointed out that the self-information of a symbol can on the average only be reduced by statistical dependence on preceding symbols. The same is true of the mutual information provided by a received symbol about transmitted symbols for a channel with no storage. If the transmitter wants to maximize the average amount of mutual information received, he can do no better than to select successive transmitted symbols independently from some distribution  $p(u_i)$ . Then the problem of choosing a transmitter strategy reduces to the problem of choosing a first order distribution  $p(u_i)$  for the transmitted symbols. Then  $p(u_i)$  and the  $q(v_j | u_i)$  together determine channel operation completely.

**Capacity of a Noisy Channel.** The *channel capacity*  $C$  of a given noisy channel is defined as the maximum value of the transmission rate  $R$  which can be obtained by varying the transmitter strategy. For a channel with no storage, this is the maximum  $R$  which can be obtained by varying  $p(u_i)$ , with  $q(v_j|u_i)$  fixed. Thus from eq. (69),

$$(70) \quad C = \max_{p(u_i)} R = \max_{p(u_i)} \left\{ \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} p(u_i) q(v_j|u_i) \log \frac{q(v_j|u_i)}{q(v_j)} \right\},$$

where the values of  $q(v_j|u_i)$  are held fixed as the  $p(u_i)$  are varied, and the variation of  $q(v_j)$  is determined from the relation

$$(71) \quad q(v_j) = \sum_{i=1}^{n_u} p(u_i) q(v_j|u_i).$$

The maximization is carried out by differentiating eq. (70) for  $R$  with respect to each of the  $p(u_i)$ , subject to the constraint that

$$(72) \quad \sum_{i=1}^{n_u} p(u_i) = 1.$$

However, this maximization may lead to negative values for some of the  $p(u_i)$ . It is then necessary to eliminate one or more of the input symbols by setting its probability at zero, and to maximize again with a smaller input set, until a maximum  $R$  is obtained with all  $p(u_i)$  non-negative (30).

### Interpretations of Capacity

One interpretation of the capacity  $C$  of a noisy channel is provided by its definition. In Example 1, Fig. 8, the binary erasure channel has a rate of transmission  $R = p$  when 0's and 1's are transmitted with equal probabilities, as shown by eq. (63). This is the maximum rate attainable for this channel, and is therefore its capacity,  $C = p$ . One bit of information per symbol is supplied to the channel, and on the average  $p$  bits of information *about the transmitted symbol* are received. But the transmission process is not reliable. Since information is being supplied to the channel at a rate greater than the channel capacity, not all of it can get through, and the channel determines in a random fashion which bits will be lost and which will be saved.

**Feedback Interpretation.** In the binary erasure channel, suppose that the transmitter can look over the receiver's shoulder and can see which of the transmitted symbols have been erased in the channel. This can be accomplished by having a noiseless feedback channel from receiver to transmitter. Every time a transmitted digit is erased the transmitter can repeat it, going on to the next digit as soon as the first unerased version of the preceding one has been received.

The transmitter is now supplying the channel with information at an average rate of  $p$  bits per transmitted symbol. The repeated digits do not give additional information about the message to be transmitted, but only about where erasures have occurred in transmission. The receiver receives information at the same average rate of  $p$  bits per symbol, and receives each message symbol once. Here channel capacity has the interpretation it has in the case of a noiseless channel: it is the maximum average rate at which information can be transmitted reliably over the channel. In fact noiseless channel coding Theorem I: *Controlled Source* (Sect. 5) applies verbatim to this noisy channel.

**Coding Interpretation.** A noiseless feedback channel is not usually available. Fortunately it turns out that it is not needed. It is possible to obtain reliable transmission over a noisy channel, at any rate less than channel capacity, by proper encoding, without making use of any feedback information. This is the primary justification of "capacity" as a significant parameter for a noisy channel. Indeed it is perhaps the most important single justification for the definition of mutual information and the whole structure of information theory. The formal expression of this fact is the:

**NOISY CHANNEL CODING THEOREM.** *Given: a discrete source of type I, II, or III (Sect. 5) with average rate  $R_t$  bits per second; a discrete noisy channel without storage of capacity  $C_t$  bits per second; and any  $\delta > 0$ . If  $R_t < C_t$ , it is possible to encode sequences of source symbols for transmission over the channel so that the probability that such a sequence will be incorrectly decoded is  $< \delta$ . This is not possible if  $R_t > C_t$ .*

**Relation to Noiseless Case.** This theorem is essentially the second Noiseless Channel Coding Theorem, with a few minor modifications. Both of these theorems may be strengthened to give relations between the error probability  $\delta$ , the difference  $C_t - R_t$  between rate and capacity, and the length  $L$  of the sequence of source symbols which must be encoded. Results like the Tehebysheff Inequality and the Central Limit Theorem must be applied to both the source self-information distribution and the channel mutual information distribution. Some work has been done on these problems recently (Refs. 31-34).

**Implications.** The Noisy Channel Coding Theorem shows that a lack of reliability in a channel does not impose a corresponding lack of reliability on the received and decoded messages. This alone is not surprising. For *example* in the binary erasure channel, transmitting at rate  $R = 1/N$  bits per symbol by repeating each message binit  $N$  times for transmission, the error probability per message binit is

$$(73) \quad q^N = q^{(1/R)} = 2^{-(1/R) \log(1/q)},$$

since the receiver can decode each message binit unless all  $N$  repetitions are erased. The probability in eq. (73) can be made arbitrarily small, but only by letting  $R \rightarrow 0$ .

The theorem also states, however, that error probability can be made arbitrarily small *without* decreasing rate  $R$ , so long as  $R < C$ , the channel capacity. This requires proper encoding of long sequences of source symbols.

**Construction of Codes.** There is as yet no analog to the Huffman code for noisy channels. Considerable work has been done in designing codes for the binary symmetric channel of Fig. 9 (Refs. 35-43). However, no simple, explicit procedure has yet been found for transmitting at rates arbitrarily close to channel capacity with arbitrarily small error probability.

The only constructive procedure available transmits at rates less than capacity. However, if the rate is kept fixed, it is possible for the receiver to set the error probability as low as he desires, but this depends on how much delay he is willing to tolerate. This procedure has been discussed for the binary symmetric channel (Ref. 40). It will be illustrated here for the binary erasure channel of Fig. 8.

**Error-Free Coding for the Binary Erasure Channel**

In the binary erasure channel of Fig. 8 the error probability can be reduced by using some of the input binit as information symbols and some as check symbols. Such a coding procedure is illustrated in Fig. 10.

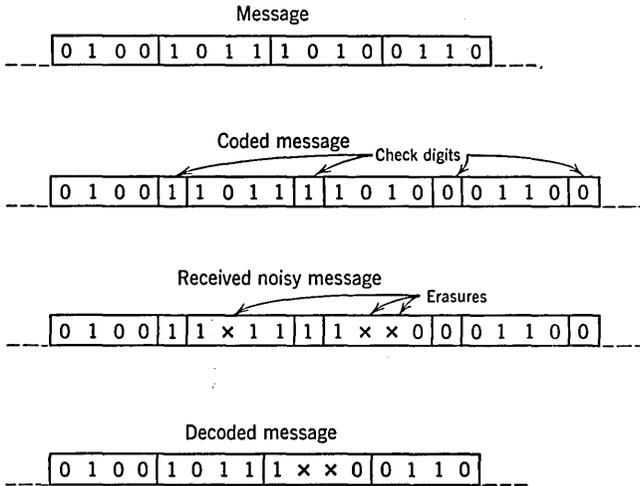


FIG. 10. Parity check coding for the binary erasure channel.

**Parity Check Codes.** Assume that the probability  $q$  of erasure per digit in the channel is 0.05. In Fig. 10, each group of four successive message digits has added to it by the coder a fifth digit for checking. The added digit is selected to be a 0 or a 1 so as to make the total number of 1's in the block of five coded digits even. A check digit of this type is called a *parity check* (Ref. 35). If the channel erases only one digit or none in each block of five, the receiver can correct the erasure by filling it in with a 0 or 1 so as to make the total number of 1's in the block even again.

The channel may erase two or more digits in a single block of five, as shown in the third block in Fig. 10. The receiver cannot correct that block. However, with  $q = 0.05$ , so that on the average only 5 per cent of the transmitted symbols are erased, the receiver will be able to correct more than three-quarters of the erasures. In fact, the average number of erasures per block of  $N = 5$  remaining after correction is equal to  $Nq$ , the average number before correction, less  $Np^{N-1}q$ , the contribution to the average due to blocks containing a single erasure. This gives

$$(74) \quad Nq - Np^{N-1}q = Nq[1 - p^{N-1}] = Nq[1 - (1 - q)^{N-1}] \\ < Nq[1 - (1 - q)^N] < Nq[1 - (1 - Nq)] = (Nq)^2.$$

Thus the number of erasures remaining is reduced from  $Nq$  to  $(Nq)^2$ , by a factor of  $Nq = 5(0.05) = 0.25$ .

**Behavior of First Stage.** In the coding procedure illustrated in Fig. 10, input information is being supplied at an average rate of  $\frac{4}{5} = 0.80$  bit per input binit. The capacity of the channel is  $C = p = 1 - q = 0.95$  bit per symbol. The resultant probability of remaining erasures is  $< \frac{1}{4}$  of the original probability of erasure in the channel. Without reducing the input information rate it is possible to reduce the probability of remaining erasures by iterating this kind of checking procedure. The second step in such an iteration is illustrated by Fig. 11.

**Iteration.** In Fig. 11, the size of the basic block has been increased from 5 to  $N_1 = 10$  digits, the first nine being message digits and the tenth being a parity check on the preceding nine. After nineteen such blocks of ten, the coder adds a twentieth check block. The first digit in the check block is a parity check on the first digits in each of the nineteen preceding blocks, i.e., digits 1, 11, 21,  $\dots$ , 181 in time order. The second digit in the check block checks all the second digits in preceding blocks and so on; the last digit in the check block checks the nine preceding check digits, and it is in fact a parity check on the whole group of  $(10)(20) = N_1N_2 = 200$  digits. This is visualized most easily as in Fig. 10, in which the blocks are aligned below one another, and each digit in the check block

checks the column above it. The last digit in the check block checks both the row to its left and the column above it.

The receiver decodes each row which has no more than a single erasure as soon as the check digit at the end of that row is received. If a row has more than one erasure, it can still be decoded properly when the check block arrives, if each *column* has only one erasure left after the first decoding step.

1 0 1 1 1 0 0 0 1	1 1 0 1 1 1 0 0 0 1	1 1 0 1 1 1 0 0 0 1	1 1 0 1 1 1 0 0 0 1
1 1 0 1 1 0 1 0 0 1	1 1 0 1 1 0 × 0 0 1	1 1 1 0 1 1 0 1 0 0 1	1 1 1 0 1 1 0 1 0 0 1
0 0 1 1 1 1 1 0 0 1	0 0 1 1 1 1 1 0 0 1	1 0 0 1 1 1 1 1 0 0 1	1 0 0 1 1 1 1 1 0 0 1
1 1 1 1 1 0 0 1 0 0	1 1 1 1 × 0 × 1 0 0	0 1 1 1 1 × 0 × 1 0 0	0 1 1 1 1 0 0 1 0 0
0 1 1 0 0 1 0 1 0 0	0 1 1 0 0 1 0 1 0 0	0 0 1 1 0 0 1 0 1 0 0	0 0 1 1 0 0 1 0 1 0 0
0 0 0 1 1 1 0 1 0 0	0 0 0 1 1 1 0 1 0 0	0 0 0 1 1 1 0 1 0 0	0 0 0 1 1 1 0 1 0 0
0 0 1 0 1 1 0 1 1 1	0 0 1 0 1 1 0 1 1 1	1 0 0 1 0 1 1 0 1 1 1	1 0 0 1 0 1 1 0 1 1 1
1 1 1 0 0 0 0 1 0 0	1 1 1 0 0 0 0 1 0 0	0 1 1 1 0 0 0 0 1 0 0	0 1 1 1 0 0 0 0 1 0 0
0 1 1 1 1 0 0 1 0 1	1 0 1 1 1 1 0 0 1 0 1	1 0 1 1 1 1 0 0 1 0 1	1 0 1 1 1 1 0 0 1 0 1
0 1 1 0 1 0 1 1 1 0	0 0 1 1 × 1 0 1 1 1 0	0 0 1 1 0 1 0 1 1 1 0	0 0 1 1 0 1 0 1 1 1 0
1 0 0 1 0 1 1 0 0 0	1 0 0 1 0 1 1 0 0 0	0 1 0 0 1 0 1 1 0 0 0	0 1 0 0 1 0 1 1 0 0 0
0 1 1 1 0 1 0 1 1 0	0 1 1 1 0 1 0 1 1 0	0 0 1 1 1 0 1 0 1 1 0	0 0 1 1 1 0 1 0 1 1 0
0 1 1 0 1 1 1 0 1 0	0 0 1 1 0 1 1 1 0 1	0 0 1 1 0 1 1 1 0 1 0	0 0 1 1 0 1 1 1 0 1 0
1 1 0 0 1 0 1 1 1 0	1 1 1 0 0 1 0 1 1 0	1 1 1 0 0 1 0 1 1 1 0	1 1 1 0 0 1 0 1 1 1 0
1 1 1 1 0 0 1 0 0 1	1 1 1 1 1 0 0 1 0 0	1 1 1 1 1 0 0 1 0 0 1	1 1 1 1 1 0 0 1 0 0 1
0 0 0 1 0 1 1 0 1 0	0 0 0 0 1 0 1 1 × ×	0 0 0 0 1 0 1 1 × ×	0 0 0 0 1 0 1 1 0 1 0
0 0 1 0 0 1 0 1 1 0	0 0 1 0 0 1 0 1 1 0	× 0 0 1 0 0 1 0 1 1 0	0 0 0 1 0 0 1 0 1 1 0
0 1 0 1 1 1 0 1 0 1	1 0 1 0 1 1 1 × 1 0 1	1 0 1 0 1 1 1 0 1 0 1	1 0 1 0 1 1 1 0 1 0 1
0 1 0 0 1 0 0 1 1 0	0 1 × 0 1 0 0 1 1 1 0	0 1 0 0 1 0 0 0 1 1 0	0 1 0 0 1 0 0 0 1 1 0
0 0 0 0 1 1 0 1 1 0	0 0 0 × 1 1 0 1 1 1 0	0 0 0 0 0 1 1 0 1 1 0	0 0 0 0 0 1 1 0 1 1 0

Transmitted  
message

Received noisy  
message

After correction  
by rows

After correction  
by columns

FIG. 11. Iterated checking for the binary erasure channel.

**Erasure Probability.** Since none of the digits appearing in a single column have ever been together in a check group before, they are statistically independent, and the distribution of erasures in each column is binomial again. Define  $q$  as the erasure probability in the channel,  $q_1$  as the average erasure probability remaining after correction of rows,  $q_2$  as the average erasure probability remaining after checking by columns,  $p_1 = 1 - q_1$ ,  $p_2 = 1 - q_2$ .

$$(75) \quad \begin{aligned} N_1 q_1 &= N_1 q - N_1 p^{N-1} q < (N_1 q)^2, & q_1 &< (N_1 q) q, \\ N_2 q_2 &= N_2 q_1 - N_2 p_1^{N-1} q_1 < (N_2 q_1)^2, & q_2 &< (N_2 q_1) q_1. \end{aligned}$$

For  $q = 0.05$ ,  $N_1 = 10$ , this gives  $q_1 < \frac{1}{2}q$ . For  $N_2 = 20$ ,  $q_2 < \frac{1}{2}q_1$ .

**Further Iteration.** The next step, keeping  $N_1 = 10$  and  $N_2 = 20$ , is to add a check layer of 200 check digits after 39 layers of 20 blocks of ten digits each have been transmitted. This third order check will again multiply the erasure probability by a factor  $< \frac{1}{2}$ . This procedure can

be repeated indefinitely, giving for the  $k$ th order check a remaining erasure probability

$$(76) \quad q_k < (N_k q_{k-1}) q_{k-1} < \frac{1}{2} q_{k-1}, \quad \text{or} \quad q_k < 2^{-k} q_1 \\ N_k = 2N_{k-1} = 2^{k-1} N_1.$$

In the limit as  $k \rightarrow \infty$ , the remaining erasure probability becomes arbitrarily small. The rate of transmission in bits per symbol is just the fraction of input symbols which are message digits and not check digits. This is

$$(77) \quad R = (1 - \frac{1}{10})(1 - \frac{1}{20})(1 - \frac{1}{40}) \cdots \\ > 1 - (\frac{1}{10} + \frac{1}{20} + \frac{1}{40} + \cdots) \\ > 1 - (\frac{1}{10})(1 + \frac{1}{2} + \frac{1}{4} + \cdots) \\ > 1 - \frac{2}{10} = 0.80.$$

Thus the rate is at least as great as the rate in the simple block checking scheme of Fig. 10, but the error probability is as low as the receiver cares to set it if the transmitter adds the check digits of all orders, and if the receiver is willing to wait long enough for a sufficiently high order check to come along before decoding.

**Relation between Error Probability, Rate, and Delay.** The iterative coding procedure just discussed is not optimum. However, it shares two characteristics with optimum systems.

1. The reliability attained increases, for fixed rate, as the permitted coding delay increases.

2. The reliability attained increases, for fixed delay, as the required transmission rate decreases.

For any noisy channel there is a trading relationship between the probability  $P_e$  of residual error, the permissible delay  $N$ , the transmission rate  $R$ , and the channel capacity  $C$ . Here  $N$  is the number of symbols delay permitted between the transmission of a given symbol and the computation of its decoded version. The best terms of trade can be shown to give an approximately exponential decrease of error probability with delay:

$$(78) \quad P_e \doteq e^{-x(C,R)N},$$

in the sense that

$$(79) \quad \lim_{N \rightarrow \infty} \left( -\frac{\log P_e}{N} \right) = x(C, R)$$

exists for  $C > R$  as a positive number. The function  $x(C, R)$  is called the *exponent* of the error probability. For  $C - R$  small but positive, the ex-

ponent is approximately given by

$$(80) \quad x(C, R) \doteq \frac{(C - R)^2}{2\sigma_I^2},$$

where  $\sigma_I^2$  is the variance of the mutual information distribution for the given channel, and for the transmitter distribution  $p(u_i)$  which attains capacity (Refs. 31, 32, 34, 41).

## 8. THE CONTINUOUS CASE

### Continuous Sources

A waveform like that of Fig. 12 shows two kinds of continuity. It takes on a continuum of amplitude values, and its amplitude changes continuously with time.

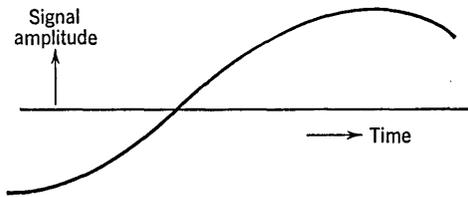


FIG. 12. A continuous waveform.

**Quantization.** The amplitude continuity may be removed by amplitude quantization, as in Fig. 13. This may be accomplished by a *quantizer*, which has an amplitude transfer characteristic of the staircase type as illustrated. The output is a waveform whose amplitude values are selected

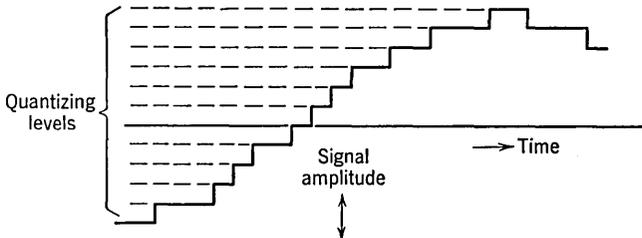


FIG. 13. An amplitude-quantized waveform.

from a discrete set, but whose jumps from one value to another occur at arbitrary times. The difference between the input signal and the quantized output is the *quantization noise* (Ref. 44).

**Sampling.** The time continuity may be removed by making periodic observations of the amplitude of the waveform, deriving from the con-

tinuous time function a sequence of *sample values*. The period of the sampling is called the *sampling interval*. The resultant samples still have amplitudes selected from a continuous set.

**SAMPLING THEOREM.** *If a waveform  $x(t)$  is bandlimited to frequencies between 0 and  $W$  cycles per second, then it is completely determined by its samples  $x(kT)$  taken at a sampling interval  $T = 1/(2W)$  seconds. The function  $x(t)$  may be re-created from its sample values by the expansion*

$$(81) \quad x(t) = \sum_{k=-\infty}^{\infty} x(kT) \frac{\sin(t - kT)}{(t - kT)}.$$

Only bandlimited signals will be discussed here. The expansion eq. (81) makes it possible to replace all such functions by sequences of sample values (Ref. 45). Some work has been done recently on the nonband-limited case (Ref. 46).

**Sampled and Quantized Signals.** A waveform which has been sampled and quantized produces a sequence of sample values, each of which has been selected from a discrete set. Such a waveform is illustrated in Fig. 14. Definitions of self-information and source rate carry over to such a sample sequence directly from the discrete case. If the waveform being sampled is very smooth, the redundancy present in the sample sequence may have a particularly simple structure, and simple conditional coding procedures may be quite efficient in removing it (Ref. 47.)

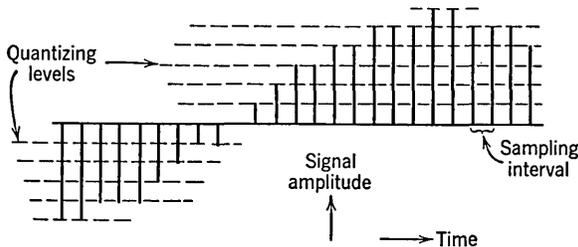


FIG. 14. Time samples from an amplitude-quantized waveform.

**Binary Pulse-Code Modulation.** If a function  $X(t)$ , bandlimited to  $W$  cycles per second, is quantized to  $M = 2^m$  levels, the sampled and quantized output consists of selections from a set of  $M = 2^m$  possible sample values. Each level may be represented uniquely by a sequence of  $m$  binary digits, giving the number of the level in binary notation. The binary number may then be encoded for transmission as a sequence of pulses and spaces, and transmitted in a bandwidth of  $mW$  cycles per second. This process is illustrated in Fig. 15. *The resulting signal, although requiring more bandwidth than the original, may be transmitted over a noisier*

channel, since the receiver need distinguish only the presence or absence of a pulse (Ref. 48).

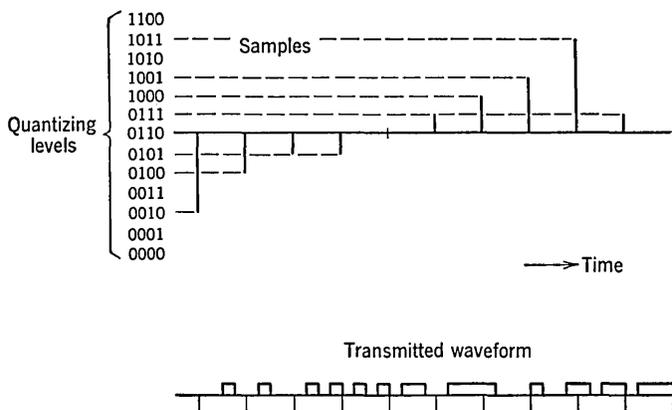


FIG. 15. Binary pulse—code modulation.

**Self-Information of Continuous Signals.** In a sampled sequence which is not quantized, each sample value is selected from an infinite set and may have infinite self-information. If the samples  $x(kT)$  are statistically independent and have the probability density  $p(x)$ , then  $-\log p(x)$  and its average value

$$(82) \quad H(X) = - \int_{-\infty}^{\infty} p(x) \log p(x) dx$$

are still definable in many cases, but they no longer represent information values. The quantity  $H(X)$  is still called the entropy of the distribution with density  $p(x)$ , but is no longer the average self-information of the source per sample. The entropy function in the continuous case is not invariant under a change in the scale by which the amplitude  $x$  is measured.

The infinite self-information associated with a selection from a continuous set arises from the fact that the selection of a single real number between 0 and 1 is equivalent to the selection of an infinite sequence of binary digits, namely the binary expansion of the real number, and conversely.

**Continuous Noisy Channels**

Only stationary channels, without storage, and with bandwidth limited from 0 to  $W$  cycles per second will be considered. Such a channel is defined by a conditional probability density function  $q(y|x)$ . For a given value  $x$  of the transmitted sample,  $q(y|x)$  gives the density of the distribution of possible received values  $y$ .

EXAMPLE. *Additive Noise.*

Let  $z$  be a noise voltage selected with probability density  $r(z)$ , and let the received signal  $y$  be the sum of the transmitted signal and the noise,  $y = x + z$ . Then

$$(83) \quad q(y|x) = q(x + z|x) = r(z) = r(y - x).$$

Thus a continuous channel with bandwidth  $W$  and additive noise is completely specified by the distribution of the noise which is added.

**Mutual Information and Rate.** If each transmitted sample value  $x$  is selected from a probability density  $p(x)$ , and the channel is specified by the conditional density  $q(y|x)$ , the joint density  $p(x, y) = p(x)q(y|x)$  defines both the channel and the transmitter strategy, in analogy to the discrete case. The probability density  $q(y)$  of the received sample values is then given by

$$(84) \quad q(y) = \int_{-\infty}^{\infty} p(x, y) dx.$$

The random variable

$$(85) \quad I(x; y) = \log \frac{p(x, y)}{p(x)q(y)}$$

is again defined as the mutual information between  $x$  and  $y$ , and its average value

$$(86) \quad R = I(X; Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) \log \frac{p(x, y)}{p(x)q(y)} dx dy$$

is the average rate of transmission of mutual information, in bits per sample. This measure retains its informational significance while self-information does not, because mutual information is invariant to a change in the scale on which both the transmitted sample  $x$  and the received sample  $y$  are measured.

EXAMPLE. *Additive Noise.* As before, let  $y = x + z$ , where  $z$  is an added noise, statistically independent of  $x$ . Then by eqs. (83) and (86),

$$\begin{aligned} (87) \quad R = I(X; Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) \log \frac{q(y|x)}{q(y)} dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) \log \frac{r(x - y)}{q(y)} dx dy \\ &= - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) \log q(y) dx dy \\ &\quad + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) \log r(y - x) dx dy \\ &= H(Y) - H(Z). \end{aligned}$$

**Channel Capacity.** The *capacity*  $C$  of a *noisy continuous channel* is the maximum value of the rate  $R$  which may be obtained in eq. (86) by varying the probability density  $p(x)$ , with which the transmitted sample values are chosen. The variation is usually constrained so as to keep constant a given peak or mean square value of the time function  $x(t)$ , which is determined through eq. (81) by the sample values. In general, finding  $C$  is a difficult variational problem.

**EXAMPLE. Additive Noise.** By eq. (87), the rate  $R$  in a channel with independent additive noise is the difference between the entropy of the distribution of received signal values  $y$  and the entropy of the distribution of the additive noise  $z$ . For a given channel, the noise distribution is fixed, so that maximizing the rate reduces to the maximization of  $H(Y)$  by variation of  $p(x)$ : thus from eqs. (83) and (84),

$$(88) \quad \max_{p(x)} H(Y) = \max_{p(x)} \left\{ - \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} p(x)r(y-x) dx \right] \times \right. \\ \left. \log \left[ \int_{-\infty}^{\infty} p(x)r(y-x) dx \right] dy \right\},$$

subject to constraints on peak or average transmitter power. This problem is difficult.

**Entropy of Gaussian Distribution.** If the sample values  $x(kT)$  of a bandlimited function  $x(t)$  are selected with statistical independence from a distribution with density  $p(x)$ , the mean square value of the time function  $x(t)$  is equal to the mean square value of the samples (Ref. 44), which is given by

$$(89) \quad \overline{x^2} = \int_{-\infty}^{\infty} x^2 p(x) dx.$$

Thus  $\overline{x^2}$  is the signal power  $S$ . For  $\overline{x^2} = S$  fixed, the distribution with maximum entropy is Gaussian (Ref. 16), with

$$(90) \quad p(x) = (1/\sqrt{2\pi S}) e^{-x^2/2S},$$

and entropy, from eq. (82), given by

$$(91) \quad H(X) = - (1/\sqrt{2\pi S}) \int_{-\infty}^{\infty} e^{-x^2/2S} (\log [1/\sqrt{2\pi S}] - x^2/2S) dx \\ = \log \sqrt{2\pi S} + \frac{1}{2} \log e \\ = \frac{1}{2} \log 2\pi e S.$$

**Gaussian Additive Noise.** If an added bandlimited noise  $z$  is statistically independent of the signal  $x$ , and either  $x$  or  $z$  has zero mean value,

$$(92) \quad \overline{y^2} = \overline{x^2} + \overline{z^2} = S + N,$$

where  $N$  is the mean noise power. Thus if the channel is given and the average transmitter power is constrained, the received power is determined by eq. (92). The entropy  $H(Y)$  in eq. (87) will then be maximized if  $q(y)$  is Gaussian with variance  $S + N$ .

However, the sum of two independent random variables cannot have a Gaussian distribution unless both random variables themselves have Gaussian distributions (Ref. 49). Thus only if the noise is Gaussian can the transmitter select a (Gaussian)  $p(x)$  which will lead to a maximum  $H(Y)$ . The rate will then be

$$(93) \quad \begin{aligned} R &= H(Y) - H(Z) \\ &= \frac{1}{2} \log 2\pi(S + N) - \frac{1}{2} \log 2\pi S \\ &= \frac{1}{2} \log (1 + S/N) \text{ bits per sample.} \end{aligned}$$

Rewriting eq. (93) on a bits-per-second basis gives:

**CAPACITY OF A CHANNEL WITH ADDITIVE GAUSSIAN NOISE.** Given a channel bandlimited from 0 to  $W$  cycles per second, with an average transmitter power  $S$ , perturbed by additive white Gaussian noise of total power  $N$ , its capacity is

$$(94) \quad C = W \log (1 + S/N) \text{ bits per second.}$$

The restriction to white noise (noise which has a uniform spectral density in the interval 0 to  $W$  cycles) is required in order that successive samples of noise be statistically independent. If they are not, the capacity will be greater than that given by eq. (94).

**Dependence of Capacity on Bandwidth.** Holding  $S$  fixed and increasing  $W$ , the noise power  $N$  increases with  $W$ , since noise power in

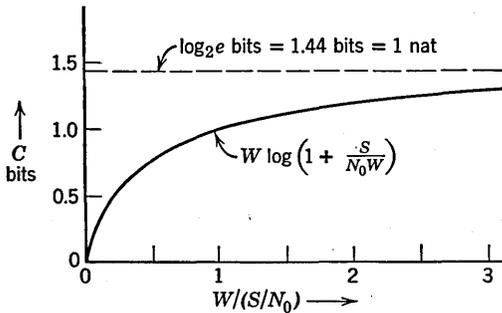


FIG. 16. Channel capacity and bandwidth.

frequencies previously rejected now enters the channel. For thermal noise and shot noise, the noise power  $N$  is directly proportional to  $W$ :

$$(95) \quad N = N_0 W \text{ watts,}$$

where  $N_0$  is the noise power per cycle bandwidth.

Substituting this in eq. (94) gives

$$(96) \quad C = W \log (1 + S/N_0 W),$$

which is plotted in Fig. 16.

### Interpretations of Capacity

Some interpretation of the capacity of a continuous noisy channel is required, since the channel can accept input information at an infinite rate, but can only transmit information about its input to the receiver at a finite rate.

**Discrete Input Interpretation.** One interpretation is provided by the fact that the noisy channel coding theorem still applies to the continuous noisy channel. The output of a discrete source may be encoded for transmission over a continuous noisy channel at any rate less than channel capacity, and the receiver can then decode the received signal with arbitrarily small error probability. In this case the transmitted signals may be continuous waveforms, but they are selected from a finite set, and therefore have finite self-information (Ref. 45).

**Quantization Interpretation.** If the receiver cannot distinguish between transmitted waveforms which are very near to one another, it is not necessary to transmit the precise waveform generated by a continuous source: some "near-by" waveform will do. The quantization process discussed earlier shows one procedure for selecting a near-by waveform. The transmitted waveforms of any finite duration are then a discrete set, and one may be selected by the receiver with small error probability despite the noisy channel. Other measures of distance, or fidelity of reproduction, have been introduced and studied (Refs. 16, 46).

**Reduction of Ignorance Interpretation.** A final interpretation also carries over from the discrete case. If successive samples are statistically independent, the receiver knows a priori that  $x$  will be selected from  $p(x)$ . A posteriori the true value of  $x$  is selected from  $p(x|y)$ , a narrower distribution with less entropy. The change in entropy,

$$(97) \quad H(X) - H(X|Y)$$

measures the average reduction in the receiver's ignorance of the value of the transmitted sample (Refs. 16, 29).

**Generalizations.** The analysis of the continuous case has been extended to cases of mixed type, i.e., to distributions which have discrete probabilities as well as densities (Refs. 16, 50, 51), and some discussion has been given of the nonbandlimited case (Ref. 46).

## REFERENCES

1. Y. Bar-Hillel and R. Carnap, Semantic information, Jackson, Editor, *Communication Theory*, Butterworths, London, 1953.
2. W. J. McGill, Multivariate information transmission, *Trans. I.R.E.*, **PGIT-4**, 93-111, Sept. 1954.
3. S. Kullback, An application of information theory to multivariate analysis, *Ann. Math. Stat.*, **23**, 88-102, March 1952.
4. B. Mandelbrot, Simple games of strategy occurring in communication through natural languages, *Trans. I.R.E.*, **PGIT-3**, 124-137, March 1954.
5. M. P. Schutzenberger, On some measures of information used in statistics, C. Cherry, Editor, *Information Theory*, Butterworths, London, 1956.
6. R. A. Fisher, Theory of statistical estimation, *Proc. Cambridge Phil. Soc.*, **22**, 700-725 (1925).
7. S. O. Rice, Mathematical analysis of random noise, *Bell System Tech. J.*, **23**, 282-332 (July 1944); **24**, 46-156 (Jan. 1945). Reprinted in N. Wax, Editor, *Noise and Stochastic Processes*, Dover, 1954.
8. N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, Technology Press and Wiley, New York, 1949.
9. D. Middleton and D. Van Meter, Detection and extraction of signals in noise from the point of view of statistical decision theory, *J. Soc. Ind. Appl. Math.* **3**, 192-253 (Dec. 1955); **4**, 86-119 (June 1956).
10. W. W. Peterson, T. G. Birdsall, and W. C. Fox, The theory of signal detectability, *Trans. I.R.E.*, **PGIT-4**, 171-212, Sept. 1954.
11. C. Cherry, *On Human Communication*, Technology Press and Wiley, 1957, esp. p. 247, footnote.
12. R. Jacobson, G. Fant, and M. Halle, Preliminaries to speech analysis, *M.I.T. Acoust. Lab. Rept. 13*, 1952.
13. R. V. L. Hartley, Transmission of information, *Bell System Tech. J.*, **7**, 535-563 (July 1928).
14. R. M. Fano, *Statistical Theory of Information*, Technology Press, Cambridge, Mass., 1957.
15. M. J. E. Golay, Bits and binitis, *Proc. I.R.E.*, **42**, 1452 (Sept. 1954).
16. C. E. Shannon, A mathematical theory of communication, *Bell System Tech. J.*, 1948 as reprinted in C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, Univ. of Illinois Press, Urbana, 1949. See p. 28.
17. R. M. Fano, The transmission of information, *M.I.T. Research Lab. Electronics Tech. Rept. 65*, March 1949.
18. D. A. Huffman, A method for the construction of minimal-redundancy codes, *Proc. I.R.E.*, **40**, 1098-1101 (Sept. 1952).
19. A. A. Sardinas and G. W. Patterson, A necessary and sufficient condition for unique decomposition of coded messages, *I.R.E. Convention Record*, Pt. 8, 104-109, March 1953.

20. B. Mandelbrot, On recurrent noise limiting coding, in *Proceedings of the Symposium on Information Networks*, Polytechnic Institute of Brooklyn, New York, 1955.
21. M. P. Schutzenberger, On an application of semi-group methods to some problems in coding, *Trans. I.R.E.*, **IT-2**, 47-60, Sept. 1956.
22. L. K. Kraft, A Device for Quantizing, Grouping and Coding Amplitude-Modulated Pulses, S.M. Thesis, Elec. Eng. Dept., M.I.T., 1949.
23. B. McMillan, Two inequalities implied by unique decipherability, *Trans. I.R.E.*, **IT-2**, 115-116, Dec. 1956.
24. B. Mandelbrot, Diagnostic et transduction en l'absence de bruit, Institut de Statistique de l'Université de Paris, Paris, 1955.
25. Shannon, *op. cit.* in (Ref. 16), p. 22.
26. N. Chomsky, Three models for the description of language, *Trans. I.R.E.* **IT-2**, 113-124, Sept. 1956.
27. N. Chomsky, *Syntactic Structures*, Mouton and Co., London, 1957.
28. Shannon, *op. cit.* in (Ref. 16), p. 26.
29. P. M. Woodward, *Probability and Information Theory, with Applications to Radar*, McGraw-Hill, New York, 1953.
30. S. Muroga, On the capacity of a discrete channel, *J. Phys. Soc. Japan*, **8**, 484-494 (1953).
31. A. Feinstein, A new basic theorem in information theory, *Trans. I.R.E.*, **PGIT-4**, 2-22, Sept. 1954.
32. C. E. Shannon, The rate of approach to ideal coding (abstract only), *I.R.E. Convention Record*, Pt. 4, 47, March 1955.
33. P. Elias, Coding for noisy channels, *I.R.E. Convention Record*, Pt. 4, 37-46, March 1955.
34. C. E. Shannon, Certain results in coding theory for noisy channels, *Information and Control*, **1**, 6-25 (Sept. 1957).
35. R. W. Hamming, Error detecting and error correcting codes, *Bell System Tech. J.*, **29**, 147-160 (1950).
36. M. Plotkin, Binary codes with specified minimum distance, *Univ. Penna. Moore School Research Div. Rept. 51-20*, 1951.
37. M. J. E. Golay, Binary coding, *Trans. I.R.E.*, **PGIT-4**, 23-28, 1954.
38. E. N. Gilbert, A comparison of signalling alphabets, *Bell System Tech. J.*, **31**, 504-522 (1952).
39. I. S. Reed, A class of multiple error-correcting codes and the decoding scheme, *Trans. I.R.E.*, **PGIT-4**, 38-49, Sept. 1954.
40. P. Elias, Error-free coding, *Trans. I.R.E.*, **PGIT-4**, 29-37, Sept. 1954.
41. P. Elias, Coding for two noisy channels, in C. Cherry, Editor, *Information Theory*, Butterworths, London, 1956.
42. D. Slepian, A class of binary signalling alphabets, *Bell System Tech. J.*, **35**, 203-234 (Jan. 1956).
43. D. Slepian, A note on two binary signalling alphabets, *Trans. I.R.E.*, **IT-2**, 84-86, June 1956.
44. W. R. Bennet, Spectra of quantized signals, *Bell System Tech. J.*, **27**, 446-472 (July 1948).
45. C. E. Shannon, Communication in the presence of noise, *Proc. I.R.E.*, **37**, 10-21 (Jan. 1949).
46. A. N. Kolmogorov, On the Shannon theory of information transmission in the case of continuous signals, *Trans. I.R.E.*, **IT-2**, 102-108, Dec. 1956.
47. P. Elias, Predictive coding, *Trans. I.R.E.*, **IT-1**, 16-33, March 1955.

48. B. M. Oliver, J. R. Pierce, and C. E. Shannon, The philosophy of PCM, *Proc. I.R.E.*, **36**, 1324-1331 (1948).
49. H. Cramer, Random variables and probability distributions, *Cambridge Tracts in Math. No. 36*, Cambridge, England, 1937.
50. S. Kullback and R. A. Liebler, On information and sufficiency, *Ann. Math. Stat.*, **22**, 79-86 (March 1951).
51. K. H. Powers, A unified theory of information, *M.I.T. Research Lab. Electronics Tech. Rept. 311*, Feb. 1956.
52. J. L. Doob, *Stochastic Processes*, Wiley, New York, 1953, esp. p. 89.

## Smoothing and Filtering

Pierre Mertz

1. Definitions: Smoothing and Prediction. Symbols	17-01
2. Definitions: Correlation	17-05
3. Relationship between Correlation and Signal Structure	17-09
4. Design of Optimum Filter	17-13
5. Extensions of Procedure	17-19
6. Network Synthesis	17-25
References	17-32

### 1. DEFINITIONS: SMOOTHING AND PREDICTION. SYMBOLS

**Time Sequence of Data.** A plot is shown in Fig. 1 of a small portion of a time sequence of data,  $f(t)$ . Such a time sequence may also be represented by an electrical signal, in which the variable is a voltage or current.

The sequence of data may be taken only at successive discrete intervals of time, instead of continuously. This is illustrated by the discrete ordinates indicated in Fig. 1.

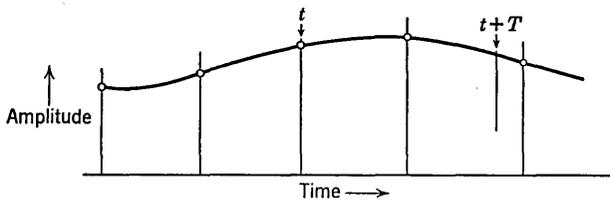


FIG. 1. Variation of a physical quantity with time.

17-01

**Stationary Time Sequence of Data.** The variation of a physical quantity with time constitutes a continuous time sequence of data. The values form a distribution. If this distribution does not show a long-range trend with time, the time sequence is said to be *stationary*. (See Chap. 13.)

*Quasi-stationary time sequence* is a distribution that is statistically stationary (i.e., shows no trend) in the short range but not in the long range.

**Errors or Noise.** There is usually a random error in the determination of a given physical quantity, or in its representation by a given electrical

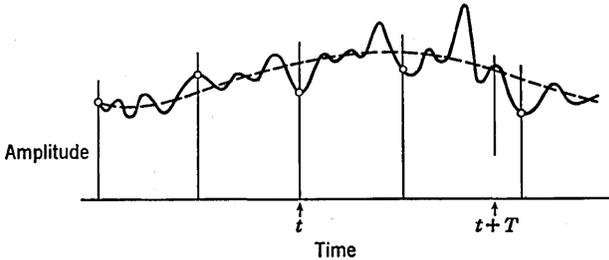


FIG. 2. Variation of a physical quantity, with superposed error, with time.

signal. This is illustrated by the erratic solid line of Fig. 2. The dotted line is the same plot as Fig. 1.

A random error may be considered as added to the actual physical quantity. If there is no long time trend in the distribution of the random error, it is also a *stationary sequence*.

In electrical signals the added sequence is usually called "noise," and in industrial processes, "disturbances."

**Smoothing Problem.** For data with random errors such as Fig. 2, an averaging process could be used to reduce the error. This assumes that the physical sequence is "smoother" than the data sequence. In an electrical signal representing the data such an averaging process can be carried out by a filter.

**Optimum Filter.** There is likely to be some filter design which has an optimum frequency-response characteristic. If the filter suppresses the rapid departures too much, it also suppresses some real variations in the physical quantity represented by the data. If, on the other hand, it does not suppress them sufficiently, it is not reducing the error as much as is feasible. For a classical theoretical analysis by Wiener see Ref. 1. The optimum filter as designed by mathematical theory is not usually critical. In practice an elementary filter is generally devised which approximates it and gives almost equal performance.

**Predicting Problem.** It is occasionally desirable not only to smooth the data, but also to extrapolate or predict the data. For *example* in Fig. 2 it may be needed, at time  $t$ , to predict the most likely signal which will occur at time  $t + T$ . This is feasible because the variation of the physical quantity is restrained by physical laws, and it does not have complete or random liberty of action.

Wiener's analysis indicates that prediction may also be effected with a filter. An optimum design is secured from nearly the same formulation as that used for the smoothing process.

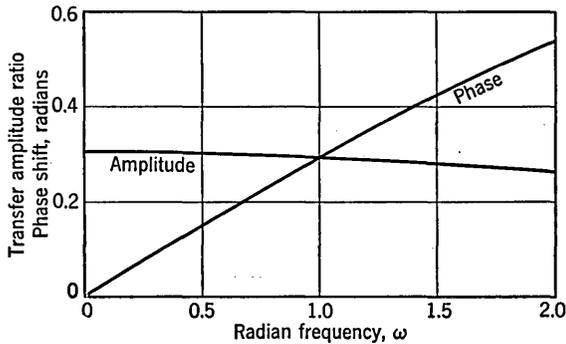


FIG. 3. Amplitude and phase characteristics vs. frequency of an optimum filter.

**EXAMPLE. Prediction Filter.** The transfer amplitude response and transfer phase shift of a smoothing and predicting filter designed according to the Wiener theory are presented in Fig. 3. The ordinates are shown as

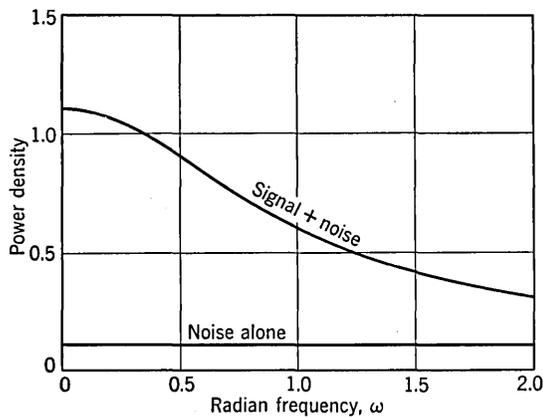


FIG. 4. Signal and noise power spectra.

functions of the radian frequency  $\omega$ . The filter is designed for a prediction time of 1 second, and for the signal and noise power spectrum illustrated in Fig. 4. More details regarding the problem and the design are given below.

**Prediction: Discrete Data.** When the data are discrete rather than continuous, the solution is not simply embodied in the form of an electrical filter. The solution describes instead a mathematical process that accomplishes an analogous averaging and predicting effect.

*Basis of Treatment.* In the present exposition, the Wiener theory is followed (Ref. 1). Advantage is taken of the treatments of Levinson (Ref. 2) and of Bode and Shannon (Ref. 3) in simplifying the presentation.

It is to be recognized that much other work has been done on the subject. At the close of the chapter some of this work, particularly recent development, is noted.

## Symbols

$A, B$	constants
$A_n$	numerator of partial fraction
$A(\omega)$	amplitude part of transfer response, in nepers
$a, b, c$	portions of an integral
$B(\omega)$	phase of transfer response, in radians
$b_j$	coefficient of $p^j$ in polynomial
$C$	parameter
$C, D, E, F$	numerators of partial fractions, sometimes with subscript indices
$C_j$	capacitance of $j$ th element in filter
$e$	base of Napierian logarithms
$F(\omega)$	signal correlation function
$f, f(t)$	signal amplitude
$g, g(t)$	noise amplitude
$g_j$	coefficient of $p$ in continued fraction
$h(t)$	total instantaneous wave amplitude
$I, \Pi$	integrals, forming part of more extensive formulas
$i$	$\sqrt{-1}$
$I(\omega)$	current, function of frequency
$J, H$	limits of summation indices
$j, h$	summation indices
$K(\omega)$	transfer response function of frequency, of Wiener filter
$K_0(x)$	Bessel function of second kind, pure imaginary argument
$k(t)$	transfer response function of time, of Wiener filter
$L_j$	inductance of $j$ th element in filter
$M, N$	limits of summation indices
$m, n$	indices
$p$	Heaviside operator = $i\omega$
$Q(\omega)$	factor of $\Phi(\omega)$ , with zero phase angle

$q(t)$	Fourier transform of $Q(\omega)$
$R$	resistance
$s$	variable of integration, for radian frequency
$T$	prediction time
$t$	time variable
$u$	variable of integration, for radian frequency
$V, V(\omega)$	voltage, function of frequency
$v, v(t)$	voltage, function of time
$v_0$	voltage amplitude
$X, Y$	numerators of partial fractions
$Y(\omega)$	transfer response function of frequency
$Y(\omega)$	admittance, function of frequency
$Z(\omega)$	impedance, function of frequency
$Z_1(p)$	driving point impedance, function of operator $p$
$Z_T(p)$	transfer impedance, function of operator $p$
$Z_T(\omega)$	transfer impedance, function of frequency
$\alpha_1, \alpha_2 \dots \beta_1, \beta_2 \dots$	zeros of polynomials, with imaginary parts $> 0$
$\alpha_1^*, \alpha_2^* \dots \beta_1^*, \beta_2^* \dots$	complex conjugates of corresponding $\alpha$ 's and $\beta$ 's
$\epsilon$	constant
$\theta$	integration limit on time variable
$\kappa$	constant
$\lambda(\omega)$	natural logarithm of $\Psi(\omega)$
$\mu, \nu$	indices
$\tau$	time variable, for correlation or integration
$\Phi(\omega)$	correlation spectrum, Fourier transform of $\phi(\tau)$
$\Phi_n(\omega)$	correlation spectrum of $n$ th derivative of input function
$\Phi_{ff}(\omega), \Phi_{gg}(\omega)$	autocorrelation spectra
$\phi$	phase shift
$\phi(\tau)$	correlation function
$\phi_n(\tau)$	correlation function of $n$ th derivative of input function
$\phi_{ff}(\tau), \phi_{gg}(\tau)$	autocorrelation functions
$\phi_{fg}(\tau), \phi_{gf}(\tau)$	cross-correlation functions
$\Psi(\omega)$	factor of $\Phi(\omega)$ for which singularities have imaginary parts $> 0$
$\Psi^*(\omega)$	complex conjugate of $\Psi(\omega)$
$\Psi_n(\omega), \Psi_n^*(\omega)$	corresponding factors of $\Psi_n(\omega)$
$\psi(\tau), \psi^*(\tau)$	Fourier transforms of $\Psi(\omega), \Psi^*(\omega)$
$\omega$	radian frequency
$\omega_j$	zero of polynomial, with imaginary part $> 0$
$\omega_j^*$	complex conjugate of $\omega_j$

## 2. DEFINITIONS: CORRELATION

**Autocorrelation Function.** The amplitude of a signal at any given time is not wholly independent of its value at other times. The correlation that exists may be expressed in terms of an autocorrelation function. In Fig. 5 the signal amplitude  $f$  is measured at times  $t$  and  $t + \tau$ . The autocorrelation function  $\phi(\tau)$  of a signal is the average product of the signal at time  $t$  and the signal at time  $t + \tau$ , averaged over a period of

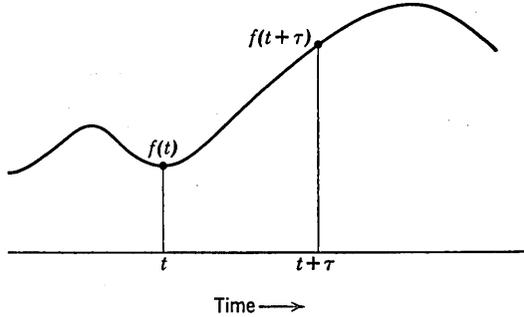


FIG. 5. Data taken for determination of autocorrelation coefficient.

time long enough to smooth out instantaneous fluctuations. With the overscribed bar to indicate averaging,

$$\phi(\tau) = \overline{f(t)f(t + \tau)},$$

or

$$(1) \quad \phi(\tau) = \lim_{\theta \rightarrow \infty} \frac{1}{2\theta} \int_{-\theta}^{\theta} f(t)f(t + \tau) dt,$$

where  $\tau =$  finite time shift.

The autocorrelation function is a measure of the extent to which the value of  $f(t)$  at any given time can be used to predict  $f(t)$  at a time interval  $\tau$  later.

EXAMPLE. An autocorrelation function is illustrated in Fig. 6. This is the one assumed for the signal whose spectrum is illustrated in Fig. 4.

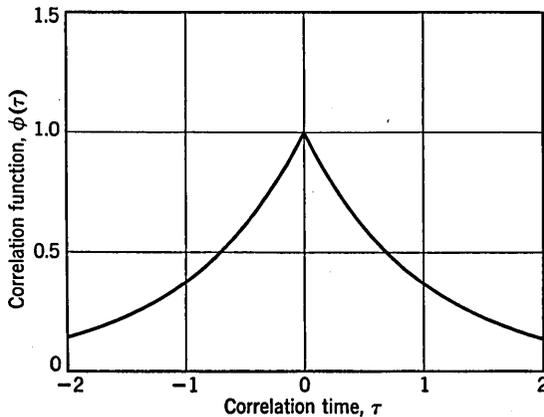


FIG. 6. Autocorrelation function.

The autocorrelation function in general is at a maximum at  $\tau = 0$ , and it usually drops off to zero or nearly zero for large values of  $\tau$ . It shows even symmetry about  $\tau = 0$ . The peak is broad when  $f(t)$  contains primarily low frequencies and narrow when  $f(t)$  contains primarily high frequencies.

**Power Density Spectrum.** It is useful to deal with the Fourier transform of the autocorrelation function  $\phi(\tau)$  which may be called the *autocorrelation spectrum*,  $\Phi(\omega)$ . This is

$$(2) \quad \Phi(\omega) = \int_{-\infty}^{\infty} \phi(\tau) e^{-i\omega\tau} d\tau,$$

and reciprocally

$$(3) \quad \phi(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi(\omega) e^{i\omega\tau} d\omega.$$

Wiener (Refs. 1 and 4) has identified the autocorrelation spectrum  $\Phi(\omega)$  with the power density spectrum of the signal (with a suitable normalizing

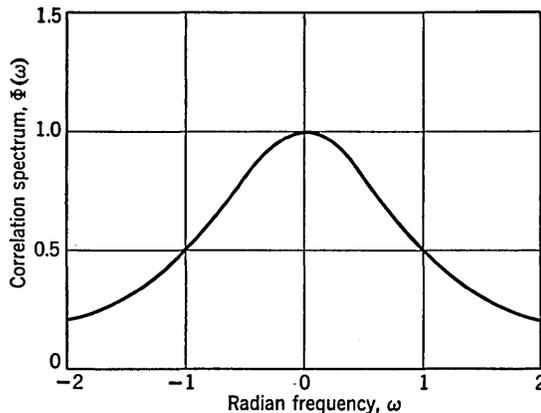


FIG. 7. Autocorrelation spectrum.

factor). That is, the function  $\Phi(\omega)$  is a measure of the mean power in the signal per unit of band width for each frequency.

The power density spectrum for the autocorrelation function of Fig. 6 is shown in Fig. 7. For simplicity a normalizing factor, to be discussed later, is introduced. This has the effect of causing both plots to reach a maximum ordinate of 1. Both plots show functions having even symmetry.

The power density is expressed in terms of coefficients of sine and cosine terms of positive frequencies. The coefficient, for any given frequency, is twice the numerical value of  $\Phi(\omega)$  of eq. (3) for that frequency. This

corresponds to the addition of  $\Phi(-\omega)$  to  $\Phi(\omega)$ . This factor represents part of the arbitrary factor mentioned in the paragraph above.

**Autocorrelation of Noise.** In the same way as the signal, the noise shows autocorrelation as defined by eq. (1) and illustrated in Figs. 5 and 6.

For the purposes of the illustration in Fig. 4, the autocorrelation of the noise is assumed to extend over a shorter time than for the signal, i.e., over only a negligible interval of time about  $\tau = 0$ .

The autocorrelation function for the noise also has a Fourier transform as defined by eq. (2) and illustrated in Fig. 7. In the example of Fig. 4 the assumption just mentioned greatly stretches the frequency scale of the noise curve in Fig. 7. Thus the power spectral density has been taken as substantially a constant over the frequency range of interest.

The *cross-correlation function* between two different signals is similarly defined as the average product of one signal  $g(t)$  and a second signal  $f(t)$  at time  $t + \tau$ . Thus

$$\phi_{fg}(\tau) = \overline{f(t + \tau)g(t)},$$

or

$$(4) \quad \phi_{fg}(\tau) = \lim_{\theta \rightarrow \infty} \frac{1}{2\theta} \int_{-\theta}^{\theta} f(t + \tau)g(t) dt$$

The cross-correlation in the reverse sense is

$$(5) \quad \phi_{gf}(\tau) = \phi_{fg}(-\tau)$$

**EXAMPLES.** If as in Fig. 8 the signal  $f$  is measured at time  $t + \tau$  and the noise  $g$  at time  $t$ , eq. (5) is the cross-correlation between signal and noise. If  $f(t)$  represents an input to a system and  $g(t)$  its output, then  $\phi_{fg}$  represents the cross-correlation of input and response.

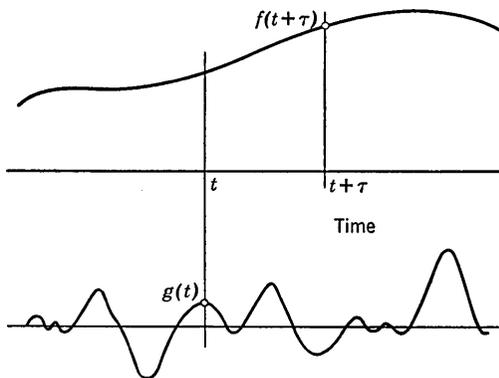


FIG. 8. Data taken for determination of cross-correlation coefficient.

**Resultant Correlations in Signal Mixed with Noise.** The auto-correlation  $\phi(\tau)$  of the function  $h(t) = f(t) + g(t)$  is made up of the auto-correlations of the component functions that may be denoted respectively by  $\phi_{ff}(\tau)$  and  $\phi_{gg}(\tau)$ , and the cross-correlations. Thus

$$(6) \quad \phi(\tau) = \phi_{ff}(\tau) + \phi_{fg}(\tau) + \phi_{gf}(-\tau) + \phi_{gg}(\tau).$$

*In most practical systems there is no correlation between signal  $f$  and noise  $g$ , and also the average value of the noise is zero.* Thus

$$\phi_{fg}(\tau) = 0, \quad \phi_{gf}(-\tau) = 0,$$

and

$$(7) \quad \phi(\tau) = \phi_{ff}(\tau) + \phi_{gg}(\tau).$$

**Cross-Correlation Spectrum.** The corresponding relations hold for the Fourier transforms, and the *cross-correlation spectrum* is

$$(8) \quad \Phi(\omega) = \Phi_{ff}(\omega) + \Phi_{fg}(\omega) + \Phi_{gf}(\omega) + \Phi_{gg}(\omega).$$

In most practical systems, as discussed above

$$(9) \quad \Phi(\omega) = \Phi_{ff}(\omega) + \Phi_{gg}(\omega).$$

This last equation means that the power density spectra of the signal and of the noise add to form that of the total wave. This addition has been performed in Fig. 4.

### 3. RELATIONSHIP BETWEEN CORRELATION AND SIGNAL STRUCTURE

**Latitude in Interpretation of Power Spectrum.** In general, there is a unique relation between a function and its Fourier transform. The power density spectrum of the signal does not, however, define a unique signal because it gives no phase information for the spectral components. Thus the power density spectrum imposes restrictions only on the signal, and it does not define it.

**Construction of Signal from Bandlimited or Shaped Impulses.** The signal may be considered as originating from a succession of idealized impulses sent out according to the information contained in the message. This succession of impulses is then transmitted through a bandlimiting or shaping network. The output of the network becomes the signal.

The character of the shaping network affects the power density of the spectrum. Consider an especially simple specific case (Ref. 5). Information is assumed as coming at periodic intervals. It is further assumed that the signal is of sufficiently random character to have an average amplitude

of zero and to give a constant power density over the spectrum. In this simple case the power density spectrum is simply the square of the transfer amplitude response spectrum of the shaping network. The corresponding amplitude response curve of Fig. 7 is shown in Fig. 9.

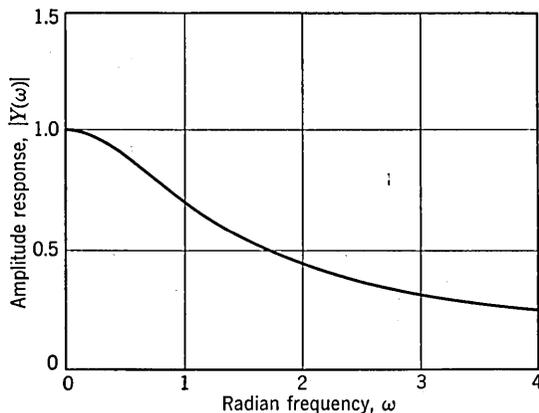


FIG. 9. Amplitude vs. frequency response of shaping network.

**Influence of Phase Response of Shaping Network.** The square of the amplitude response does not as yet specify the phase characteristic of the shaping network. One may simply assume this phase shift to be zero throughout the spectrum. In such a case the Fourier transform of the amplitude characteristic of Fig. 9 would be that illustrated in Fig. 10. This gives the output response of such a network to a unit idealized impulse

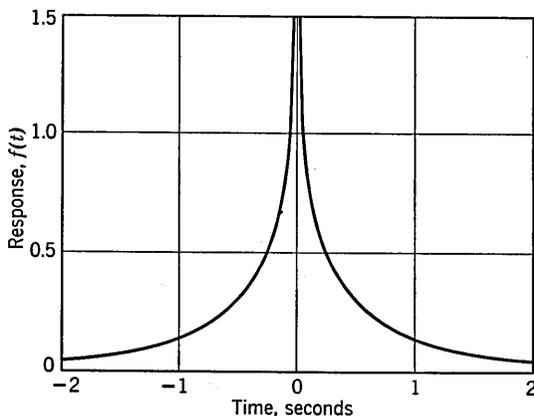


FIG. 10. Transient response of hypothetical shaping network having zero phase shift.

signal input at time  $t = 0$ . It shows a response before the occurrence of the impulse, which indicates that the particular network is not physically realizable.

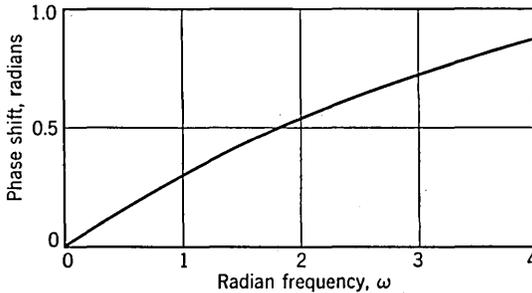


FIG. 11. Phase shift of physically realizable shaping network.

For the network to be physically realizable it must at least meet the minimum phase condition (Ref. 6). For a transfer response spectrum of

$$(10) \quad Y(\omega) = e^{A(\omega) + iB(\omega)}$$

the minimum phase condition is that

$$(11) \quad B(\omega) = \frac{2\omega}{\pi} \int_0^{\infty} \frac{A(s) - A(\omega)}{s^2 - \omega^2} ds.$$

The phase shift of any realizable network is equal to or greater than this. The phase shift meeting the condition for the amplitude response characteristic of Fig. 9 is plotted in Fig. 11.

For a response having the amplitude characteristic of Fig. 9 and the phase of Fig. 11, the Fourier transform is presented in Fig. 12. This no

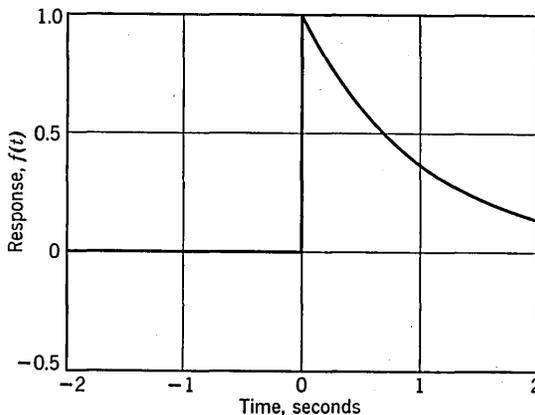


FIG. 12. Transient response of physically realizable shaping network.

longer shows an advance response to the impulse signal input. The response now starts at  $t = 0$ , the same time as the signal input.

**Representation of Network Properties in Complex Frequency Plane.** If the square of the *transfer amplitude response* is a rational function of  $\omega$ , it may be expressed thus:

$$(12) \quad |Y(\omega)|^2 = \Phi(\omega) = \kappa^2 \frac{(\omega - \alpha_1)(\omega - \alpha_1^*)(\omega - \alpha_2)(\omega - \alpha_2^*) \cdots}{(\omega - \beta_1)(\omega - \beta_1^*)(\omega - \beta_2)(\omega - \beta_2^*) \cdots},$$

where  $\kappa =$  a constant,

$\alpha, \beta =$  zeros of numerator or denominator with imaginary parts  $> 0$ ,

$\alpha^*, \beta^* =$  complex conjugates of  $\alpha, \beta$ .

The  $\alpha$ 's are the zeros, and the  $\beta$ 's poles, of the function  $|Y(\omega)|^2$ . A plot of these, for the function plotted in Fig. 7, is given in Fig. 13. There

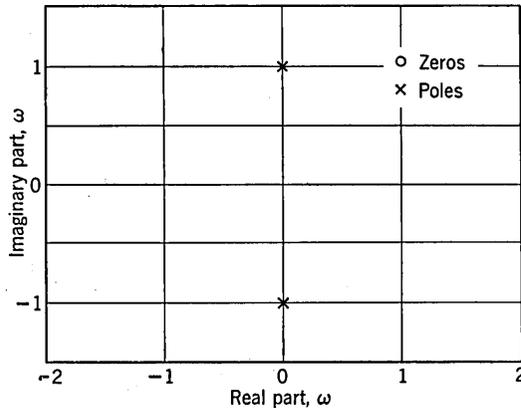


Fig. 13. Singularities of rational correlation function.

are only poles, one being above, and the other below, the axis of real  $\omega$ 's.

The response indicated by Fig. 9, with zero phase shift, is not represented by a rational function, and hence it cannot be described simply by zeros and poles in the complex frequency plane.

**Location of Zeros and Poles for Minimum Phase Network.** From Bode and Shannon (Ref. 3), the minimum phase network has the transfer response function

$$(13) \quad Y(\omega) = \kappa \frac{(\omega - \alpha_1)(\omega - \alpha_2) \cdots}{(\omega - \beta_1)(\omega - \beta_2) \cdots}.$$

The zeros and poles are all in the upper half-plane of the complex frequency space. A plot of the zeros and poles which applies to the amplitude

response of Fig. 9 and phase shift of Fig. 11 is shown in Fig. 14. In this specific case there are no zeros and only one pole, which is in the upper half-plane.

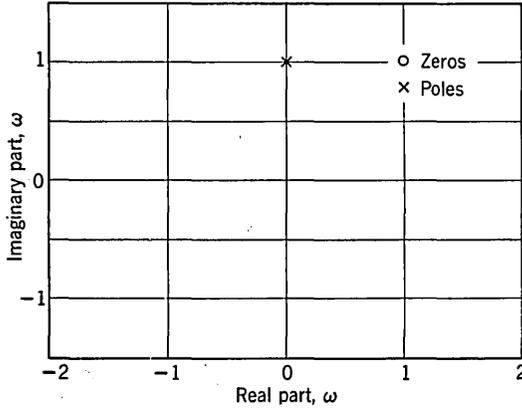


FIG. 14. Singularities of physically realizable shaping network.

#### 4. DESIGN OF OPTIMUM FILTER

**Criterion of Optimization.** A criterion of performance is necessary to judge when one filter design is better than another. That used by Wiener is based on a comparison between the filter output and the actual signal, freed of noise, at the extrapolated time. The difference, or error, is measured as a function of time, and its root-mean-square value determined. *The optimum filter is taken as that for which the root-mean-square error (Chap. 13) is a minimum.* This assumption is important to the development of the theory.

**Wiener Solution, Smoothing and Prediction.** The optimum filter proposed by Wiener has a frequency response which may be designated as  $K(\omega)$ . This has a Fourier transform  $k(t)$ . The reciprocal relations between the two are:

$$(14) \quad K(\omega) = \int_{-\infty}^{\infty} k(t)e^{-i\omega t} dt$$

$$(15) \quad k(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} K(\omega)e^{i\omega t} d\omega.$$

In the simple Wiener solution the transfer response of the optimum filter has the value

$$(16) \quad K(\omega) = \frac{1}{2\pi\Psi(\omega)} \int_0^{\infty} e^{-i\omega t} dt \int_{-\infty}^{\infty} \frac{\Phi_{ff}(u)}{\Psi^*(u)} e^{iu(t+T)} du$$

Here  $\Phi_{ff}(\omega)$  represents, as in eq. (8), the autocorrelation spectrum of  $\omega$  of the signal alone. The  $\omega$  is changed to  $u$  as a variable of integration.

The quantities  $\Psi(\omega)$  and  $\Psi^*(\omega)$  represent the factors of the autocorrelation spectrum of the signal plus noise,  $\Phi(\omega)$  of eqs. (8) and (9). The factors are taken, as in eqs. (12) and (13), such that

$$(17) \quad \Phi(\omega) = \Psi(\omega)\Psi^*(\omega),$$

where  $\Psi(\omega)$  has zeros and poles only in the upper half-plane of the complex frequency space, and  $\Psi^*(\omega)$  only in the lower half-plane. The quantity  $T$  is the *prediction time*.

In the limiting case where no noise is added to the signal  $\Phi_{ff}(u) = \Phi(u) = \Psi(u) \cdot \Psi^*(u)$ . Thus in the integrand of eq. (16), the ratio  $\Phi_{ff}(u)/\Psi^*(u) = \Psi(u)$ .

In this limiting case eq. (16) takes on the simpler form

$$(18) \quad K(\omega) = \frac{1}{2\pi\Psi(\omega)} \int_0^\infty e^{-i\omega t} dt \int_{-\infty}^\infty \Psi(u)e^{iu(t+T)} du.$$

**Factorization Problem.** The key to the solution of the filter is of course the solution of the factorization problem expressed in eq. (17).

The formal solution, in terms of the  $\Phi(\omega)$  assumed known, is given by Levinson (Ref. 2) as

$$(19) \quad \Psi(\omega) = e^{\lambda(\omega)}$$

$$\lambda(\omega) = \frac{i}{\pi} \int_0^\infty \frac{\omega \log \Phi(s)}{\omega^2 - s^2} ds.$$

$$(20) \quad \Psi^*(\omega) = \Phi(\omega)/\Psi(\omega).$$

**Identification with Minimum Phase Network.** The connection of eq. (17) with eqs. (12) and (13) identifies  $\Psi(\omega)$  with the transfer response characteristic  $Y(\omega)$  of the minimum phase shaping network which has the amplitude response characteristic

$$|Y(\omega)| = \sqrt{\Phi(\omega)}.$$

The eq. (19) shows some similarity to eq. (11). The two differ because eq. (11) determines the phase only of  $\Psi(\omega)$  (or  $Y(\omega)$  of eq. (10)) whereas eq. (19) determines the entire complex exponent in the equation just above it.

**Alternative Formulation.** In actual practice eqs. (16) and (19) lead to substantial mathematical difficulties even in simple cases. One practical approach is, as in eq. (12), to take  $\Phi(\omega)$  as a rational function. When this

is done  $\Psi(\omega)$  is given by eq. (13) and  $\Psi^*(\omega)$  by its complex conjugate. That is,

$$(21) \quad \Psi(\omega) = \kappa \frac{\prod_{m=1}^M (\omega - \alpha_m)}{\prod_{n=1}^N (\omega - \beta_n)},$$

$$(22) \quad \Psi^*(\omega) = \kappa \frac{\prod_{m=1}^M (\omega - \alpha_m^*)}{\prod_{n=1}^N (\omega - \beta_n^*)}.$$

A further transformation is convenient because integrands in the form of sums are more easy to handle than as products.

The products may be expanded to partial fractions, thus, for first order poles

$$\Psi(\omega) = \sum_{n=1}^N \frac{C_n}{\omega - \beta_n}.$$

For poles of order  $\mu$ ,

$$\Psi(\omega) = \sum_{n=1}^N \frac{C_{\mu,n}}{(\omega - \beta_n)^\mu}.$$

In a similar manner the principal factor of the integrand at the right in eq. (16), if it has no poles of higher order than the first, may be expanded thus

$$(23) \quad \frac{\Phi_{ff}(\omega)}{\Psi^*(\omega)} = \sum_{j=1}^J \frac{D_j}{\omega - \omega_j} + \sum_{j=1}^J \frac{E_j}{\omega - \omega_j^*}.$$

Where poles exist of order  $\mu$ , this becomes

$$(24) \quad \frac{\Phi_{ff}(\omega)}{\Psi^*(\omega)} = \sum_{j=1}^J \left( \frac{D_{\mu,j}}{(\omega - \omega_j)^\mu} + \frac{E_{\mu,j}}{(\omega - \omega_j^*)^\mu} \right).$$

Wiener has shown that when these are substituted in eq. (16) the solution becomes, for the case with poles only of the first order in both  $\Phi_{ff}(\omega)/\Psi^*(\omega)$  and  $\Psi(\omega)$ ,

$$(25) \quad K(\omega) = \sum_{j=1}^J \frac{D_j e^{i\omega_j T}}{\omega - \beta_j} \bigg/ \sum_{n=1}^N \frac{C_n}{\omega - \beta_n}.$$

For poles of order  $\mu$ , this is

$$(26) \quad K(\omega) = \frac{\sum_{j=1}^J D_{\mu,j} e^{i\omega_j T} \sum_{\nu=0}^{\mu-1} \frac{(iT)^\nu}{\nu! (\omega - \omega_j)^{\mu-\nu}}}{\sum_{n=1}^N \frac{C_{\mu,n}}{(\omega - \beta_n)^\mu}}$$

Empirical solutions of the factorization problem are mentioned in Sect. 5.

EXAMPLE 1. A simple illustration of prediction given by Wiener (Ref. 1) has been presented in Figs. 3 to 12. The equations for this illustration may now be noted.

For the signal alone:

$$(27) \quad \Phi_{ff}(\omega) = \frac{1}{\omega^2 + 1}$$

The Fourier transform is found in Campbell and Foster (Ref. 7). This is tabulated with the argument  $p = i\omega$  instead of  $\omega$ , so that

$$(28) \quad \Phi_{ff}(p) = \frac{-1}{p^2 - 1}$$

From pair 444 (Ref. 7)

$$(29) \quad \phi_{ff}(\tau) = \frac{1}{2} e^{-|\tau|}$$

This is represented in Fig. 6, ignoring the factor  $1/2$ .

For the noise alone:

$$(30) \quad \Phi_{gg}(\omega) = \frac{\epsilon^2}{(\omega/C)^2 + 1} \rightarrow \epsilon^2,$$

where  $C \rightarrow \infty$ .

$$(31) \quad \begin{aligned} \Phi(\omega) &= \Phi_{ff}(\omega) + \Phi_{gg}(\omega) \\ &= \frac{1}{\omega^2 + 1} + \frac{\epsilon^2}{(\omega/C)^2 + 1} \rightarrow \frac{1}{\omega^2 + 1} + \epsilon^2, \end{aligned}$$

$$(32) \quad \phi(\tau) = \frac{1}{2} (e^{-|\tau|} + \epsilon C e^{-|\tau/C|}).$$

For the conditions illustrated in Figs. 9 and 10, where it is assumed that the phase response of the shaping network is zero (Sect. 3),

$$(33) \quad Q_{gg}(\omega) = \sqrt{\Phi_{ff}(\omega)} = \frac{1}{\sqrt{\omega^2 + 1}},$$

$$(34) \quad Q_{gg}(p) = \frac{1}{\sqrt{1 - p^2}}.$$

From pair 558 (Ref.7),

$$(35) \quad q_{gg}(\tau) = (1/\pi)K_0(|\tau|).$$

Here  $K_0$  is the Bessel function of the second kind, and imaginary argument (the actual argument being  $i|\tau|$ ). It is tabulated in Watson (Ref. 8). This is illustrated in Fig. 10.

For the shaping network assumed as physically realizable:

$$(36) \quad \Psi_{ff}(\omega) = \frac{-i}{\omega - i},$$

$$(37) \quad \Psi_{ff}(p) = \frac{1}{p + 1}.$$

From pair 438 (Ref. 7),

$$(38) \quad \Psi(t) = e^{-t}, \quad t > 0.$$

This is illustrated in Fig. 12:

$$(39) \quad \Psi_{ff}^*(\omega) = \frac{i}{\omega + i},$$

$$(40) \quad \Psi_{ff}^*(p) = \frac{-1}{p - 1}.$$

From pair 439 (Ref. 7),

$$(41) \quad \Psi_{ff}^*(t) = e^t, \quad t < 0.$$

$$(42) \quad \Psi(\omega) = \frac{\epsilon[\omega - (i\sqrt{1 + \epsilon^2/\epsilon})]}{\omega - i},$$

$$\Psi^*(\omega) = \frac{\epsilon[\omega + (i\sqrt{1 + \epsilon^2/\epsilon})]}{\omega + i},$$

$$\begin{aligned} \frac{\Phi_{ff}(\omega)}{\Psi^*(\omega)} &= \frac{1}{\epsilon[\omega - i][\omega + (i\sqrt{1 + \epsilon^2/\epsilon})]} \\ &= \frac{D}{\omega - i} + \frac{E}{\omega + (i\sqrt{1 + \epsilon^2/\epsilon})}, \end{aligned}$$

where

$$\begin{aligned}
 D &= \frac{-i}{\epsilon + \sqrt{1 + \epsilon^2}}, \\
 E &= \frac{i}{\epsilon + \sqrt{1 + \epsilon^2}}, \\
 K(\omega) &= \frac{De^{-T}/(\omega - i)}{\epsilon[\omega - (i\sqrt{1 + \epsilon^2}/\epsilon)]/(\omega - i)}, \\
 (43) \quad K(\omega) &= \frac{e^{-T}}{(\epsilon + \sqrt{1 + \epsilon^2})(\sqrt{1 + \epsilon^2} + i\omega)} \\
 &= e^{-T} \cdot \frac{1}{\epsilon + \sqrt{1 + \epsilon^2}} \cdot \frac{\exp[-i \tan^{-1}(\epsilon\omega/\sqrt{1 + \epsilon^2})]}{\sqrt{1 + \epsilon^2} + \epsilon^2\omega^2}
 \end{aligned}$$

The form of eq. (43) indicates the nature of the variables in  $K(\omega)$ . The first factor depends only upon the prediction time  $T$ , and the second only upon the noise spectral density  $\epsilon^2$ . The denominator of the third factor is a combined function of noise density and frequency. All three of these quantities are real and hence affect only the amplitude response of  $K(\omega)$ . The numerator of the third factor is complex. It has a modulus of unity and expresses the phase shift of  $K(\omega)$ . The amplitude response and the phase are the two quantities plotted in Fig. 3.

EXAMPLE 2. A second illustration, by Wiener, assumes no noise. The function required of the filter is prediction. Consider

$$(44) \quad \Phi(\omega) = \frac{1}{(\omega^2 + 1)^2},$$

$$(45) \quad \Phi(p) = \frac{1}{(p^2 - 1)^2}.$$

From pair 433 (Ref. 7),

$$(46) \quad \phi(t) = [(|t| + 1)/4]e^{-|t|}.$$

$$\Psi(\omega) = \frac{-1}{(\omega - i)^2}, \quad \Psi^*(\omega) = \frac{-1}{(\omega + i)^2}, \quad \Psi(p) = \frac{1}{(p + 1)^2}.$$

From pair 442 (Ref. 7),

$$\psi(t) = te^{-t}, \quad t > 0.$$

From eq. (18)

$$\begin{aligned}
 (47) \quad K(\omega) &= \frac{1}{\Psi(\omega)} \int_0^\infty \psi(t+T)e^{-i\omega t} dt \\
 &= \frac{1}{\Psi(\omega)} \int_0^\infty (e^{-T}te^{-t} + Te^{-T}e^{-t})e^{-i\omega t} dt \\
 &= \frac{(\omega - i)^2}{-1} \left[ \frac{-e^{-T}}{(\omega - i)^2} - \frac{iTe^{-T}}{\omega - i} \right].
 \end{aligned}$$

$$(48) \quad K(\omega) = e^{-T}(1 + T) + i\omega Te^{-T}.$$

This solution is also reached from the alternative formulation of eq. (26).

The solution consists of a term independent of  $\omega$  and a term which comprises a differentiation of the input wave. This was also the case in the previous example, except that there the total band was limited by the denominator. In the present case the total band is infinite.

In polar form the solution is

$$(49) \quad K(\omega) = e^{-T} \sqrt{1 + 2T + T^2 + \omega^2 T^2} \exp \{i \tan^{-1} [\omega T / (1 + T)]\}.$$

## 5. EXTENSIONS OF PROCEDURE

**Filters with Lag.** On occasion the urgency in time of reproduction of a signal mixed with noise is not so great as the need for greatest feasible reduction of the noise in the reproduction.

In such cases the prediction time  $T$  is advantageously changed into a lag, that is,  $T$  becomes negative. The optimum filter formula of eqs. (16) and (26) still holds generally, but there are difficulties in carrying out the second integration, with the lower limit of zero.

Wiener (Ref. 1) suggests an approximation thus:

$$(50) \quad e^{i\omega T} \approx \left[ \frac{1 + (i\omega T/2\nu)}{1 - (i\omega T/2\nu)} \right]^{\nu}.$$

The single case is considered where the ratio of eq. (24) contains only one pole in the upper half-plane of  $\omega$ . This is assumed of the first order, at  $\omega_1$ . A possible pole in the lower half-plane may be at  $\omega_2^*$ . Then

$$(51) \quad \frac{\Phi_{ff}(\omega)e^{i\omega T}}{\Psi^*(\omega)} \approx \left[ \frac{1 + (i\omega_1 T/2\nu)}{1 - (i\omega_1 T/2\nu)} \right]^{\nu} \frac{F}{(\omega - \omega_1)(\omega - \omega_2^*)}$$

$$(52) \quad \approx \sum_{n=1}^N \frac{A_n}{[1 - (i\omega_1 T/2\nu)]^n} + \frac{X}{(\omega - \omega_1)} + \frac{Y}{(\omega - \omega_2^*)}.$$

In eq. (52) the first summation term in the partial fraction expansion is approximated only to  $n = N \leq \nu$ . Then

$$(53) \quad K(\omega) \approx \left( \sum_{n=1}^N \frac{A_n}{[1 - (i\omega_1 T/2\nu)]^n} + \frac{X}{\omega - \omega_1} \right) / \Psi(\omega).$$

In the example which was illustrated by eqs. (36) to (43), Wiener gives the result as

$$(54) \quad K(\omega) = \frac{T^2/2}{[1 + (T/2)][(T/2)\sqrt{1 + \epsilon^2} - \epsilon]} + \frac{1 + i\omega}{(\sqrt{1 + \epsilon^2} + i\omega)[1 - (i\omega T/2)]} + \frac{1 - (T/2)}{[1 + (T/2)](\epsilon + \sqrt{1 + \epsilon^2})(\sqrt{1 + \epsilon^2} + i\omega)}.$$

Where the noise level is high and the signal weak a simple formula is obtained for the optimum filter with lag. Let

$$(55) \quad \Phi_{ff}(\omega) = \epsilon F(\omega), \quad \Phi_{gg}(\omega) = 1,$$

$$(56) \quad \Phi(\omega) = 1 + \epsilon F(\omega),$$

$$(57) \quad K(\omega) \approx \frac{\Phi_{ff}(\omega)}{\Phi(\omega)} \approx \frac{\epsilon F(\omega)}{1 + \epsilon F(\omega)},$$

$$(58) \quad \approx \epsilon F(\omega).$$

That is, the optimum filtering in such a case merely follows the spectral power density weighting of the signal. This corresponds to conventional procedure.

**Input Functions with Correlation in a Derivative.** On occasion the input function that is dealt with shows a strong correlation in one of its derivatives. By this it is meant that the correlation function is sustained significantly, out to a time of the order of the prediction time or longer.

**EXAMPLES.** This condition occurs when the function represents one of the coordinates of the motion of a massive body. If the forces on this are not too large, the velocity can show correlation out to a substantial time. Another case is that of a massive body acted on by forces which change only by small amounts from instant to instant. In such a situation the acceleration can show correlation over a useful time range.

Cases of this type may assume significant practical importance, and in such situations it may be useful to base the filter design on the correlation function in the appropriate derivative. The autocorrelation function of the  $n$ th derivative may be denoted by  $\phi_n(\tau)$ , and its Fourier transform by

$\Phi_n(\omega)$ . In the same way that  $\Phi(\omega)$  was factored above in eq. (17),  $\Phi_n(\omega)$  may be factored as

$$(59) \quad \Phi_n(\omega) = \Psi_n(\omega)\Psi_n^*(\omega).$$

Then the Wiener formula for the optimum filter is, for the prediction time  $T$ ,

$$(60) \quad K(\omega) = 1 + i\omega T + \dots + \frac{(i\omega T)^{n-1}}{(n-1)!} \\ + \frac{\omega^n}{2\pi\Psi_n(\omega)} \int_0^\infty e^{-i\omega t} dt \int_{-\infty}^\infty \frac{\Psi_n(u)\epsilon^{iut}}{u^n} \times \\ \left[ (\epsilon^{iuT} - 1) - (iuT) - \dots - \frac{(iuT)^{n-1}}{(n-1)!} \right] du.$$

This is the form of the equation in which no noise is assumed in the input. Where there is noise the  $\Psi_n(u)$  in the integrand would be replaced by the expression  $\Phi_{ff}(\omega)/\Psi^*(\omega)$  in eq. (16).

EXAMPLE. Assume that the correlation spectrum of eq. (44) describes the correlation in the second derivative of the input function and that no noise is present. That is,

$$(61) \quad \phi_2(\tau) = [(|\tau| + 1)/4]e^{-|\tau|},$$

or

$$\Phi_2(\omega) = \frac{1}{(\omega^2 + 1)^2}.$$

Then the solution for the filter for the prediction time  $T$ , as given by eq. (60), is

$$K(\omega) = 1 + i\omega T + \frac{\omega^2}{2\pi\Psi_2(\omega)} \int_0^\infty e^{-i\omega t} dt \int_{-\infty}^\infty \frac{\Psi_2(u)}{u^2} e^{iut} (e^{iut} - 1 - iuT) du.$$

As in eq. (46),

$$\Psi_2(\omega) = \frac{-1}{(\omega - i)^2}.$$

Let

$$I = \frac{-1}{2\pi} \int_{-\infty}^\infty \frac{e^{iut}}{u^2(u-i)^2} (e^{iuT} - 1 - iuT) du, \\ I = \frac{-1}{2\pi} \int_{-\infty}^\infty \left[ \frac{e^{iu(t+T)}}{u^2(u-i)^2} - \frac{e^{iut}}{u^2(u-i)^2} - \frac{iTe^{iut}}{u(u-i)^2} \right] du$$

The integration of the third term of the integrand is accomplished by the

combination of pairs 210 and 442 (Ref. 7) and gives

$$c = -Tte^{-t} - Te^{-t} + T, \quad t > 0.$$

The integration of the first and second terms is accomplished by the further application of pair 210, and gives

$$a = -(t + T)e^{-(t+T)} - 2e^{-(t+T)} - (t + T) + 2, \quad t + T > 0,$$

$$b = te^{-t} + 2e^{-t} + t - 2, \quad t > 0.$$

Thus

$$I = a + b + c$$

$$= -te^{-t}(e^{-T} + T - 1) - e^{-t}(Te^{-T} + 2e^{-T} + T - 2), \quad t > 0.$$

Continuing to the second integration of eq. (60)

$$\begin{aligned} \text{II} &= \int_0^{\infty} I e^{-i\omega t} dt \\ &= \int_0^{\infty} (Ate^{-t} + Be^{-t})e^{i\omega t} dt \\ &= \frac{-A}{(\omega - i)^2} - \frac{iB}{\omega - i}. \end{aligned}$$

$$\frac{\omega^2 \text{II}}{\Psi_2(\omega)} = -\omega^2(\omega - i)^2 \text{II} = \omega^2 A + i\omega^2(\omega - i)B,$$

$$K(\omega) = 1 + i\omega T + \omega^2 A + \omega^2 B + i\omega^3 B.$$

$$(62) \quad K(\omega) = 1 + i\omega T + (i\omega)^2(Te^{-T} + 3e^{-T} + 2T - 3)$$

$$+ (i\omega)^3(Te^{-T} + 2e^{-T} + T - 2).$$

In this solution, there are three successive differentiations of the input wave to add to the constant term. As in the previous example, characterized by eq. (48), the absence of noise in the assumptions leads to an infinite band in the filter.

**“Filters” for Discrete Data.** The signal which has been assumed in the discussion up to this point has been continuous. However, it could be expressed as the amplitude of a succession of discrete pulses, such as the maximum daily temperatures at a given location.

Where the data are *discrete*, in electrical form or not, they cannot be passed through a physical electrical filter to carry out the operations which

have been discussed. *The operations can, however, be expressed in terms of mathematical processes, with only small changes from the previous description.*

For discrete data eq. (1) becomes

$$(63) \quad \phi(n) = \lim_{M \rightarrow \infty} \frac{1}{2M+1} \sum_{m=-M}^M f(m)f(m+n),$$

and eq. (2) becomes

$$(64) \quad \Phi(\omega) = \sum_{n=-\infty}^{\infty} \phi(n)e^{-i\omega n}.$$

The factorization problem of eq. (17) is modified somewhat, because  $\Phi(\omega)$  is more likely to be an empirically determined rather than an analytically expressed quantity.

Wiener (Ref. 1) has indicated various means for handling this problem; one method is the following.

Equation (17) may be written:

$$(65) \quad \log \Phi(\omega) = \log \Psi(\omega) + \log \Psi^*(\omega).$$

Also  $\log \Phi(\omega)$  may be expanded into a Fourier series, as

$$(66) \quad \log \Phi(\omega) = \sum_{n=-\infty}^{-1} a_n e^{-i\omega n} + a_0 + \sum_{n=1}^{\infty} a_n e^{-i\omega n}$$

With  $\Phi$ , and therefore  $\log \Phi$ , real, the series shows symmetry between positive and negative  $n$ 's. From the discussions regarding eqs. (10), (11), (12), and (17) to (20), and recognition that the first term at the right of eq. (66) shows no response at positive times ( $n > 0$ ), and the third term no response at negative times ( $n < 0$ ), the terms of eq. (66) may be identified with those of eq. (65). That is,

$$(a_0/2) + \sum_1^{\infty} a_n e^{-i\omega n}$$

are identified with  $\log \Psi(\omega)$ , and

$$(a_0/2) + \sum_{-\infty}^{-1} a_n e^{-i\omega n}$$

with  $\log \Psi^*(\omega)$ .

This permits the computation of  $\Psi(\omega)$  and  $\Psi^*(\omega)$  from the Fourier series of eq. (66). The Fourier series itself may be obtained empirically by numerical computation or by the use of a harmonic analyzer.

Other empirical solutions of the factorization problem have been presented. Some of these relate to the determination of the minimum phase

of a network which shows an empirical transfer amplitude response characteristic (Refs. 6 and 9).

Additional empirical solutions for the factorization have been presented, related to other problems (Ref. 10).

**Alternatives to Wiener Criterion of Optimization.** Lee (Ref. 11) has explored the possibilities of an alternative to the Wiener criterion. He minimizes the integral square cross-correlation error (integrated with respect to time). He expresses the conditions in terms of an autocorrelation of the autocorrelation function (as if the latter were itself a signal) and a cross-correlation of the autocorrelation function and the cross-correlation function. In these terms the specifications for the optimum filters are completely analogous to those of Wiener.

Zadeh and also Middleton and van Meter have outlined possibilities in the design of an optimum filter which uses the methods of decision theory. (See Ref. 12.) This utilizes all the information known to the designer about the signal and the noise, and optimizes the decision. This optimum minimizes the integrated risk of wrong decisions. The results obtained are chiefly of conceptual value, because the computational problems are formidable even in relatively simple situations.

Other authors have also applied criteria of or akin to those used in decision theory (Ref. 13). The procedures are particularly effective conceptually where the signal interpretation occurs under extremely unfavorable noise conditions, such as for the signals from fringe areas of search radars. The paper of Middleton and van Meter (Ref. 12) contains a complete bibliography.

Zadeh and Ragazzini (Ref. 14) have extended the Wiener theory to the case where the data are described by a nonstationary time series. Particularly they assume its approximation by a polynomial of a given order in time, with unknown coefficients. Such cases have a certain practical interest. The treatment uses an approach suggested in a report by Bode, Blackman, and Shannon, in 1948, to the Research and Development Board.

Chang (Ref. 15) has enlarged the Wiener criterion by considering also integrated squares of errors in the frequency domain. The integration further weights these errors according to arbitrary functions of the frequency. He develops two theorems, a minimization theorem and a separation theorem. The first is an extension of the Wiener theorem in terms of frequency, and the second represents an alternative procedure to the factorization methods of Wiener.

Still other authors (Ref. 16) have given consideration to extending the hypotheses under which the Wiener criterion can be used. In particular they have extended the treatment to include nonstationary noise and a system having time-varying linear parameters.

**Nonlinear Prediction.** The discussion so far has centered on a filter which performs linear operations on the signal input to it. That is, it multiplies the Fourier components of that input by a given numerical factor and shifts their phase by a given angle. Both of these vary with the frequency of the component, but they are independent of the amplitude of the component.

Some thought has been given by Bode and Shannon (Ref. 3) and others (Ref. 17) to the possibilities of nonlinear prediction, in which the operation on the signal would vary with the signal amplitude. At the expense of functional complications, this permits improvement in the accuracy of prediction under certain conditions. The problem has not been worked out to anything like the analytical detail devoted to linear prediction.

## 6. NETWORK SYNTHESIS

**Introduction.** Many of the problems discussed so far lead to a solution in the form of an electrical filter. Specification of the transfer properties of this filter comes out as the end product of the solution in terms of frequency as  $K(\omega)$ , or of time as  $k(t)$ . In an actual case a further step is necessary, namely the network synthesis. The filter or network must be built, and it needs specification in terms of components.

This is an art with an extensive background and innumerable ramifications. The scope of the present discussion is limited to electric networks, with some references for amplification. In data signal processing, on occasion the mechanical properties of equipment may affect signal propagation. Some discussions of electromechanical elements have been given by Everitt and Anner, and by Graham (Ref. 18).

The stage of analysis considered at this point bridges some of the steps from a theoretical toward a schematic design of the equipment. Some of the solutions which have been advanced in the present chapter, particularly to problems which are largely of prediction and exclude noise, are essentially formulas for analog computation (see Vol. 2, Analog Computers).

**Components of Electric Networks.** The elements composing electric networks within the limited scope of this treatment consist of resistances, inductances, and capacitances. These elements are marked by various relations between voltage across them and current flow.

The voltage may be set up as a function of time, as

$$v(t) = v_0 \cos(\omega t + \phi)$$

or as a function of radian frequency, as

$$V(\omega) = v_0 e^{i\phi}.$$

Here it is taken as a complex quantity. The current may be similarly expressed.

The relationship between the two, for a resistance, is

$$V(\omega) = RI(\omega),$$

where  $R$  is the value of the resistance (idealized as independent of frequency). For an inductance the relation is

$$V(\omega) = i\omega LI(\omega)$$

and for a capacitance it is

$$V(\omega) = I(\omega)/(i\omega C).$$

The quantity

$$Z(\omega) = V(\omega)/I(\omega)$$

is called the *impedance*, and its reciprocal

$$Y(\omega) = 1/Z(\omega) = I(\omega)/V(\omega)$$

is called the *admittance*.

The impedance (or admittance) can be expressed for any aggregation of interconnected elements ending in two terminals. As such it may be called the "driving point impedance" (or admittance) of that network at the specified pair of terminals. In an aggregation it is also possible to measure the voltage at one pair of terminals, and the current at another pair. Here the voltage to current ratio is called the *transfer impedance* between the two pairs of terminals, and the current to voltage ratio the *transfer admittance*.

**Specification in Terms of Transfer Response.** The properties required in filters such as have been specified in Fig. 3, or in eqs. (16), (26), (43) and others, have been expressed as a transfer response. When a function of frequency, it has been called  $K(\omega)$ , and when of time,  $k(t)$ . According to the circumstances of the particular equipment considered, the response as a function of frequency may be set up as a transfer impedance or a transfer admittance. It may also be a transfer ratio merely of voltages, or of currents.

For simplicity, consideration here is limited to a case practical in vacuum tube circuitry, where the input to the filter is taken as a current, and the output from it a voltage. That is,  $K(\omega)$  is identified with a transfer impedance, or  $Z_T(\omega)$ .

**Cauer's Method of Synthesis.** In Fig. 15 a vacuum tube gives output  $I$  into the filter. The voltage  $V$  across the terminating resistance  $R$  drives a succeeding vacuum tube. The figure shows a generally practical case of the filter both starting and ending with a bridged capacitance. The filter itself therefore comprises an odd number of elements. If in any case

it is desired to omit one of the end capacitances it may be assumed to approach zero.

For a filter of this type, the transfer admittance is a rational function of  $\omega$  with  $n$  poles and no zeros, and can be written as

$$(67) \quad Z_T(\omega) = \frac{Z_0}{(\omega - \omega_1)(\omega - \omega_2) \cdots (\omega - \omega_n)}$$

Cauer (Ref. 19) has indicated a method for synthesizing the network from this function, which has been noted by Peless and Murakami (Ref. 20).

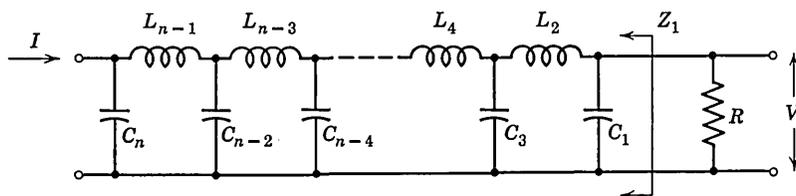


FIG. 15. Low-pass ladder filter.

For this,  $Z$  is changed to a function of  $p = i\omega$ , thus

$$(68) \quad Z_1(p) = \frac{R}{b_n p^n + b_{n-1} p^{n-1} + \cdots + b_1 p + 1}$$

The normalizing factor becomes  $R$ .

Then the driving point impedance  $Z_1(p)$  of Fig. 15 (excluding the terminating resistance  $R$ ) is found by dividing the even part of the denominator by the odd part, and multiplying the whole by the normalizing factor  $R$ .

$$(69) \quad Z_1(p) = R \frac{b_{n-1} p^{n-1} + b_{n-3} p^{n-3} + \cdots + b_2 p^2 + 1}{b_n p^n + b_{n-2} p^{n-2} + \cdots + b_1 p}$$

$Z$  is then expanded into a continuing fraction, as

$$(70) \quad Z_1(p) = \frac{R}{g_1 p + \frac{1}{g_2 p + \frac{1}{g_3 p + \cdots \frac{1}{g_n p}}}}$$

$Z_1$  is found from the elements of Fig. 15 as

$$(71) \quad Z_1(p) = \frac{1}{pC_1 + \frac{1}{pL_2 + \frac{1}{pC_3 + \dots + \frac{1}{pC_n}}}}$$

Thus

$$(72) \quad \begin{aligned} C_1 &= g_1/R, \\ L_2 &= g_2R, \\ C_3 &= g_3/R, \\ &\dots \\ C_n &= g_n/R. \end{aligned}$$

ILLUSTRATION. The solution of the illustration expressed by eq. (43) represents an especially simple case.

$$(73) \quad Z_T(p) = \frac{\kappa R}{(\epsilon/\sqrt{1 + \epsilon^2})p + 1}$$

Here  $\kappa$  represents a constant factor or gain adjustment to be set when lining up the equipment.

$$(74) \quad Z_1(p) = \frac{R}{(\epsilon/\sqrt{1 + \epsilon^2})p}$$

$$(75) \quad C_1 = (\epsilon/\sqrt{1 + \epsilon^2})/R.$$

**Butterworth-Thomson Filters.** A few general characteristics of the performance of filters like Fig. 15 may be noted. When a stepped wave signal, as indicated in Fig. 16a, is used as input, the output signal has the essential character of Fig. 16b. Where the input amplitude is current, and the output amplitude is voltage, this trace in Fig. 16b is called the *indicial impedance* of the filter.

The trace is distinguished by a *rise time*, which measures the duration between crossings of 0.1 and 0.9 of the final amplitude. (Somewhat different ranges are occasionally used.) The trace is also distinguished by an *overshoot*. This is measured as a per cent of the final amplitude.

In the design of such filters for general purposes (and in the absence of a specific formulation like the Wiener equations) it is usually desirable to conserve the shape of the input signal as much as feasible. This is obtained with a short rise time and low overshoot.

It is similarly desirable to conserve frequency space. In terms of a response characteristic, such as illustrated in Fig. 9, the region where the response is large is called the *passband*. The region where it is small is called the *elimination band*. An intermediate region is called the *rolloff*

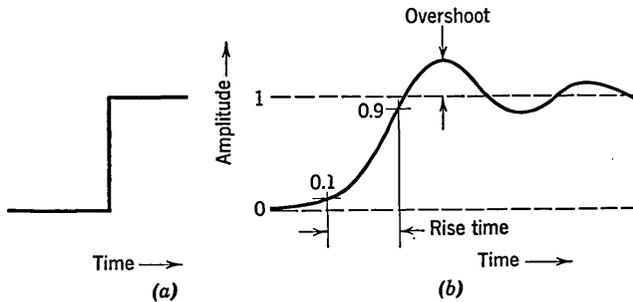


FIG. 16. Passage of step function signal through filter: (a) input, (b) output.

*band*. Conserving frequency space consists in limiting the total band, comprising passband and rolloff band together. For a given passband it means limiting the ratio of the rolloff bandwidth to the passband width. Such conservation is significant in that it represents a system cost, in money, bulk of apparatus, and other factors, to maintain transmission over a wider frequency band than necessary.

Within a given frequency space (passband plus rolloff band) rise time and overshoot conditions are mutually antagonistic. A plot of one versus the other, for some filter designs of the type of Fig. 15, is shown in Fig. 17. Here the rise time is normalized in a manner discussed below.

A series of filter designs was presented by Butterworth (Ref. 21) characterized by a minimum of curvature of the response characteristic in the passband and called *maximally flat amplitude*. This is plotted in Fig. 17 for  $m = 0$  (Butterworth). Here  $n$  has the same meaning as in Fig. 15. The design condition leads to generally short rise time, but fairly high overshoot.

A similar series of designs was presented by Thomson (Ref. 22) characterized by a minimum of curvature in the phase characteristic and called *maximally flat envelope delay*. It is plotted in Fig. 17 for  $m = 1$  (Thomson). This design condition leads to generally higher rise times, but lower overshoots.

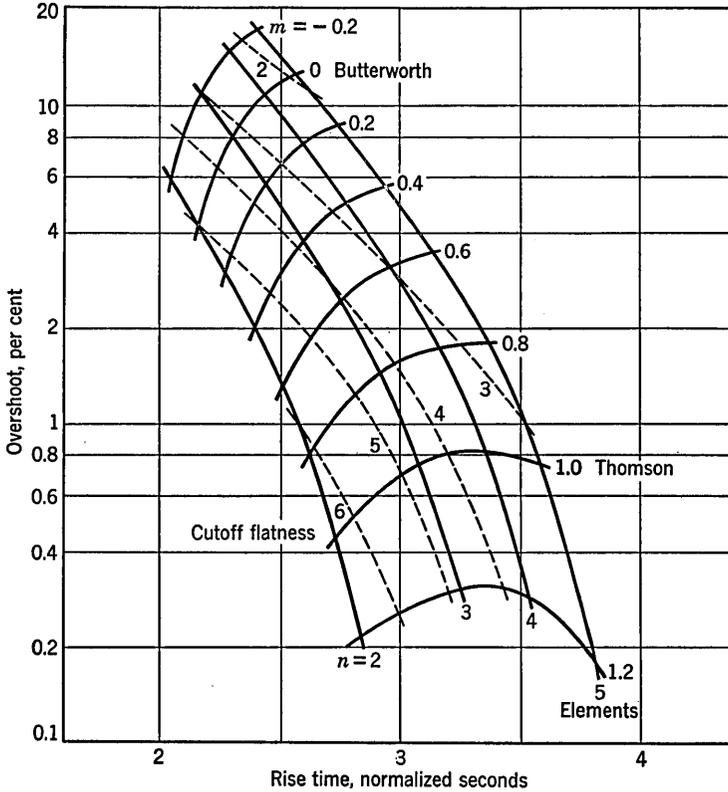


Fig. 17. Design curves, transitional Butterworth-Thomson filters.

**Transitional Butterworth-Thomson Filters.** Peless and Murakami (Ref. 20) have prepared a series of designs intermediate between these two, by degrees indicated by the parameter  $m$ . Their rise time versus overshoot performance is plotted in Fig. 17.

In all these filter designs the ratio of rolloff bandwidth to the passband width tends to reduce as the number of elements is increased. This is why the better compromises between rise time and overshoot in Fig. 17 appear for the smaller numbers of elements.

A rough measure of the utilization is given by the ratio of the frequency for which the drop in response is large (say 30 db, or an amplitude ratio of 1 to 31.6) to the frequency for which the drop is small (say 3 db, or an amplitude ratio of 1 to 1.41). In the figure this quantity is called the *cutoff flatness*; it is plotted in dotted lines. Each line indicates a locus of compromises, between rise time and overshoot, for a given fixed degree of frequency band conservation.

The normalized rise time has been taken, in the Butterworth filters, to indicate the rise time of a filter whose response has dropped 3 db at a radian frequency of 1 radian per second (or a cyclic frequency of  $1/(2\pi)$  cycles per second). The rise times for the other filters are for designs whose amplitude response characteristics in the elimination band are asymptotic to those of the Butterworth filters. Where  $m$  is positive, the 3-db points of the filters come at lower frequencies than for the Butterworth filter. The exact amounts are indicated by  $\omega$  in Tables 1 to 4.

TABLE 1. TWO-ELEMENT FILTERS

$m$	-0.2	0	0.2	0.4	0.6	0.8	1.0	1.2
$\omega$	1.054	1	0.949	0.902	0.859	0.820	0.786	0.756
$C_1\omega R$	0	0	0	0	0	0	0	0
$L_2\omega/R$	0.747	0.707	0.673	0.643	0.618	0.596	0.577	0.561
$C_3\omega R$	1.338	1.414	1.486	1.554	1.618	1.677	1.732	1.782

TABLE 2. THREE-ELEMENT FILTERS

$m$	-0.2	0	0.2	0.4	0.6	0.8	1.0	1.2
$\omega$	1.064	1	0.933	0.868	0.810	0.756	0.712	0.671
$C_1\omega R$	0.525	0.500	0.478	0.458	0.441	0.425	0.411	0.398
$L_2\omega/R$	1.393	1.333	1.287	1.252	1.224	1.201	1.184	1.168
$C_3\omega R$	1.368	1.500	1.624	1.743	1.854	1.958	2.055	2.148

TABLE 3. FOUR-ELEMENT FILTERS

$m$	-0.2	0	0.2	0.4	0.6	0.8	1.0	1.2
$\omega$	1.064	1	0.924	0.845	0.774	0.712	0.659	0.617
$C_1\omega R$	0	0	0	0	0	0	0	0
$L_2\omega/R$	0.399	0.383	0.368	0.354	0.342	0.330	0.320	0.311
$C_3\omega R$	1.127	1.082	1.043	1.008	0.978	0.951	0.928	0.907
$L_4\omega/R$	1.643	1.577	1.535	1.508	1.492	1.484	1.481	1.482
$C_5\omega R$	1.353	1.531	1.698	1.856	2.004	2.143	2.273	2.395

TABLE 4. FIVE-ELEMENT FILTERS

$m$	-0.2	0	0.2	0.4	0.6	0.8	1.0	1.2
$\omega$	1.064	1	0.916	0.824	0.740	0.671	0.617	0.572
$C_1\omega R$	0.321	0.309	0.298	0.288	0.279	0.270	0.262	0.255
$L_2\omega/R$	0.928	0.894	0.864	0.836	0.811	0.788	0.767	0.747
$C_3\omega R$	1.435	1.382	1.338	1.300	1.269	1.243	1.221	1.203
$L_4\omega/R$	1.770	1.695	1.656	1.640	1.638	1.646	1.659	1.676
$C_5\omega R$	1.323	1.545	1.751	1.945	2.125	2.295	2.453	2.602

**Design Data.** The information given above permits making a general compromise choice of filter for any given situation. The specific compromise choice depends upon what ultimate use is made of the signal. Where

the timing indication is important, the compromise would favor low rise time as against low overshoot. Where amplitude indication is important, the reverse holds. Where frequency conservation is important as compared with a more favorable rise time versus overshoot compromise, or where other specifications indicate need of sharpness in cutoff, enough elements to do this may be chosen.

The element values, taken from the Peless-Murakami paper (Ref. 20), are listed in Tables 1 to 4. The values are normalized as for the rise times, and are also normalized to a termination resistance  $R$ . Thus the entry of 1 for  $C\omega R$  means that  $C$  alone is 1 farad divided by  $2\pi$  times the cyclic frequency of the Butterworth 3-db cutoff, and again divided by  $R$ . If that cutoff is 1000 cycles, and  $R$  is 1000 ohms,

$$(76) \quad \begin{aligned} C &= 1 \times (2\pi \times 10^3)^{-1} \times 10^{-3} \text{ farad} \\ &= 0.159 \text{ microfarad.} \end{aligned}$$

Similarly the entry 1 for  $L\omega/R$  means that  $L$  alone is 1 henry times  $R$  divided by  $2\pi$  times the cyclic frequency of the Butterworth 3-db cutoff. Again for a cutoff of 1000 cycles, and  $R$  of 1000 ohms.

$$(77) \quad L = 1 \times (2\pi \times 10^3)^{-1} \times 10^3 = 0.159 \text{ henry.}$$

The even element filters are all designed for  $C_1 = 0$ .

**Tchebysheff-Darlington Filters.** More general discussion of the synthesis of networks has been presented by Darlington (Ref. 23) and by Grossman (Ref. 24). The procedures there employed lead to the use of mathematical contributions of Tchebysheff, and the filters have been called Tchebysheff-Darlington filters. Essentially in the papers referred to, the design is specifically applied to filters in which tolerances are placed on a permissible variation in transmission over the passband, and on a required completeness of suppression in the elimination band.

## REFERENCES

1. N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, Wiley, New York, 1949. (Published in a Classified Report in 1942.)

See also:

A. Kolmogoroff, Interpolation und Extrapolation von stationären zufälligen Folgen, *Bull. acad. sci. U.R.S.S., Sér. math.*, 5, 3-14 (1941).

H. Jacot, Théorie de la Prévision et du filtrage des séries aléatoires stationnaires selon Norbert Wiener, *Ann. Télécommunications*, 7, 241-249, 297-303, 325-335 (1952).

2. N. Levinson, A heuristic exposition of Wiener's mathematical theory of prediction and filtering, *J. Math. and Phys.*, 26 (2), 110-119 (1947). (Reprinted in Ref. 1.)

3. H. W. Bode and C. E. Shannon, A simplified derivation of linear least square smoothing and prediction theory, *Proc. I.R.E.*, **38**, 417-425 (1950).
4. N. Wiener, Generalized harmonic analysis, *Acta Math.*, **55**, 117-258 (1930).
5. H. Nyquist, Certain topics in telegraph transmission theory, *Trans. Am. Inst. Elec. Engrs.*, **47**, 617-644 (1928).
6. H. W. Bode, *Network Analysis and Feedback Amplifier Design*, Van Nostrand, Princeton, N. J., 1945.
7. G. A. Campbell and R. M. Foster, *Fourier Integrals for Practical Applications*, Collected Papers, American Telephone and Telegraph Company, New York, 1937; also *Bell System Tech. J.*, **7**, 639-707 (1928).
8. G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Macmillan, New York, 1944.
9. D. E. Thomas, Tables of phase associated with a semi-infinite unit slope of attenuation, *Bell System Tech. J.*, **26**, 870-899 (1947).
10. E. O. Powell, An integral related to the radiation integrals, *Phil. Mag.*, **34** (7), 600-607 (1943).
  - A. Fletcher, Notes on tables of an integral, *Phil. Mag.*, **35** (7), 16-17 (1944).
  - F. W. Newman, *The Higher Trigonometry, Superrationals of Second Order*, Macmillan and Bowes, Cambridge, England, 1892.
  - A. Fletcher, J. C. P. Miller, and L. Rosenhead, *An Index of Mathematical Tables*, Scientific Computing Service, Ltd., London, 1946.
11. Y. W. Lee, On Wiener filters and predictors, *Proceedings of the Symposium on Information Networks*, April 12-14, 1954, Vol. III, pp. 19-29, Polytechnic Institute of Brooklyn, New York.
12. L. A. Zadeh, General filters for separation of signal and noise, *Proceedings of the Symposium on Information Networks*, April 12-14, 1954, Vol. III, pp. 31-49, Polytechnic Institute of Brooklyn, New York.
  - D. Middleton and D. van Meter, Detection and extraction of signals in noise from the point of view of statistical decision theory, Pts. I and II, *J. Soc. Ind. and Appl. Math.*, **3**, 192-253 (1955); **4**, 86-119 (1956).
13. P. M. Woodward and I. L. Davies, Information theory and inverse probability in telecommunication, *Proc. Inst. Elec. Engrs. (London)*, **99**, Pt. III, 37-44 (1952).
  - I. L. Davies, On determining the presence of signals in noise, *Proc. Inst. Elec. Engrs. (London)*, **99**, Pt. III, 45-51 (1952).
  - D. O. North, *Analysis of the Factors Which Determine Signal to Noise Discrimination in Radar*, Rept. PTR-6C, RCA Laboratories, June 1943.
  - G. W. Preston, The design of optimum transducer characteristics using the method of statistical estimation, *Proceedings of the Symposium on Information Networks*, April 12-14, 1954, Vol. III, pp. 51-59, Polytechnic Institute of Brooklyn, New York.
  - L. A. Zadeh and J. R. Ragazzini, Optimum filters for the detection of signals in noise, *Proc. I.R.E.*, **40**, 1223-1231 (1952).
  - J. L. Lawson and G. E. Uhlenbeck, *Threshold Signals*, Mass. Inst. Technol. Radiation Laboratory Series, Vol. 24, McGraw-Hill, New York, 1950.
  - T. G. Slattery, The detection of a sine wave in the presence of noise by the use of a non-linear filter, *Proc. I.R.E.*, **40**, 1232-1236 (1952).
14. L. A. Zadeh and J. R. Ragazzini, An extension of Wiener's theory of prediction, *J. Appl. Phys.*, **21**, 645-655 (1950).
15. S. S. L. Chang, Two network theorems for analytical determination of optimum-response physically realizable network characteristics, *Proc. I.R.E.*, **43**, 1128-1135 (1955).

16. R. C. Booton, An optimization theory for time-varying linear systems with non-stationary statistical inputs, *Proc. I.R.E.*, **40**, 977-981 (1952).
- R. C. Davis, On the theory of prediction of non-stationary stochastic processes, *J. Appl. Phys.*, **23**, 1047-1053 (1952).
- J. Bendat, A general theory of linear prediction and filtering, *J. Soc. Ind. and Appl. Math.*, **4**, 131-151 (1956).
17. A. G. Bose, A theory for the experimental determination of optimum non-linear systems, *I.R.E. Convention Record*, Pt. 4, pp. 21-30, March 1956.
- R. Drenick, A non-linear prediction theory, *Trans. I.R.E.*, **PGIT-4**, 146-152, (Sept. 1954).
18. W. L. Everitt and G. E. Anner, *Communication Engineering*, McGraw-Hill, New York, 1956.
- R. E. Graham, Linear servo theory, *Bell System Tech. J.*, **25**, 616-651 (1946).
19. W. Cauer, Ausgangsseitig Leerlaufende Filter, *ENT*, **16**, 161-163 (1939).
- E. A. Guillemin, A summary of modern methods of network synthesis, in *Advances in Electronics*, Vol. 3, pp. 261-303, Academic Press, New York, 1951.
20. Y. Peless and T. Murakami, Analysis and synthesis of transitional Butterworth-Thomson filters and band pass amplifiers, *RCA Rev.*, **18**, 60-94 (1957).
21. S. Butterworth, On the theory of filter-amplifiers, *Exptl. Wireless and Wireless Eng.*, **7**, 536-541 (1930).
- V. D. Landon, Cascade amplifiers with maximal flatness, *RCA Rev.*, **5**, 347-362 (1941).
22. W. E. Thomson, Networks with maximally-flat delay, *Wireless Eng.*, **29**, 255-263 (1952).
- J. Laplume, Amplificateurs moyenne fréquence à distortion de phase réduite, *L'Onde Electrique*, **31**, 357-362 (1951).
23. S. Darlington, Synthesis of reactance fourpoles, *J. Math. Phys.*, **18**, 257-353 (1939).
24. A. J. Grossman, Synthesis of Tchebycheff parameter symmetrical filters, *Proc. I.R.E.*, **45**, 454-473 (1957).

## Data Transmission

*Pierre Mertz*

1. Introduction and Symbols	18-01
2. Formation and Use of the Electrical Signal	18-07
3. Transmission Impairment	18-18
References	18-30

### 1. INTRODUCTION AND SYMBOLS

**Basic Considerations.** Data which are generated at a given point, either as a result of collecting original information or at the output of a computer as a result of the processing of other data, often have to be transmitted to some other point in order to be used for further data processing or remote control. Two basic parameters determine the extent of the undertaking which this transmission involves.

1. The order of magnitude of the distance between the points of origination and utilization.

2. The nature of the data that are to be transmitted. This includes the information content of the data and the frequency band which is required to handle it in the transmission medium. For fundamental discussion on this point see Chap. 16. The treatment here analyzes principally the current practical art, in which the efficiency of utilization of the frequency band is much lower than ideal.

The adaptability of the data signals to transmission over available facilities is a practical factor of great importance. There are extensive

systems of communication already set up, reaching over vast areas, which are in current commercial use.

**Transmission Distances Involved.** The engineering effort required to set up a system of data transmission varies considerably with the distance involved. Some concrete illustrative gradations are:

- A few inches or feet
- One to a few hundred feet
- One to a few miles
- One to several hundred miles
- Several hundred to several thousand miles
- International or intercontinental facilities

This discussion stresses particularly lengths in the middle regions, from a few miles to a thousand miles.

### **Nature of Transmission Facilities Available**

These facilities need to meet certain requirements, discussed below. Of these the frequency band which they are capable of transmitting is paramount, and is of chief concern here.

**Local Wiring.** The band which this wiring will handle is indefinite, and it varies with the physical structure of the conductors and their inductive exposure to other electrical circuits. Bands have been handled from less than 100 cycles to television bands of a few megacycles.

**Telephone Facilities.** These include all of the plant which has been developed for telephonic purposes, and hence they comprise a wide variety of facilities. They are sometimes nominally characterized as capable of a 3-kc band. This full width band is not usable for data transmission, partly because some of the facilities cut off below this and partly because the lower frequency region, below 1000 cycles, is not likely to be effectively employed in the data transmission (Ref. 1). See Sect. 2 for more quantitative details regarding a usable band.

There are telegraph facilities of narrower band, but since these are usually multiplexed on telephone facilities, they are not considered separately.

**Program Transmission.** These circuits are commercially used for the interconnection of radio broadcast stations. They have frequency bandwidths, in round numbers, of 3, 5, 8, and 15 kc (Ref. 2). As in the case of the telephone facility, the full band cannot be expected to be utilizable for data transmission. Also the commercial demand for the 8- and 15-kc bands is very low, so that there is at present a substantial network of only the 3- and 5-kc bands.

**Television Transmission.** An extensive network exists at present interconnecting television broadcast stations and studios and facilities for theater television (Ref. 3). The bandwidth of these facilities generally runs to a little over 4 Mc. However, on older coaxial cable facilities the bandwidth is only 2.7 Mc. Some experimental facilities of broader frequency band than 4 Mc have been furnished for short period tests, but not on a commercial basis.

**Other Wide Band Conductor Transmission.** For economy, telephone facilities are frequently gathered together in more or less large groups. The combined signal for the entire group is then handled over a wire circuit much as a single signal (Ref. 4). *Groups* of 48-kc bandwidth, and *super-groups* usually of 5 groups, or 240 kc, are handled in this manner. Also, on other types of system, bands of 16 kc are found.

The use of these types of bands would, of course, require the development of arrangements for extending them from the terminals, at offices of the common carriers where they are located, to other premises.

Carrier current facilities are also multiplexed by power companies on power lines. These are suitable for data transmission (Ref. 32).

**Radio Facilities.** Radio facilities naturally present certain elements of flexibility in their use compared with facilities provided over conductors. The limits to this flexibility are, however, set by allocation problems and by the propagation characteristics of the frequency region used (Ref. 5). The frequency bandwidths used run from those for individual channels to large aggregations of multiplexed channels which may include television channels (Ref. 6).

The utilizable bandwidth for the individual channels is not necessarily set by the adjacent allocations. It is often actually set by multipath echo effects. It tends to run from something under a telephone bandwidth (3 kc), up to the general order of magnitude of television channels (6 Mc).

Radio channels that form part of a large aggregation, particularly those leased from common carriers, tend to run at telephone or television bandwidth, and differ little from similar circuits over conductor facilities. Similarly, group and super-group bands of intermediate width are transmitted, but the use of these again requires the development of arrangements for extending them from the terminals.

### Nature of the Data

Data consist fundamentally of two types of information (Ref. 7).

1. Choices among a group of possible conditions. A single datum, such as a room temperature, represents the single choice out of an established gamut. The total possible number of choices in that gamut depends both on the range of the gamut and on the precision of the indication within the

gamut. For *example*, a range of room temperatures may be established between 50° and 90° F, and the indication may be given to individual degrees. Then the datum represents one choice out of a possible 40, and it may be this which is to be transmitted.

2. The timing of one or a series of events (Ref. 31). One might, for *example*, send the equivalent of a clock ticking from one geographical place to another, to assure the simultaneity of astronomical observations made at these places.

In many cases in practice, both types of information may be needed. For *example*, in air traffic control, both the position of a given plane and the time at which it occupies that position, are needed. The position is indicated at the same time that it is occurring. Such a datum is said to be sent in *real time*, as distinguished from sending it much later on as a component in some abstract calculation. Data sent in real time are characterized by becoming "stale," i.e., of losing their value, if delayed too long in transmission.

**Continuous Analog Data.** In the room temperature example cited above, the temperature may be represented by the position of the end of a mercury thread, or by the angular position of a shaft (dial thermometer), or again by the value of a given voltage. This position or voltage is not the actual temperature, but may be identified as analogous to it. Such data, where some different quantity varies proportionately (or according to some other appropriate law of variation) to the quantity desired, are called *analog* data. The demarcations between choices are not emphasized in the datum quantity, but they are important in a statement of the indication.

Where the analog relationship between the utilized data and the original quantity is not interrupted, the data are said to be *continuous*.

**Discontinuous Analog Data.** It may suffice to have the temperature information once every 10 minutes instead of continuously. An analog quantity may be set up in which the relationship to the original quantity is interrupted when not needed. The results are called "discrete" or discontinuous analog data. This may make it possible to interlace other data between the temperature readings.

**Multiple Speed Analog Data.** In the case of a clock it has been found convenient to use the angular position of the shaft of the minute hand to identify one out of 60 choices, or one minute in the hour. For general use, however, a range of 12 hours is desirable. It is not convenient to use a pointer that can identify one out of 720 possible choices. The problem is solved by using two shafts, one geared to the other. The minute hand identifies one out of 60 choices. The hour hand identifies one out of 12 choices, each of which corresponds to one group of the 60 choices of the

minute hand. The principle is sometimes extended by adding a third shaft and hand to read seconds. It is even further extended in conventional watt-hour and gas meters. All these are examples of "multiple speed" or "multiple shaft" analog data.

**Digital Data.** The above process can be carried to its logical conclusion, where each shaft distinguishes only one out of two choices. In this extreme case the demarcations between choices are emphasized, and the choices would more usually be indicated by two-position members rather than by shafts. The choice may be considered as identified by a sequence of binary indications or *binary digits* and the data are called *digital*. Less extreme forms are sometimes used in which one out of three or more discrete choices are indicated by each digit.

Digital information may be transmitted over a group of wires, by assigning one digit to each wire. This is known as *parallel* transmission. Or the various digits may be assigned to successive ordered pulses (or spacing intervals) on a single wire. This is called *serial* transmission, and the series of digits may be ordered in either direction. *Examples.* The digits indicating large values may come first (as in reading decimal digit Arabic numerals), or those indicating small values may come first (as in adding or multiplying operations with Arabic numbers).

**Timing Data.** This information can be indicated in a variety of ways. More usually the desired time is indicated by the wave front of an appropriate transition, say in voltage.

**Starting and Other Auxiliary Information.** In the example just given above, where successive digits are transmitted serially, it is usually desirable to identify the start of the sequence by some auxiliary information. At other times the auxiliary information is in the nature of a pilot, reference or calibration datum against which the magnitude of the utilized data are compared before actual use. Other auxiliary information is sometimes needed for error checking or possible other purposes.

**Error Standards.** It is not generally expected that data transmission will be completely perfect. For one or another reason, errors are caused. Thus in engineering a given system there is some need to give thought to what kind of error performance will be acceptable.

In the case of analog data although the boundaries between successive choices are not emphasized, the spacing between the choices is important. This spacing is obtained from the precision which is found useful in the data. It is expected that this precision will be maintained in the transmission of the data. It is common to express the error expected or experienced, in terms of its *root-mean-square* (rms) value (see Chap. 13). Occasionally a *maximum* error is noted, often say three times the rms

value. In a Gaussian distribution, errors larger than this occur with a frequency of about 1 part in 370.

Timing errors are measured by a similar rms or maximum displacement, but in a timing variable rather than an amplitude parameter (Ref. 31).

In the case of digital data an elementary measure of the error is the frequency of occurrence of errors in the binary digits in the data at the receiver. On occasion, a more sophisticated measure is desirable which takes account of the distribution of the errors in time. This is because in general, when one digit in a specific group of digits is in error, the usefulness of the entire group is vitiated. Thus errors in close succession, in such cases, do not cause as much ultimate impairment as when they are more scattered. Measures of impairment in such cases are not easily established without some detailed knowledge of the entire scheme for setting up and using the data that are transmitted.

Where special measures are incorporated into the signals for error checking, it is usually convenient to count the frequency of occurrence of both the detected and undetected errors. Of these, the first are apt to constitute only a minor impairment but the second are serious. *Undetected errors* are those not detected during the test, but obtained from some later comparison between the signals actually sent and those actually received.

## Symbols

$a$	relative echo amplitude
$D$	envelope delay
$f$	cyclic frequency
$I$	wave amplitude
$k$	normalizing constant, equal to mean square value of $I$
$M$	matrix of resistances
$p(x)$	probability density of variable $x$
$R$	resistance, with subscripts for specific cases
$r$	amplitude ratio
$r(\omega)$	amplitude ratio at radian frequency $\omega$
$T$	pulse duration time
$t$	time variable
$V$	voltage
$v$	instantaneous signal voltage
$\theta$	pulse separation time (front edge to front edge)
$\lambda$	wavelength of ripple along cyclic frequency scale ( $= \Delta f$ )
$\tau$	echo delay
$\phi, \phi(\omega)$	phase shift at radian frequency $\omega$
$\omega$	radian frequency ( $= 2\pi f$ )

## 2. FORMATION AND USE OF THE ELECTRICAL SIGNAL

**Encoding and Decoding**

The first step in the preparation of a signal for transmission consists in expressing the variable that is intended for transmission into some sort of code that can be used to form the electrical configuration.

**Analog Data.** There is not very much latitude for coding such data, aside from transferring from one type of physical quantity into another. Thus a temperature or a distance may be transformed into a shaft rotation or a voltage. The principal modification that can be introduced is the insertion of some sort of nonlinear relation between the one quantity and the other.

**Digital Codes.** A simple code into which an analog quantity may be converted is the binary digital code (see Chap. 16). A diagram of the 8 choices for a 3 binary digit code is illustrated in Fig. 1. The dark areas

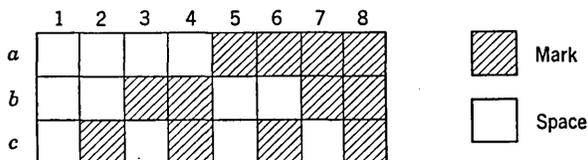


FIG. 1. Diagram of 3-digit binary code selections.

indicate, say, voltage (or current) "on," and the white areas, voltage (or current) "off." They are termed respectively *marking* and *spacing*. A variation of this code sometimes used to simplify an encoding mechanism is the *reflected binary* or *Gray code* (Ref. 8). This is shown in Fig. 2. The simplification in mechanism comes essentially because the change from any given choice to the next adjacent choice involves the change of only one binary digit. Other variations of this simple code type have been devised. One such is a coding to include negative values of the encoded quantity (Ref. 9).

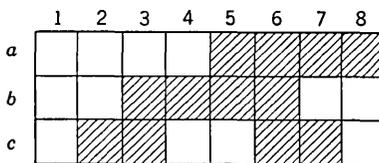


FIG. 2. Diagram of 3-digit reflected binary (Gray) code selections.

More complicated codes have been devised in which present code designation is a function also of past history of the value being encoded, or of more than one variable (Ref. 10).

These complications are conceived principally to condense the information to be transmitted into the most compact code possible. They do involve an increase in cost of equipment, and a loss in time at both en-

coder and decoder that may be important where the transmission operates in real time (see Sect. 1, Nature of the Data). Some economic study has been made of such points (Ref. 11). See also Chap. 16.

**Processes of Digital Encoding.** Only some elementary principles can be mentioned here (Ref. 9). More details of these processes are given in Vol. 2, Chap. 20.

1. A basic method of encoding consists in laying out the analog input along one dimension of a two-dimensional code matrix and reading the coded output along the other dimension. This can be utilized for any arbitrary code. As an illustration the diagrams of Figs. 1 and 2 may represent plates, with holes punched through the shaded squares, in a cathode ray tube (Ref. 12). The electron beam is deflected along the horizontal coordinate by analog voltage input. A subsequent vertical deflection then gives the coded signal, in serial form, on an electrode beyond the plate. The beam goes through the punched holes, but is stopped where no holes exist.

2. A second basic method consists in encoding the analog quantity first into a *unit-counting code*. For each value of the analog quantity to be transmitted a counting mechanism counts and cumulates unit increments up to a value nearest to the input quantity. A unit-counting code is not efficient for transmission since the number of binary digits sent is large. It can be converted into a binary digital code by successive scale of two counting dividers (see Vol. 2, Chap. 20). If other codes are desired a further conversion can be made.

3. A third basic method uses the general principle that any decoder may be used for encoding by associating it with an appropriate inverse feedback path. An arbitrary code indication is set up, say the last previous transmission. This is decoded, and the result is compared with the input. The inverse path mechanism uses the error to step the code in the direction to reduce the error. The stepping mechanism continues until the error is less than the smallest choice interval.

Coding mechanisms in which the present output depends on more than the single present input exhibit a greater variety of types and will not be discussed here.

**Processes of Digital Decoding and Smoothing.** The decoder in general has two broad functions.

1. Decoding proper is to convert the digital indication into an analog indication. In nearly all cases this appears as an individual analog indication when the code is received. In some cases this is the only function needed.

2. To *hold* or *store* and possibly *smooth* the analog indications are needed where individual indications are required at more frequent intervals than the code permits, or where continuous analog indications are required.

The *decoding function* proper may be classified by types of mechanism, as for the encoder. For the moment these are limited to the case where a single input leads to a single output.

(a) In a basic type of decoder the choices indicated by the respective binary digits lead to the single element of an arbitrarily prearranged *matrix*. This element translates to its prearranged analog output.

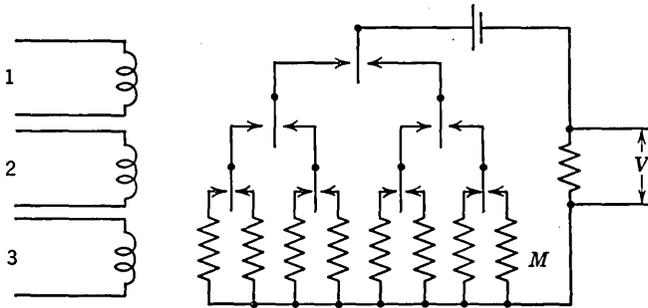


FIG. 3. Relay matrix for 3-digit code.

An *example* of this is shown in Fig. 3 in terms of relays. The respective digits operate relays 1, 2, and 3. Any given choice leads to some resistance of the matrix  $M$ . These are chosen in advance to yield the desired analog voltage  $V$  at the output, for the given choice.

(b) A variation of this is applicable to codes where the successively ordered digits contribute proportioned *weights* to a cumulation of the

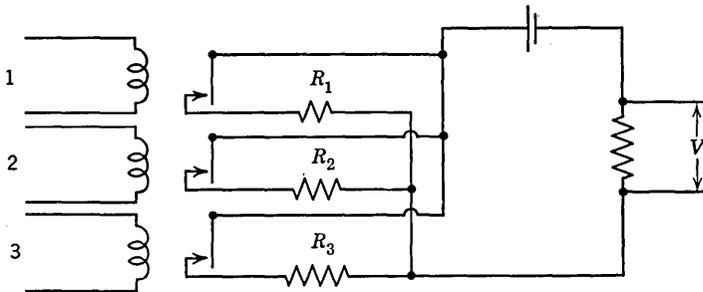


FIG. 4. Relay matrix for 3-digit binary code.

analog total. This occurs in the binary digital code. An *example* is shown in Fig. 4. The successively ordered digits choose respective resistances  $R_1$ ,  $R_2$ , and  $R_3$ . These are proportioned to cumulate currents, in the progressive ratios of 4, 2, and 1, in the output resistance, which must be

low compared to the  $R$ 's to keep the contributions independent. The output voltage  $V$  gives the analog for the binary digital choice.

(c) If instead of current contributions, successive pulse counts are cumulated, the process leads to a translation from a binary digital to a unit-counting code. This can then translate further to the final analog quantity. Such an arrangement is the inverse of the second basic encoding process.

(d) Finally an encoder may be inserted in an inverse feedback path for conversion into a decoder. An arbitrary analog output is set up, say the last value just previously decoded. This is encoded, and the code indication is compared with the present input code. The inverse feedback path uses the difference to change the analog value in the direction to reduce the difference. Several steppings of the code may be needed until identity is secured. This basic method is not easily applicable to arbitrary codes.

*Holding the analog signal* requires some form of temporary storage (Ref. 13). Where the error objective calls for a more accurate interpolation between the discrete values, still more equipment is needed. The process has been here called *interpolation*, but it is clear that after one discrete value has been obtained and before the next is available, the process really required is *extrapolation* or *prediction*.

The principles are described in Chap. 17 for the optimizing properties required in the above processes. More than just an electrical filter may be needed, because of the discrete character of the values.

Where the data are such that the best correlation occurs between successive values of the analog variable, a mere holding, or zero order prediction, is optimum. Where, as is quite possible in practice, a good correlation holds between successive rates of change (or velocities), a first order predictor is better. This predicts from the velocity as derived from past data. Where a good correlation held on the accelerations, a second order predictor would be called for.

This second function of the decoder may be used by itself in cases where the data were merely sampled as discrete analog data, and not digitally encoded.

**Error Detection Codes.** It is possible to introduce deliberate redundancy into the code used in the data transmission path. This establishes auxiliary relationships. At the receiver a test may be made for these auxiliary relationships. When they are found missing, the fact is an indication of error in the transmission.

A simple form of this redundancy is the *parity check* (Ref. 14). For this the message is divided into successive groups of binary digits, and an extra digit is provided at the end of each group for the redundant informa-

tion. The number of marks in the group is noted as being even or odd. If even, the added digit is made marking; if odd it is made spacing, to make the total always odd. Hence the reception of a total even number of marks indicates an error. Undetected errors can exist when two errors conspire to maintain the total odd. The system can also be arranged to make the correct total always even.

Another *example* is the *2 out of 5 code*. Here 5 binary digits are always disposable for the signal, and 2 of these are always made marking. This gives 10 combinations, which is very handy for translation from and to a decimal digit code. When the receiver receives any other than 2 marks, it indicates an error. Two errors can combine here also to evade detection. The code may also be used with 3 marks out of the 5 binary digits.

The redundancy may be increased to the point where the specific digit in error may itself be located in the signal, and therefore corrected. This is an *error-correcting code*. Some combinations of errors exist that can be detected by this, but not corrected. Even rarer combinations are possible which evade detection.

### Modulation and Multiplexing Methods

Several steps must be considered in these processes.

**Baseband Signal.** The information which is to be transmitted from one point to the other eventually appears in the form of an electric amplitude (say a voltage or a current) before it is propagated over the transmission medium. In a continuous analog system it appears, say, as a continuously varying d-c voltage. In a discrete analog system it appears as a succession of pulses of varying voltage amplitudes. In a binary digital system it appears as a succession of pulses each individually of either marking or spacing voltage amplitude.

The signal in this form is called a *baseband signal*. It has a frequency spectral distribution of power which goes down to and includes zero frequency (or d-c). Its amplitude distribution depends upon the shaping of the individual pulses or *shape factor* (where pulses are involved) and upon the sequence of amplitudes which codifies the information or *discrimination factor* (Ref. 15).

Nyquist has shown (Ref. 15) that the *complex amplitude at each frequency is equal to the product of these two factors*, each of which is complex. In a code that gives sufficient randomness to the signal, and that permits positive and negative voltage values, with an average of zero, the long time average of the power distribution in the discrimination factor is flat over the frequency range. In such a system, therefore, the signal power distribution is equal to that for the shape factor, or for a single pulse (aside from a normalizing factor).

For an idealized pulse with rectangular sides, such as shown at *A* in Fig. 5, the frequency band is infinite, as illustrated by the full line. However, most of the power is located below the frequency  $1/T$ , where  $T$  is the pulse duration.

Practical pulses are in general rounded somewhat as shown at *B* of Fig. 5. For these, all but a negligible proportion of the power is located

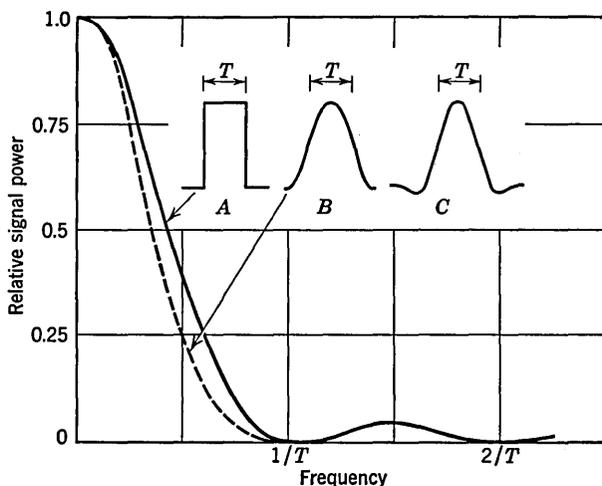


FIG. 5. Power spectra of various pulses.

below frequency  $1/T$ , as illustrated by the dotted line. The pulse form at *C* indicates that obtained from the full line spectrum, cut off to zero above frequency  $1/T$ .

**Nominal Effective Band.** Nyquist has further shown (Refs. 10 and 15) that the *minimum frequency band required to transmit independent amplitudes for each pulse, where the pulse separation is  $\theta$ , is  $1/(2\theta)$* . This may be called a *nominal effective band*. In practice a somewhat larger band is generally used. If the successive pulses are set up edge to edge (that is,  $\theta$  made equal to  $T$ ), then in the full line of Fig. 5, the nominal effective band reaches from 0 to  $1/(2T)$ . The band actually used in practice usually reaches to  $1/T$ , or twice as far. The part of the band between  $1/(2T)$  and  $1/T$  transmits a portion of the signal that has low power, and may be designated as *rolloff band*.

Occasionally a narrower rolloff band than that reaching to  $1/T$  may be used. This results in oscillatory transients or "ringing" before and after each pulse of the signal.

Where short pulses are transmitted at infrequent intervals,  $\theta \gg T$ . In such cases a wider frequency band is used than necessary for the informa-

tion, and  $1/T \gg 1/\theta$ . Additional pulses from other channels of information can be interlaced in between, to use the frequency space more fully. It is found in Sect. 3, Tolerances, that this leads to the need for high fidelity in the transmission.

**Amplitude Modulation.** Except for direct wire transmission over short distances it is not usually practicable to transmit a baseband signal of the spectral distribution illustrated in Fig. 5. This is because it involves transmission all the way down to and including d-c.

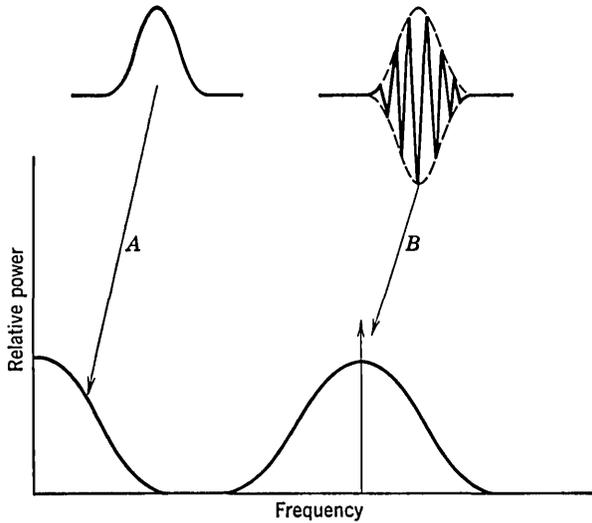


FIG. 6. Spectra of baseband and carrier pulses.

A simple procedure to avoid the need of d-c is to use the baseband signal to form the envelope of a carrier wave, as shown at *B* in Fig. 6. Here the spectrum becomes that of the carrier frequency and two symmetrical sidebands. Each of these has the same shape as the baseband. The baseband or envelope signal is recovered at the receiver usually by the use of a rectifier.

Certain precautions are needed when using a carrier signal in this way. Interferences are likely to develop between the lower sideband and the baseband, if the carrier is placed at such a low frequency that they overlap. Each of these interferences can, if needed, be reduced at the source. The more usual procedure, however, is to allocate the spectrum to avoid such an overlap.

Another point to be noted is that over certain types of telephone facilities, in which another carrier wave is used for the transmission, the data

signal carrier may not be reproduced at its exact frequency. The received signal may be displaced up to two cycles per second from that transmitted (in some older facilities this may be some 20 or 30 cycles per second). Thus the system cannot be designed in such fashion that the reproduction of this exact frequency is critical. For *example*, the carrier frequency cannot be depended upon for use as a synchronizing frequency.

**Vestigial Sideband Transmission.** The information carried in one sideband is duplicated in the other. Thus only one of the two is necessary

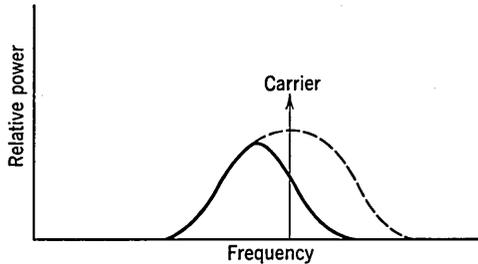


FIG. 7. Vestigial sideband spectrum.

for transmission, and a saving in frequency space required in the transmission medium is achieved by suppressing the other sideband. However, cutting off a band sharply at the carrier is difficult in data transmission, where the power spectrum contains frequencies close to and including the carrier. To solve this Nyquist (Ref. 15) has indicated a sloping cutoff as shown in Fig. 7. This retains a "vestige" of the suppressed band, and is, therefore, called *vestigial sideband transmission*. The cutoff through the carrier region introduces interfering spurious components to the signal. These are called "quadrature" components because in their interfering effect they add in quadrature to the undistorted signal. The interference usually causes an impairment in signal-to-noise ratio, which is discussed more fully in Sect. 3, Quadrature Component in Vestigial Sideband Transmission.

When all the precautions which have been discussed are allowed for, it is found that with double sideband modulation *a signaling speed of about 650 signal elements (each of duration  $T = 1.54$  milliseconds in Figs. 5 and 6) per second may be transmitted over a substantial proportion of telephone circuits* (Ref. 16).

*With vestigial sideband transmission about 1600 signal elements per second have been transmitted over selected and suitably treated telephone circuits* (Ref. 17). This is slightly more than double that with double sideband. The increase comes mostly from the use of the vestigial band, but in part from selection and treatment of circuits.

**Frequency Modulation.** Characteristics of the carrier wave other than its envelope amplitude may be varied in accordance with the base-band signal. A common example is the variation of its instantaneous frequency. This can have certain advantages, for *example*, when transmitting over a medium whose amplitude response at the receiver varies from instant to instant. A more detailed analysis also shows that transmission by frequency modulation is less subject to impairment from noise than by amplitude modulation (Ref. 18).

**Other Methods.** Still other characteristics of the carrier wave may be varied to indicate the signal (Ref. 19). Phase modulation may be used. Or the data signal (whether itself constituted of pulses or not) may be transmitted over a medium that uses a pulse code form of modulation (Ref. 20). In this case the instantaneous amplitudes of the data signal are reproduced by a secondary pulse code. However, the requirements for this mode of transmission have not as yet been worked out. The range of different possibilities is very great.

**Frequency Division Multiplex.** In Fig. 6 a second carrier with its sidebands may be placed at a higher frequency than  $B$ , and transmit an

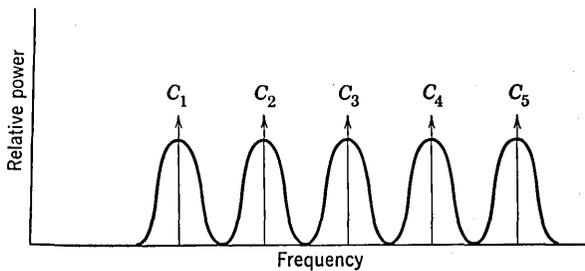


FIG. 8. Carrier signal spectra, frequency discrimination.

independent signal. More carriers can, of course, continue to be added, as suggested in Fig. 8. The limit is the frequency band available on the transmission system. This is known as *multiplexing by frequency division*.

A consequence of this form of multiplexing lies in modulation products which are generated between pairs (and larger groups) of the simultaneously operating channels. These products can cause interference into the channels in which they fall. The modulation arises from nonlinearity in the transmission process, at a possible variety of points according to the details of the facility used. Engineering precautions are needed to keep the interference down to an acceptable level.

**Time Division Multiplex.** Where a signal uses a basic pulse which is repeated at much longer intervals than its own duration, other independent

signals may use pulses which are inserted intermediate between these. A scheme of five such channels is indicated in Fig. 9. The limit is, of course, fixed by the relative durations of the individual pulses, the spacing interval between pulses of the same channel and the guard space which is required between pulses of one channel and of its nearest neighbors to prevent mutual interference:

The choice of whether frequency or time division multiplexing is preferable in a given case depends upon the nature of the transmission impair-

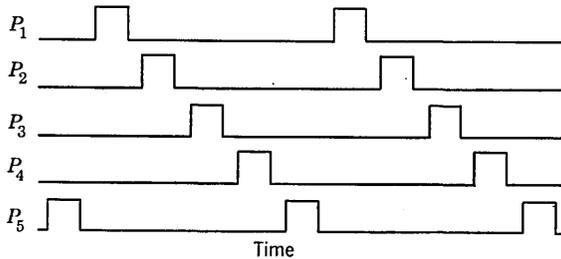


FIG. 9. Pulse signal profiles, time discrimination.

ments to be expected and upon the relative costs. Both methods are in extensive use.

**Other Forms of Multiplexing.** In a generalized study of multiplexing (Ref. 21) it is found that *n independent channels can be multiplexed on a given signal through the superposition of n mutually orthogonal functions.* The arrangements discussed above represent two possible solutions, but there are many others. A single *example* is the use of independent amplitude modulation channels on the sine and cosine waves of a carrier.

Although these other orthogonal function solutions offer possibilities in the art, they have not at the present time received as much design effort on actual embodiments as have the frequency division or the time division multiplexing.

### Auxiliary Signals

It was noted in Sect. 1, Nature of the Data, that when the data are presented in certain ways it is desirable to include some starting or other auxiliary information to mark out specific blocks of the information or to give other reference conditions. This auxiliary information needs to be distinguished in some way from the primary information. It comes regularly in the organization of the transmission, so that while the system is in normal operation the distinction need not be particularly conspicuous. However, for one cause or another the transmission may occasionally be interrupted. When this occurs, reestablishment is likely to be quicker, the more distinctive the auxiliary signals are.

*Multiplexing* the auxiliary signal with the principal signal may be done in a large variety of ways. A simple form is used in the standard teletypewriter. Here the distinction is secured by setting up a pattern of marking and spacing in the binary signal that is not duplicated in any



FIG. 10. Stop and start pulses in teletypewriter signal.

portion of any character. As indicated in Fig. 10, this pattern consists of a marking stop signal that in duration is equal to or greater than 1.4 signal elements, followed by a spacing start signal of one signal element duration. At *a* and *b* are shown stop signal elements of minimum duration,

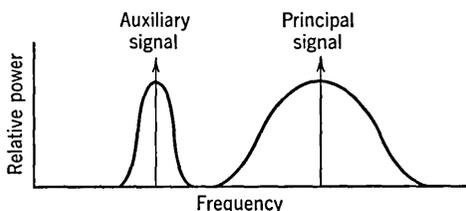


FIG. 11. Double sideband signal with auxiliary word start channel.

at *c* one of longer than minimum duration. There is, of course, some possibility that the excess over the minimum duration would bring the stop signal to 2 signal elements. This could then be duplicated in portions of characters, which could then be confused with the stop-start pattern.

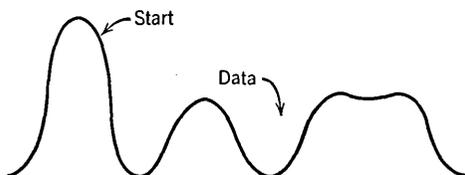


FIG. 12. Use of amplitude discrimination for word start channel.

This does occur on occasion, and the return to normal operation takes several characters.

At the other extreme, distinctiveness in the auxiliary signal is obtained by the use of an extra, narrow band carrier channel for it. One *example* of this is illustrated in Fig. 11. Another method is to use amplitude discrimination (Ref. 17). This is illustrated by the signal of Fig. 12.

With this arrangement the amplitude range permitted for the signal is reduced in comparison with the power capacity of the transmission medium, since the maximum capacity is used for the auxiliary signal. Thus the effective signal-to-noise ratio is less than it could be if the principal signal utilized the full capacity.

There are numerous other methods of introducing the auxiliary signals.

### 3. TRANSMISSION IMPAIRMENT

Electric circuits do not reproduce signals with complete fidelity. A basic element of the engineering of a system consists, therefore, in establishing tolerances on the permissible impairment of the signal consistent with acceptable performance.

The influence of limitation of the frequency bandwidth has already been discussed in Sect. 2.

#### Noise

No electric communication circuit is ever free from varying currents and voltages which are uncorrelated with the transmitted signal (except possibly in some statistical manner) and which tend to be confused with it at the receiver. These erratic waves have been perceived and studied in telephony. They have there been named *noise* because they end in audible noise in the receiver. The term has, however, been generally extended in the art to cover the effect in other types of communication.

Since noise is unpredictable in detail, it has in general to be dealt with in a statistical manner. Extensive study has been made of its statistical and other properties (Ref. 22). The discussion here is confined to a simple exposition of what one can expect in signal transmission media.

**Single Frequency Noise.** The noise wave may consist of a sustained single frequency. On the time scale this is a simple harmonic variation in voltage or current. On the frequency scale it is a single line spectrum.

**Single Impulse Noise.** On the other hand, the wave may consist of a sharp impulse at a given time. On the frequency scale it consists of a density of components which is uniform in amplitude out to some frequency beyond which it drops and approaches zero. This frequency depends upon the duration of the impulse. For an infinitesimal duration the frequency is infinity. The phases of the components are closely correlated.

**Cumulation to Gaussian Noise.** It is clear that the single frequency and the single impulse represent opposite extreme types of noise. In practice one can encounter a cumulation of a number of different single frequencies, each of different amplitude and phase. One can also encounter a cumulation of different single impulses, each of different amplitude and timing.

Each of these cumulations, as it becomes more extensive, and with sufficient randomness in its components, approaches "white" Gaussian noise (Ref. 23). *White Gaussian noise* may be defined in simple terms as random noise which has a Gaussian distribution of amplitude in the time domain and uniform distribution of power in the frequency domain. (To keep the total power finite, the uniformity need extend only somewhat beyond the frequency range under consideration for the signal.) The term "white" stems from its analogy to Rayleigh-Jeans radiation in optics (Ref. 24). This is somewhat of a misnomer, however, in that this radiation

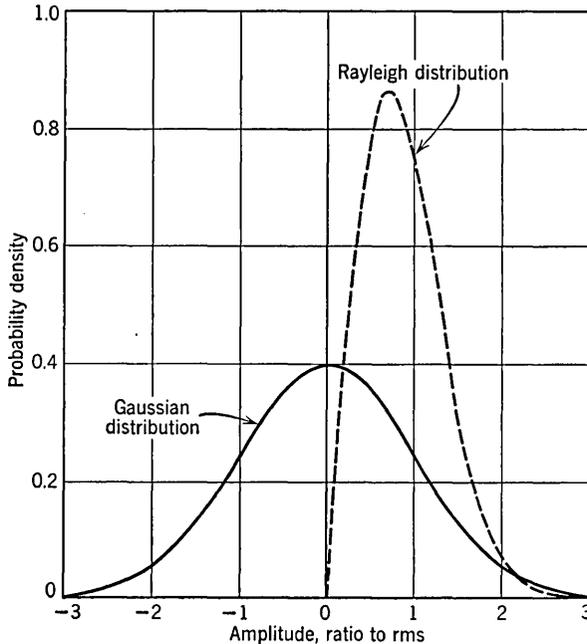


FIG. 13. Normalized Gaussian and Rayleigh noise distributions.

has a uniform distribution of power in the *frequency* spectrum. A white radiation used by colorimetrists has uniform power distribution in the *wavelength* spectrum. Thus Rayleigh-Jeans radiation, which has more power than this in the blues and less in the reds, is not white but really blue.

The *Gaussian distribution* in a normalized form is expressed by the equation

$$(1) \quad p \left( \frac{I}{\sqrt{k}} \right) \frac{dI}{\sqrt{k}} = \frac{dI}{\sqrt{2\pi k}} e^{-I^2/2k}.$$

This is the probability that the instantaneous amplitude lies between

$I/\sqrt{k}$  and  $(I + dI)/\sqrt{k}$ , where  $k$  is the normalizing constant, equal to the mean square value of  $I$ .

A plot of this normalized distribution is illustrated in the full line of Fig. 13. The normalization consists in referring to the amplitude  $I$  in terms of its ratio to the rms value  $\sqrt{k}$ .

Noise encountered in practice is rarely apt to be exactly any one of the three types which have been described. These are, however, much used as idealizations for mathematical and engineering purposes.

**Non-White and Non-Gaussian Noise.** On occasions where the noise actually encountered is sufficiently different from the idealization to influence a conclusion it is necessary to deal with the non-white and non-Gaussian noise as such.

The deviation of noise from the idealized white Gaussian may be characterized in a number of different ways. It may be characterized by size, as large or small; toward single frequency idealization or toward impulse idealization; by variation in spectral distribution or in distribution of amplitudes as a function of time; or other characteristics.

One frequently used variant is noise obtained from Gaussian noise by *envelope rectification*. This follows a Rayleigh distribution of amplitudes (Ref. 23). The Rayleigh distribution in normalized form is expressed by the equation

$$(2) \quad p\left(\frac{i}{\sqrt{k}}\right) \frac{dI}{\sqrt{k}} = \frac{2I}{k} e^{-I^2/k}, \quad I > 0$$

$$= 0, \quad I < 0,$$

where the symbols have the same meaning as in eq. (1).

A plot of this distribution, in the normalized form, is illustrated in the dotted lines of Fig. 13.

A second frequently used variant is *filtered white Gaussian noise*. This modifies the power spectrum of the noise. A special case occurs when the filter has a passband which becomes narrow compared to the spectrum of the signal which is being disturbed. In this case the noise approaches single frequency noise.

A third variant is the *impulsive noise* encountered in communications circuits. This is cumulated from single impulses of the type which have been considered. However, the number cumulated is small enough and not sufficiently random in amplitude or time of occurrence, so that the distribution of amplitudes (including zero amplitude, which is important) is not Gaussian. This can occur for *example* in telephone circuits exposed to some forms of dial-switching equipment or to static. It is necessary

for close engineering in such cases to determine the exact distribution of amplitudes, and sometimes of timing instants.

**Influence of Noise on Error.** The effect of noise, of course, is that it changes the received wave and tends to cause the signal for one set of data to be confused with that for another.

CASE 1. *Analog Data.* The error may vary continuously from zero to large amounts. One simple method of expressing the error is in terms of its rms value. There are other methods (see Chap. 17), but they usually lead to more complicated techniques of engineering. The optimum noise performance may be obtained in a given system when both signal and noise are filtered through the optimum filter (see Chap. 17).

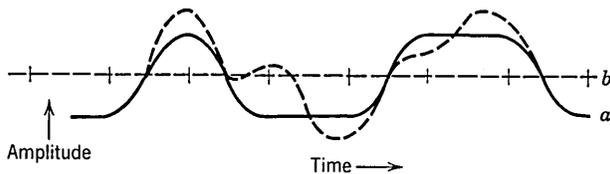


FIG. 14. Effect of noise in causing errors in data pulse signals.

CASE 2. *Digital Data.* The received signal wave may vary over a range without any misinterpretation resulting (Ref. 25). When the departure exceeds this range, however, a marking signal may be misinterpreted as a space, or a space as a mark.

A simple *illustration* of this is shown in the baseband signal of Fig. 14. At *a* is the assumed signal, consisting of a space, a mark, two spaces, two marks, and a space. The interpretation of marking or spacing is assumed to be made according to whether the wave amplitude falls above or below the critical value *b*, at sampling instants which are designated on the line *b*. The effect of a few noise pulses are indicated by dotted lines on *a*. In the simple baseband system shown, the critical level *b* is at half the marking level, or 6 db below marking. Thus *the critical signal-to-noise ratio, in terms of marking level to noise peaks, is 6 db*. For a higher signal-to-noise peak ratio, no errors in transmission are caused by the noise. For a lower ratio, errors appear.

Because of the erratic nature of noise, this is not always a convenient specification. If the noise has, say, a Gaussian distribution, no matter what the critical level *b* is, there is some finite probability that it will be exceeded (with the appropriate polarity) and cause an error. The engineering of the system then consists in first setting an acceptable error performance (see Sect. 1, Error Standards). This sets the acceptable probability for the existence of noise pulses of a given polarity 6 db below marking

level. In Fig. 13 one can tell, for Gaussian noise, how far above rms a level must be to occur with any given probability. This indicates how far below the critical level  $b$  the rms of the noise must be set. If that ratio is translated into db, then *by adding 6 db one obtains the figure which must be specified for the signal- (marking level) to-noise (rms) ratio for the system, in order to meet the desired error performance.*

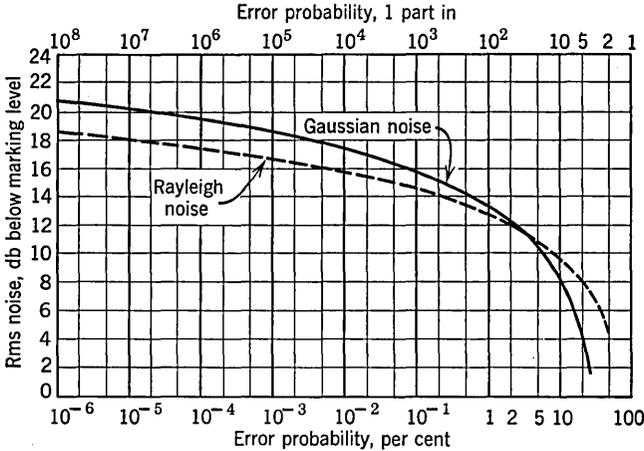


FIG. 15. Error probabilities from Gaussian and Rayleigh noise distributions.

For convenience this ratio in decibels has been plotted as the solid line in Fig. 15. A dotted line is shown for the case of noise having a Rayleigh distribution.

If the noise is of the *impulsive type* it is usually impractical to specify it in terms of its rms value, as a ratio to the signal. In such cases it is apt to be expeditious to make measurements on the noise itself, to determine the amplitude of its peaks at the error frequency which has been set. Then the marking amplitude can be set 6 db above this.

The *noise margin* of 6 db which has been discussed here is a basic figure. If the signal has three levels to be recognized, as in Fig. 12, the figure has to be increased. It will also be found (see below) that other margins need to be added to allow for other impairing effects on the signal.

**Echoes and Equalization**

**Echoes and Transfer Functions.** In general, no transmission medium reproduces a sent signal wave shape exactly in all its characteristics. The departures from exact faithfulness can be considered from two points of view, sometimes one being more convenient and sometimes the other.

According to the first point of view the departures may be considered as a succession of more or less delayed *echoes* of the original wave. Some of the echoes are of the same polarity as the original wave, and some of the opposite polarity.

According to the second point of view the signal and its transmitted reproduction may be analyzed into their *Fourier transforms*. Each Fourier component of the reproduction is obtainable from that of the original by multiplication by an amplitude response factor and displacement by a phase shift. The factor and phase shift vary from frequency to frequency over the spectrum of the signals. Since the Fourier transform is unique, the complete description in a linear system of a given case according to the one viewpoint can also be matched exactly by a complete description according to the other viewpoint. Whichever one is used is then simply a matter of engineering convenience.

Practical experience indicates that the echo treatment leads quickly to equalizer designs. This is because it forms an immediate bridge to functions of frequency, and equalizer designs are simple in such terms. The design of filters and equalizers from transfer functions of time is usually far more cumbersome.

**Equalization.** In practical transmission systems these distortions are usually reduced by what is called equalization. A network is placed in tandem with the system which again multiplies all the Fourier components, each by an amplitude response factor, and displaces each by a phase shift. The response factor of the network is designed to vary in the inverse way from that of the system, so that the products of the two are as nearly as possible constant over the frequency spectrum. For this reason the network is called an *equalizer* and the process called equalization. The phase shifts are designed to add together to a total phase shift which is proportional to frequency. Since the perceptive mechanism of the ear is not very responsive to phase shifts, the equalization of telephone circuits has generally concerned itself almost exclusively with an equalization of the amplitude response factor. This unconcern with the phase correction is occasionally of importance in the use of telephone facilities for data transmission. It sometimes requires the insertion of phase correcting networks to supplement the amplitude correction already existing for the telephone use.

It is not usually possible or economically feasible in practice to correct a system exactly. The considerations which are given below can apply equally well to a residual departure, left after such correction has been applied as is practical, or to an uncorrected facility.

**Impairment of Noise Margin.** It is clear that an echo partakes of one property in common with noise, i.e., it changes the received wave and

tends to cause the signal for one set of data to be confused with that for another. Where a given deviation has previously been set as acceptable, the echo uses up some of this possibility for deviation and leaves less of it as an allowance for the noise. In this sense, therefore, it impairs the noise margin of the system. Where, as was noted above, a margin of 6 db was necessary for the marking signal level over individual noise peaks to just avoid potential error, a greater margin is needed in the presence of echo.

The amount of this *excess margin* which can be allotted to echo, in any given case, is a matter of engineering judgment. It depends upon the relative costs of reducing the echo and the alternative of reducing the noise. One may say that, in general, an increase of 1 db in margin is small, and that as severe a limitation as this on the echo is economical where the

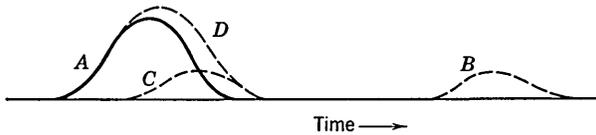


FIG. 16. Close-in and remote echoes in signal.

echo is fairly easy to reduce. At the other extreme one may say that an increase of 10 db in margin is fairly large. It is likely to be economical only where it is quite difficult to reduce the echo, or where the noise expected is very low. In the engineering of data systems it is convenient to consider steps respectively of 1, 3, 6 and 10 db in the noise margin impairment.

The amplitude of echo that can cause a given impairment depends upon how much it is delayed with respect to the original signal. As an illustration the signal *A* in Fig. 16 may be followed by a comparatively long delayed echo at *B*. Here the impairment depends upon the echo amplitude, and it does not vary much with small changes in the echo delay.

The signal may also be followed by a closely spaced echo as at *C*. The sum of signal and echo is shown at *D*. Here the major effect of the echo is to change somewhat the wave shape of the signal, but mostly it changes signal amplitude. A substantial part of the effect of the echo consists merely in changing the effective loss of the transmission facility somewhat. This part of the effect can be compensated for by an adjustment of receiving gain. The impairment from a closely following echo of a given amplitude is less than from a long-delayed echo of the same amplitude. Also the impairment can be expected to vary rather rapidly with delay, for the short delays.

**Relationship between Echoes and Equalization.** This relationship, suggested above, may be examined quantitatively (Refs. 7 and 26).

Consider a single Fourier component of the signal voltage, of frequency  $\omega/2\pi$

$$(3) \quad v = \cos \omega t.$$

When this is transmitted over a system that generates an echo of relative amplitude  $a$  and relative delay  $\tau$ , it becomes

$$(4) \quad \begin{aligned} v &= \cos \omega t + a \cos \omega(t - \tau) \\ &= \cos \omega t + a \cos \omega t \cos \omega \tau + a \sin \omega t \sin \omega \tau, \end{aligned}$$

$$(5) \quad v = (1 + a \cos \omega \tau) \cos \omega t + a \sin \omega t \sin \omega \tau.$$

If the *overall transmission* is designated as

$$(6) \quad v = r(\omega) \cos [\omega t - \phi(\omega)],$$

then  $r(\omega)$  represents the *overall amplitude ratio* and  $\phi(\omega)$  the *overall phase shift*. These quantities may be computed from eq. (5), namely as:

$$(7) \quad r(\omega) = [(1 + a \cos \omega \tau)^2 + (a \sin \omega \tau)^2]^{1/2},$$

$$(8) \quad \approx 1 + a \cos \omega \tau, \text{ when } a \text{ is small};$$

$$(9) \quad \phi(\omega) = \tan^{-1} \frac{a \sin \omega \tau}{1 + a \cos \omega \tau},$$

$$(10) \quad \approx a \sin \omega \tau, \text{ when } a \text{ is small.}$$

Thus a given echo in a description in terms of time corresponds approximately to a *ripple* in the amplitude response and in the phase, in a description in terms of frequency. The *amplitude of the ripple, in nepers or radians, is equal to the relative amplitude of the echo*. The wavelength of the ripple along the frequency scale ( $\lambda = \Delta f = \Delta\omega/2\pi$ ) is equal to the reciprocal of the delay  $\tau$ .

$$(11) \quad \begin{aligned} \Delta\omega\tau &= 2\pi, & \Delta f\tau &= 1. \\ \lambda &= \Delta f = 1/\tau. \end{aligned}$$

**Tolerances.** A given *amplitude allowance* for noise and echo together is taken as unity, and the amount allocated to noise alone is allocated at 1 to 10 db below this. This noise allocation bears a ratio of  $r$  ( $<1$ ) to the total. The remainder, or  $1 - r$ , is allocated to the echo. Figures are given in the first two columns of Table 1 to indicate the numerical values.

TABLE 1. ECHO AND RIPPLE TOLERANCES

Impairment, db	$r =$ Amplitude Ratio	Echo Tolerance		Ripple Tolerance	
		$(1 - r)/2$	db	$(1 - r)/2$ , db	degrees
1	0.89	0.054	25	$\pm 0.47$	$\pm 3.1$
3	0.71	0.146	17	$\pm 1.3$	$\pm 8.4$
6	0.50	0.25	12	$\pm 2.2$	$\pm 14$
10	0.32	0.34	9	$\pm 3.0$	$\pm 20$

The noise peak tolerance in the simple binary code transmission of Fig. 14 was one-half of the marking amplitude. Thus  $(1 - r)/2$  measures the ratio of amplitude allocated to the echo to marking amplitude. This is

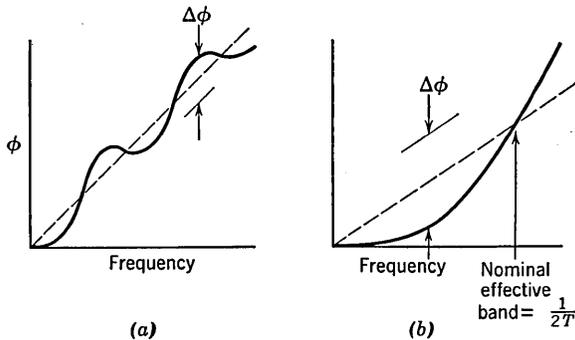


FIG. 17. Phase shift characteristics: (a) remote echo, (b) close-in echo.

listed in the third column as an echo tolerance. It is converted into decibels in the fourth column.

The tolerance may be placed instead on the ripple amplitude in the amplitude response characteristic. The numerical figure of the third column expresses this ripple excursion, in each direction, in nepers. It is converted, in the fifth column, into decibels.

The tolerance may also be placed on the *phase shift ripple*. For this purpose the quantity of the third column is assumed as measured in radians. For convenience it is converted into degrees in the sixth column.

So far, nothing has been said concerning the *absolute propagation time* of the system. When this is taken into consideration, it is found that the phase ripple really occurs about a diagonal straight line through the origin, as illustrated in Fig. 17a.

Where the echo delay is very short, and only a small portion of a ripple cycle appears in the utilized frequency band, the straight line about which

the phase deviations are to be taken is not so easily identified. A more or less arbitrary, but practical construction is given in Fig. 17*b*. Here the straight line is drawn to intersect the actual phase at the frequency which marks the top of the nominal effective band. This is the reciprocal of twice the signal element duration. The phase departure is taken as the maximum double excursion from this straight line, as indicated by  $\Delta\phi$  in the figure.

Note that in Fig. 17*a* the excursion  $\Delta\phi$  measures double the ripple amplitude and consequently double the echo amplitude. In Fig. 17*b* the double ripple amplitude is not accessible within the scope of the plot, and is larger than  $\Delta\phi$ . Figure 17*a* represents a remote echo, such as at *B* in Fig. 16, and Fig. 17*b* represents a close-up echo, such as at *C* in Fig. 16.

It is then clear that for a given excursion  $\Delta\phi$  the echo amplitude in Fig. 17*b* (close-up echo) is larger than for Fig. 17*a* (remote echo).

This variation of the actual echo amplitude for a given  $\Delta\phi$  corresponds approximately to the variation in permissible echo amplitude for a given impairment suggested in Fig. 16. Thus (Ref. 7) the specification of the phase departure, for an allotted impairment, is roughly independent of whether the departure occurs as a single long bend such as in Fig. 17*b*, or as a fine structure ripple such as in Fig. 17*a*.

The tolerances which are listed in the last two columns of Table 1 were set for a binary digital transmission. Tolerances for a continuous analog transmission have in practice been found to be of the same order of magnitude.

However, tolerances for a discrete analog system, with time division multiplexed channels interlaced, need to be much more severe (Ref. 27).

**Envelope Delay Distortion.** In practice it is often convenient to measure the phase shift characteristic of a transmission system in terms of its *envelope delay*. This represents the time of transmission of the envelope of a carrier, as the carrier frequency is varied through the spectrum. It is measured as (Ref. 28)

$$(12) \quad D = d\phi/d\omega,$$

where  $D$  = envelope delay, seconds,

$\phi$  = phase shift, radians,

$\omega$  = radian frequency, radians per second.

When this differentiation is applied to the simplified eq. (10) the result is

$$(13) \quad D = (d/d\omega) a \sin \omega\tau = a\tau \cos \omega\tau.$$

The double excursions of the ripples in eq. (13) are

$$(14) \quad \Delta D = 2a\tau.$$

It is found from eq. (13) that the application of a fixed tolerance on the envelope delay ripple irrespective of the wavelength of the ripple (or corresponding echo delay  $\tau$ ) leads to an exaggeratedly severe limitation on echo amplitude  $a$  for large values of  $\tau$ .

In other words *the use of the envelope delay for the purpose of specifying limits on phase distortion for data transmission tends to be unduly severe on fine structure excursions in the characteristic.* Thus when the envelope delay criterion is used, it is necessary to be aware of this and appropriately ignore the finer structure ripples in the characteristic.

In a general way it is found (Ref. 7) that a delay distortion of  $\pm 0.4$  signal element duration gives a noise impairment, under unfavorable conditions, of about 3 db. If one takes distortions as roughly proportional to the permissible echoes the tolerance figures are given in Table 2.

TABLE 2. ENVELOPE DELAY TOLERANCES

Impairment, db	Tolerance, Signal Element
1	$\pm 0.15$
3	$\pm 0.4$
6	$\pm 0.7$
10	$\pm 0.9$

**Quadrature Component in Vestigial Sideband Transmission.** An interfering component similar in certain respects to an echo is generated by the usual form of vestigial sideband transmission (Sect. 2, Amplitude Modulation). This is the *quadrature component*, so called because this interference adds in quadrature to the otherwise undistorted signal (Ref. 29).

Although the precise wave shape of this interfering component is different from that of an echo, it does use up signal amplitude range in much the same manner and requires an increase in signal to noise margin.

It has been shown (Refs. 29, 30) that the amplitude of this interfering component varies according to how much frequency space is allowed to the vestigial sideband, and, to some degree, to the particular shape of the cutoff. The wider this frequency space, the smaller will be the amplitude of the quadrature component. In actual data transmission practice the vestigial bandwidth used, as measured from the carrier to the frequency at which the response drops to a very low value, tends to run from some one-half to one-fourth of the nominal effective band.

The amplitude of the quadrature component can also be changed by changing the depth of modulation of the signal. The depth of modulation is reduced by sending a finite amplitude (instead of the more usual zero amplitude) of carrier during a spacing signal. This reduces the quadrature component, and to that extent it reduces the impairment which it causes in the signal to noise ratio. It does, of course, also reduce the amplitude range of the signal between marking and spacing, and to that extent also impairs the signal-to-noise ratio. As the spacing carrier rises, this impair-

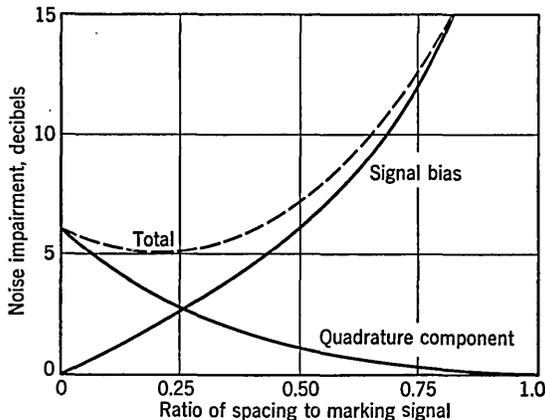


FIG. 18. Noise impairment caused by quadrature component in vestigial sideband transmission.

ment also rises, and the quadrature component impairment drops. At some value there is a minimum total impairment. A typical case has been worked out by Sunde (Ref. 30) and is illustrated in Fig. 18.

### Level Changes

Transmission systems in general show some variation with time in overall net loss (or gain). This comes from a variety of causes, such as changes in temperature (and therefore resistance) of conductors, variation in battery supply, and aging or replacement of vacuum tubes.

**Analog System.** An amplitude modulated analog system is especially vulnerable to received level change. A system engineered to a possible  $\pm 5$  per cent error does not represent a very high performance. Yet if all the error is assigned to level change, this is required to be within less than  $\pm 1/2$  db. This is a severe requirement for anything but a comparatively short direct wire circuit.

Because of this, an analog system usually uses a pilot channel of some type to transmit a reference amplitude. Also, frequency modulation is often preferred to amplitude modulation. Even in this case, however, in some carrier facilities, as was mentioned above in Sect. 2, there is a change in frequency which is analogous to a level change. Thus a pilot channel may be needed to send a reference frequency.

**Digital System.** In a binary digital system a level change cuts into the signal-to-noise margin somewhat in the same way as does an echo. If no change is assumed in the critical level distinguishing a mark from a space, then for the four grades of impairment considered before, the allowances are given in Table 3.

TABLE 3. LEVEL CHANGE TOLERANCES

Impairment, db	$r =$ Amplitude Ratio	$(r + 1)/2 =$ Amplitude Tolerance	Tolerance, db
1	0.89	0.95	0.5
3	0.71	0.86	1.4
6	0.50	0.75	2.5
10	0.32	0.66	3.6

These are still fairly severe requirements, and usually some compensating device is provided in the system. This can be an automatic adjustment of the critical level at a given fraction of marking level, or an automatic volume adjuster for the marking level, or possibly even both.

If a three-level signal is used, as in Fig. 12, the tolerances are correspondingly more severe.

## REFERENCES

1. A. B. Clark, Telephone transmission over long cable circuits, *Bell System Tech. J.*, **2**, 67-94 (1923).  
 J. T. O'Leary, E. C. Blessing, and J. W. Beyer, An improved 3-channel carrier telephone system, *Bell System Tech. J.*, **18**, 49-75 (1939).  
 H. J. Fisher, M. L. Almquist, and R. H. Mills, A new single channel carrier telephone system, *Bell System Tech. J.*, **17**, 162-183 (1938).  
 C. W. Green and E. I. Green, A carrier telephone system for toll cables, *Bell System Tech. J.*, **17**, 80-105 (1938).  
 R. S. Caruthers, The Type N-1 carrier telephone system: Objectives and transmission features, *Bell System Tech. J.*, **30**, 1-32 (1951).
2. F. A. Cowan, R. G. McCurdy, and I. E. Lattimer, Engineering requirements for program transmission circuits, *Bell System Tech. J.*, **20**, 235-249 (1941).  
 R. A. Leconte, D. B. Penick, C. W. Schramm, and A. J. Wier, A carrier system for 8000-cycle program transmission, *Bell System Tech. J.*, **28**, 165-180 (1949).

3. F. A. Cowan, Networks for theater television, *J. Soc. Motion Picture & Television Engrs.*, **62**, 306-313 (1954).

S. Doba and A. R. Kolding, A new local video transmission system, *Bell System Tech. J.*, **34**, 677-712 (1955).

C. H. Elmendorf, R. D. Ehrbar, R. H. Klie, and A. J. Grossman, L-3 Coaxial system design, *Bell System Tech. J.*, **32**, 781-832 (1953).

4. R. E. Crane, J. T. Dixon, and G. H. Huber, Frequency division techniques for a coaxial cable network, *Trans. Am. Inst. Elec. Engrs.*, **66**, 1451-1459 (1947).

K. E. Appert, R. S. Caruthers and W. S. Chaskin, Application and transmission features of a new 12-channel open-wire carrier system, *Trans. Am. Inst. Elec. Engrs.*, **73**, Pt. I, 18-27 (1954).

5. *Radio Spectrum Conservation*, Report of the Joint Technical Advisory Committee, McGraw-Hill, New York 1952.

6. A. A. Roetken, K. D. Smith, and R. W. Friis, The TD-2 microwave radio relay system, *Bell System Tech. J.*, **30**, 1041-1077 (Pt. II) (1951).

7. P. Mertz, Transmission line characteristics and effects on pulse transmission, *Proceedings of the Symposium on Information Networks*, April 12-14, 1954, Vol. III, pp. 85-114, Polytechnic Institute of Brooklyn, New York.

8. W. M. Goodall, Television by pulse code modulation, *Bell System Tech. J.*, **30**, 33-49 (1951).

9. B. Lippel, A systematic survey of codes and coders, *I.R.E. Convention Record*, Pt. 8, Information Theory, pp. 109-119, 1953.

10. A. E. Laemmel, Design of digital coding networks, *Proceedings of the Symposium on Information Networks*, April 12-14, 1954, Vol. III, pp. 309-320, Polytechnic Institute of Brooklyn, New York.

A. Feinstein, A new basic theorem of information theory, *Trans. I.R.E.*, Professional Group on Information Theory, **PGIT-4**, pp. 2-22, Sept. 1954.

P. Elias, Predictive coding, *I.R.E. Trans. on Information Theory*, **IT-1**, No. 1, pp. 16-33, March 1955.

D. Slepian, A class of binary signaling alphabets, *Bell System Tech. J.*, **35**, 203-234 (1956).

C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, Ill., 1949.

R. M. Fano, The transmission of information, *Mass. Inst. Technol., Research Lab. Electronics, Tech. Rept.*, No. 65, 1949.

11. B. M. Oliver, Efficient coding, *Bell System Tech. J.*, **31**, pp. 724-750 (1952).

N. M. Blackman, Minimum-cost encoding of information, *Trans. I.R.E.*, Professional Group on Information Theory, **PGIT-3**, pp. 139-149, March 1954.

12. R. W. Sears, Electron beam deflecting tube for pulse code modulation, *Bell System Tech. J.*, **27**, 44-57 (1948).

13. L. B. Wadel, Analysis of combined sampled and continuous-data systems on an electric analog computer, *I.R.E. Convention Record*, Pt. 4, pp. 3-7, 1955.

G. Franklin, Linear filtering of sampled data, *I.R.E. Convention Record*, Pt. 4, pp. 119-128, 1955.

S. P. Lloyd and B. McMillan, Linear least squares filtering and prediction of sampled signals, *Proceedings of the Symposium on Network Theory*, April 13-15, 1955, Vol. V, pp. 221-247, Polytechnic Institute of Brooklyn, New York.

R. M. Stewart, Statistical design and evaluation of filters for the restoration of sampled data, *Proc. I.R.E.*, **44**, 253-257 (1956).

14. R. W. Hamming, Error detecting and error correcting codes, *Bell System Tech. J.*, **29**, 147-160 (1950).

M. J. E. Golay, Binary coding, *Trans. I.R.E.*, Professional Group on Information Theory, **PGIT-4**, pp. 23–28, Sept. 1954.

P. Elias, Error-free coding, *Trans. I.R.E.*, Professional Group on Information Theory, **PGIT-4**, pp. 29–37, Sept. 1954.

I. S. Reed, A class of multiple-error-correcting codes and the decoding scheme, *Trans. I.R.E.*, Professional Group on Information Theory, **PGIT-4**, pp. 38–49, Sept. 1954.

R. A. Silverman and M. Balser, Coding for constant-data-rate systems, *Trans. I.R.E.*, Professional Group on Information Theory, **PGIT-4**, pp. 50–63, Sept. 1954.

15. H. Nyquist, Certain topics in telegraph transmission theory, *Trans. Am. Inst. Elec. Engrs.*, **47**, 617–644 (1928).

16. A. W. Horton and H. E. Vaughan, Transmission of digital information over telephone circuits, *Bell System Tech. J.*, **34**, 511–528 (1955).

17. J. V. Harrington, P. Rosen, and D. A. Spaeth, Some results on the transmission of pulses over telephone lines, *Proceedings of the Symposium on Information Networks*, April 12–14, 1954, Vol. III, pp. 115–130, Polytechnic Institute of Brooklyn, New York.

18. D. Middleton, On the theoretical signal to noise ratios in FM receivers: A comparison with amplitude modulation, *J. Appl. Phys.*, **20**, 334–351 (1949).

19. H. S. Black, *Modulation Theory*, Van Nostrand, Princeton, N. J., 1953.

20. W. M. Goodall, Telephony by pulse code modulation, *Bell System Tech. J.*, **26**, 395–409 (1947).

21. N. Marchand, Analysis of multiplexing and signal detection by function theory, *I.R.E. Convention Record*, Pt. 8, pp. 48–53, March 1953.

22. P. L. Chessin, A bibliography on noise, *I.R.E. Trans. on Information Theory*, **IT-1**, No. 2, pp. 15–31, Sept. 1955.

J. R. Pierce, Physical sources of noise, *Proc. I.R.E.*, **44**, 601–608 (1956).

W. R. Bennett, Methods of solving noise problems, *Proc. I.R.E.*, **44**, 609–637 (1956).

23. S. O. Rice, Mathematical analysis of random noise, *Bell System Tech. J.*, **23**, 282–332 (1944); **24**, 46–156 (1945).

24. F. K. Richtmyer and E. H. Kennard, *Introduction to Modern Physics*, p. 189, McGraw-Hill, New York, 1942.

25. B. M. Oliver, J. R. Pierce, and C. E. Shannon, The philosophy of PCM, *Proc. I.R.E.*, **36**, 1324–1331 (1948).

26. P. Mertz, Influence of echoes on television transmission, *J. Soc. Motion Picture & Television Engrs.*, **60**, 572–596 (1953).

H. A. Wheeler, The interpretation of amplitude and phase distortion in terms of paired echoes, *Proc. I.R.E.*, **27**, 359–385 (1939).

27. W. R. Bennett, Time division multiplex systems, *Bell System Tech. J.*, **20**, 199–221 (1941).

28. H. Nyquist and S. Brand, Measurement of phase distortion, *Bell System Tech. J.*, **9**, 522–549 (1930).

29. H. Nyquist and K. W. Pfleger, Effect of the quadrature component in single sideband transmission, *Bell System Tech. J.*, **19**, 63–73 (1940).

30. E. D. Sunde, Theoretical fundamentals of pulse transmission, *Bell System Tech. J.*, **33**, 721–788, 987–1010 (1954).

31. P. M. Woodward, Theory of radar information, *Trans. I.R.E.*, Professional Group on Information Theory, **PGIT-1**, pp. 108–113, Feb. 1953.

32. Guide to application and treatment of channels for power-line carrier, *Trans. Am. Inst. Elec. Engrs.*, Pt. III-A, **73**, 417–436 (1954).

# FEEDBACK CONTROL

## E. FEEDBACK CONTROL

*W. M. Gaines, Editor*

19. *Methodology of Feedback Control, by W. M. Gaines*
20. *Fundamentals of System Analysis, by S. J. Jennings and  
A. A. Winkeljohann*
21. *Stability, by W. E. Sollecito and S. G. Reque*
22. *Relation between Transient and Frequency Response, by C. E. Bradford  
and M. W. DeMerit*
23. *Feedback System Compensation, by P. G. Cushman*
24. *Noise, Random Inputs, and Extraneous Signals, by D. L. Lippitt*
25. *Nonlinear Systems, by W. M. Gaines*
26. *Sampled-Data Systems and Periodic Controllers, by J. E. Barnes, Jr.*



## Methodology of Feedback Control

W. M. Gaines

1. Symbols for Feedback Control	19-01
2. General Feedback Control Definitions	19-04
3. Feedback Control System Design Considerations	19-12
4. Selection of Method of Synthesis for Feedback Controls	19-19
References	19-21

### 1. SYMBOLS FOR FEEDBACK CONTROL

#### Alphabetical List by Letter Symbols

Terminology given in Table 1 is for feedback control covered in the following chapters. In the case of specific physical examples, the terminology of the particular field from which the example is taken will be used; for example, in an electrical example,  $e$  may be used for voltage and  $i$  for current. The last column of the table lists the chapter where the symbol is first used. This reference may be useful to the reader for looking up discussions of the various quantities.

The nomenclature used is patterned after the standard nomenclature and symbols of the American Standards Association (Ref. 1). Capital letters will be used to represent the Laplace transforms of the time functions; for example,  $A(s)$  is the Laplace transform of  $a(t)$ . An asterisk (\*) indicates that the quantity is in sampled form; for example,  $e^*(t)$  is the sampled form of the signal  $e(t)$ .

TABLE 1. LETTER SYMBOLS FOR FEEDBACK CONTROL

Symbols	Use or Term	First Used
$a$	Arbitrary constant and/or coefficient for differential equation	
$A$	Arbitrary constant for time response equation	
$a(t)$	Impulse response of reference input terms (function of time)	Chap. 20
$b, B$	Arbitrary constant for time response equation	
$B$	Magnitude of deadband	Chap. 25
$b(t)$	Primary feedback variable (function of time)	Chap. 20
$c(t)$	Controlled variable (function of time)	Chap. 20
$c^*(t)$	Sampled form of $c(t)$	Chap. 26
$d, D$	Arbitrary constants for time response equation	
$D$	Magnitude of negative deficiency; a denominator term; used also as a subscript	Chap. 25
$D(s)$	Polynomial in $s$ usually the denominator	Chap. 22
$e(t)$	Actuating signal (function of time)	Chap. 20
$e^*(t)$	Sampled form of $e(t)$	Chap. 26
$f$	Frequency, cycles per second (see definition of $\omega$ )	
$f(t)$	Arbitrary variable (function of time)	
$g(t)$	Impulse response of forward element (function of time)	Chap. 20
$G_D( x , \omega)$ or $G_D$	Describing function	Chap. 25
$h(t)$	Impulse response of feedback elements (function of time)	Chap. 20
$H$	Magnitude of hysteresis	Chap. 25
$i$	$i$ th term in a series, used as subscript	
$i(t)$	Ideal value of the ultimately controlled variable (function of time)	Chap. 20
$j$	Complex number, $\sqrt{-1}$	
$k$	$k$ th term in a series, used as a subscript	
$K$	Gain constant for system	
$K_0, K_1, K_2,$ etc.	Dynamic error coefficients, subscript indicates associated derivative	Chap. 20
$K_p$	Static position error coefficient	Chap. 20
$K_v$	Static velocity error coefficient	Chap. 20
$K_a$	Static acceleration error coefficient	Chap. 20
$\mathcal{L}$	Denotes application of the Laplace transform integral	
$\mathcal{L}^{-1}$	Inverse Laplace transform	
$m$	Used as a subscript to denote $m$ th term in series	
$m(t)$	Manipulated variable (function of time)	Chap. 20
$M$	Magnitude of $\frac{C}{R}(j\omega)$ , i.e., $\left  \frac{C}{R}(j\omega) \right $	Chap. 21
$M_m$	Maximum value of $\left  \frac{C}{R}(j\omega) \right $	Chap. 21
$n$	Use as a subscript to denote $n$ th term	
$n(t)$	Output of nonlinear element	
$N$	Particular value of $n$ ; a numerator term; a subscript	
$N(s)$	Polynomial in $s$ usually the numerator	Chap. 22

TABLE 1. LETTER SYMBOLS FOR FEEDBACK CONTROL (*Continued*)

Symbols	Use or Term	First Used
$p, p^2$	Differential operator, $p = d/dt, p^2 = d^2/dt^2$	Chap. 20
$p_n$	$n$ th pole	Chap. 22
$p(x)$	Probability distribution of $x$	Chap. 24
$P$	Number of poles in right half of $s$ -plane	Chap. 21
$P(x = n)$	Probability function	Chap. 24
$q(t)$	Indirectly controlled variable (function of time)	Chap. 20
$r(t)$	Reference input variable (function of time)	Chap. 20
$R$	Number of counterclockwise rotations of a vector from $-1 + j0$ to $H(j\omega)g(j\omega)$ locus as $\omega$ varies from 0 to $j\infty$ to $-j\infty$ to $-0$	Chap. 21
$S$	Magnitude or level of saturation	Chap. 25
$s$	Laplace transform operator = $\sigma + j\omega$	
$s_a, s_b, s_c$	Roots of numerator of $G(s)$ , zeros; $z_n$ also used in this case	Chap. 22
$s_1, s_2, s_3$	Roots of denominator of $G(s)$ , poles; $p_n$ also used in this case	Chap. 22
$t$	Time, seconds	
$t_r$	Rise time, seconds	Chap. 22
$t_d$	Delay time, seconds	Chap. 22
$t_p$	Time to first peak or overshoot of transient, seconds	Chap. 22
$t_s$	Settling time, seconds	Chap. 22
$T$	Time constant, seconds	Chap. 20
$u(t)$	Disturbance variable (function of time); step function	Chap. 20
$v(t)$	Desired value or command variable (function of time)	Chap. 20
$w(t)$	Impulse response of given element	
$x(t)$ }	Variables used when standard	
$y(t)$ }	terminology for feedback systems is not applicable	
$y_e(t)$	System error (function of time)	Chap. 20
$y_d(t)$	System deviation (function of time)	Chap. 20
$z(t)$	Indirectly controlled system impulse response	Chap. 20
$z$	$z$ transform operator	Chap. 26
$z_n$	$n$ th zero	Chap. 22
$Z$	Number of zeros in right half of $s$ -plane	Chap. 21
$\alpha$	Phase angle of closed loop frequency response	Chap. 21
$\gamma$	Phase margin	Chap. 21
$\delta$	Increment; Dirac function, impulse function	Chap. 20
$\Delta$	Incremental change in variable, usually used $\Delta x, \Delta y$ , etc.	
$\underline{\Delta}$	Denotes equality by definition	
$\zeta$	Relative damping, damping factor	
$\theta$	Phase of angle of open loop frequency response	Chap. 21
$\lambda$ }	Frequency, damping, used when $\sigma$ and $\omega$ are not applicable	
$\xi$ }		
$\pi$	3.14159	
$\prod_n(x)$	Product sign meaning $x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot \dots \cdot x_n$	Chap. 20
$\sigma$	Standard deviation (probability); decrement factor	Chap. 24 Chap. 20

TABLE 1. LETTER SYMBOLS FOR FEEDBACK CONTROL (*Continued*)

Symbols	Use or Term	First Used
$\sum_n (x)$	Summation: $x_1 + x_2 + x_3 + \cdots + x_n$	
$\tau$	Normalized time	
$\phi(\tau)$	Autocorrelation function	Chap. 24
$\Phi(\omega)$	Power spectral density	Chap. 24
$\psi$	Arbitrary angle	
$\omega$	Angular frequency or natural angular frequency, radians/second	
$\omega_b$	Bandpass frequency	
$\omega_c$	Crossover angular frequency, frequency at which $ G(s)H(s)  = 1$	Chap. 22
$\omega_m$	Angular frequency at which $M_m$ occurs	
$\omega_n$	$n$ th frequency term	
$\omega_0$	Undamped natural angular frequency	

## 2. GENERAL FEEDBACK CONTROL SYSTEM DEFINITIONS

### Basic Feedback Control System Elements

The basic elements of a typical feedback control system and the symbols and nomenclature used to describe them are given in Fig. 1 and defined in Table 2. The majority of the feedback control systems can be characterized by these elements although not all the basic elements will exist in every system. More complex systems can normally be represented by simply expanding the basic elements given in Fig. 1. Basic to any feedback control system are the forward elements, ( $g$ ), and feedback elements, ( $h$ ). Both the feedback and forward elements as well as the other basic elements may consist of subsystems that are feedback control systems themselves.

In cases where the standard terminology does not fit the control system under consideration, the individual authors have used the terminology they consider best suited. In such cases, the nonstandard terms and notation are defined where they are used.

The symbols and nomenclature of Fig. 1 are based upon Ref. 1. The ( $t$ ) as in  $a(t)$  can be dropped when convenient if the simpler symbol will not be misinterpreted. The usage of the symbols appearing in Fig. 1 and in Table 2 is explained in Table 1.

The techniques for manipulating block diagrams are given in Chap. 20, Sect. 3.

### Alphabetical Index for Definitions of Terms in Feedback Control

Table 2 is patterned after a similar table in *Letter Symbols for Feedback Control Systems* (Ref. 1). See also Ref. 2.

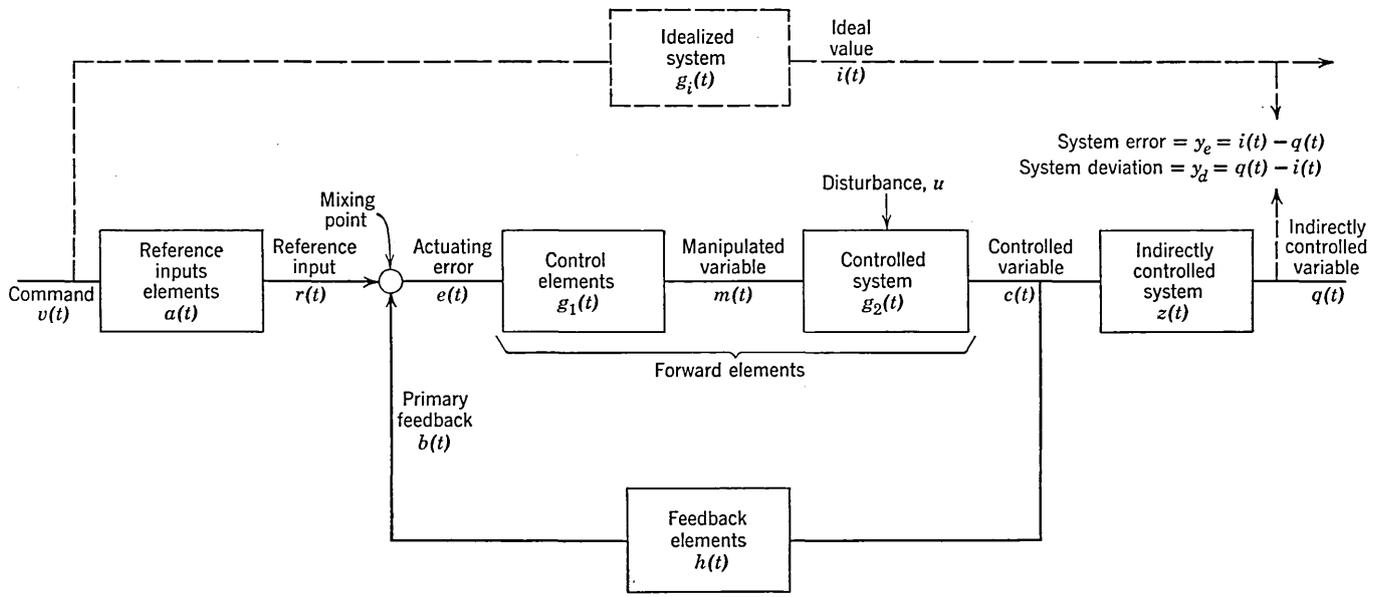


FIG. 1. Block diagram of feedback control system containing all basic elements.

TABLE 2. DEFINITIONS

Term	Definition
Actuating signal ratio	<p>Actuating signal ratio is the transform ratio of the actuating signal to the reference input.</p> <p>Under linear conditions this ratio is expressed as a Laplace transform</p> $\frac{E}{R}(s) = \frac{1}{1 + G(s)H(s)}.$
Control accuracy	Control accuracy is the degree of correspondence between the ultimately controlled variable and the ideal value.
Control area	A control area of a feedback control system is the time integral of the absolute value of the difference between the controlled variable and its final value following a specified step input or disturbance. The step input or disturbance must be specified in location and magnitude.
Control precision	Control precision is the degree of reproducibility of the ultimately controlled variable for several independent applications of the same reference input under the same operation conditions.
Control ratio	<p>Control ratio is the transform ratio of the controlled variable to the reference input. Under linear conditions, this ratio is expressed mathematically as the Laplace transform</p> $\frac{C}{R}(s) = \frac{1}{1 + G(s)H(s)}.$
Corner frequency	The corner frequency of a factor of a transfer function is the frequency at which lines asymptotic to its log magnitude curve intersect.
Dead time	Dead time is a fixed interval of time between the impression of an input on an element or system and the undistorted response to the input. Often, this is taken as the time to reach either 10% or 50% of the total change.
Error ratio	<p>The error ratio is the transform ratio of the system error with respect to the reference input. Under linear conditions, the ratio is expressed mathematically as <math>\frac{Y_e}{R}(s)</math>. In simple systems where the system error is equal to the actuating signal, the actuating ratio becomes the error ratio, i.e.,</p> $\frac{Y_e}{R}(s) = \frac{E}{R}(s) = \frac{1}{1 + G(s)}.$
Feedback control system	A feedback control system is a control system which tends to maintain a prescribed relationship of one system variable to another by comparing functions of these variables and using the relationship as a means of control.
Feedback controller	The feedback controller is a mechanism which measures the value of the controlled variable, accepts the value of the command, and, as the result of a comparison, manipulates the controlled system in order to maintain an established relationship between the controlled variable and the command.

TABLE 2. DEFINITIONS (*Continued*)

Term	Definition
Frequency response	The frequency response of a system or element is the steady-state ratio of magnitude and the difference in phase of the output with respect to a sinusoidal input. The range of frequency and conditions of operation and measurement must be specified.
Gain, magnitude, attenuation	Gain of a system or element is the ratio of magnitude of the output with respect to the magnitude of sinusoidal input. The frequency and conditions of operation and measurement must be specified.
Gain crossover	Gain crossover is a point in the plot of loop frequency response at which the magnitude of the loop frequency response is unity.
Gain margin	Gain margin is the amount by which the magnitude of the loop ratio of a stable system is different from unity at phase crossover. It is frequency expressed in decibels.
Input resolution	Input resolution between two variables of a system or element is the maximum change in the variable considered as the input which can be made without causing a change in the variable considered as the output. Resolution may be dependent upon conditions of operation and the operating point. If these are not specified, the maximum value of the resolution over the entire operating range and for all conditions of operation is implied.
Linear system	A system represented by a linear differential equation is a linear system. A system may be linear only within a certain region of operation. In such cases the boundaries of the region should be specified. The term is often used in a more limited sense to include only systems described by constant coefficient linear differential equations.
Log magnitude-angle diagram	A log magnitude-angle diagram is a plot of the log magnitude versus angle of a transfer function with frequency as a parameter.
Log magnitude and phase diagram (Bode diagram)	A log magnitude and phase diagram is a plot of the log magnitude and phase angle of a transfer function versus log frequency. Log magnitude and frequency may be multiplied by constants.
Loop gain	Loop gain is the magnitude of the loop ratio.
Loop input (or output) resolution	The loop input (or output) resolution at the specific variable is the input (or output) resolution when the loop is opened at the specified variable. When the loop is opened in a feedback control system, the dependent variable shall be called the output and the independent variable shall be called the input. When using letter symbols, the subscripts <i>o</i> and <i>i</i> shall be used with the variable to designate output and input respectively, viz., $e_o$ and $e_i$ ; indicate that the loop is opened at the actuating signal.

TABLE 2. DEFINITIONS (*Continued*)

Term	Definition
Loop ratio or open loop frequency response	<p>Loop ratio is the transform ratio or frequency response of the primary feedback to the actuating signal. Under linear conditions the ratio is expressed mathematically as the Laplace transform:</p> $\frac{B}{E}(s) = G(s)H(s).$
Mixing point	The mixing point is a symbol used in block diagrams to denote the combining of two or more signals so that the output is a function of the input signals. The direction of signal flow is indicated by arrows. The summing point is a special case of the mixing point.
Nonlinear system	A system represented by a nonlinear differential equation is a nonlinear system.
Normal time response	The normal time response is the time response with zero initial energy storage.
Nyquist diagram	A Nyquist diagram is a closed polar plot of a loop transfer function from which stability may be determined. For a single loop system it is a map on the $H(s)G(s)$ -plane of an $s$ -plane contour which encloses the entire right half of the $s$ -plane.
Output resolution	Output resolution between two variables of a system or element is the minimum change in the variable considered as the output which can be made by changing the variable considered as the input. Resolution may be dependent upon conditions of operation and the operating point. If these are not specified, the maximum value of the resolution over the entire operating range and for all conditions of operation is implied.
Parametric variation	A parametric variation is a change in system properties which affect the performance or operation of the feedback control system.
Phase crossover	Phase crossover is a point on the plot of loop ratio at which its phase angle is $-180^\circ$ .
Phase margin	Phase margin is the angle by which the phase of the loop ratio of a stable system differs from $-180^\circ$ at gain crossover.
Primary feedback ratio	Primary feedback ratio is the frequency response of the primary feedback to the reference input. Under linear conditions this ratio is expressed mathematically:
	$\frac{B}{R}(s) = \frac{G(s)H(s)}{1 + G(s)H(s)}.$
Response time, transient response	The response time of a system or element is the time required for the output to first reach a specified value after the application of a step-input or disturbance.
Rise time	The rise time of a system or element is the time required for the output to increase from one specified percentage of the final value to another following the application of a step input. Usually the specified percentages are 10% and 90%.

TABLE 2. DEFINITIONS (*Continued*)

Term	Definition
Rise time ( <i>continued</i> )	The slope at 50% of the final value is sometimes defined as rise time.
Root locus diagram	The root locus diagram is a graphical method for finding the closed loop roots of the characteristic equation. The analysis starts from the factored roots of the open loop equation and uses the complex plane for the manipulation. The root locus is the locus of the closed loop roots as a function of open loop gain.
Servomechanism	A servomechanism is a feedback control system in which the controlled variable is mechanical position.
Settling time	The settling time of a system or element is the time required for the absolute value of the difference between the output and its final value to become and remain less than a specified amount, following the application of a step input or disturbance. The specified amount is often expressed in terms of per cent of the final value (typical values: 2 or 5%).
Spurious command	A spurious command is an undesired component of the command.
Spurious reference input	A spurious reference input is an undesired component of the reference input.
Stability	Stability is the property of a system or element whose response to a stimulus dies down, if the stimulus is removed. A statement that a system is stable means that the system is stable under all normal operating conditions and for all types of stimuli normally encountered. A system may be referred to as being stable in one region of operation and not in another. If such is the case, the region of stability should be specified. A system is normally regarded as unstable if its output is oscillating in the absence of stimuli, but may be specified as being stable if the oscillations are less than a prescribed magnitude.
Steady-state error	Steady-state error is the error which remains after the transient has expired.
Stimulus	A stimulus is any type of signal which affects the controlled variable; for example, reference input and disturbance.
Summing point	The summing point is a descriptive symbol used in block diagrams to denote the algebraic summation of two or more signals. The direction of information flow is indicated by arrows and the algebraic nature of the summation by plus and minus signs.
Time constant	The time constant of an exponential component of a transient response is the time required for the component to decay from one value of $1/e$ to that value. (a) In an element or system whose response to a step input is a first order exponential, the time constant is the output change to be completed, divided by the rate of change of the output.

TABLE 2. DEFINITIONS (*Continued*)

Term	Definition
Time constant ( <i>continued</i> )	<p>If the response to a step input is not a first order exponential, this ratio is not constant with time. In such a case, the definition in the first paragraph still applies to each exponential component of the response.</p> <p>(b) Mathematically the time constant is equal to <math>T_1</math> in the transform factor <math>(1 + T_1s)^{-1}</math> and in the exponential <math>\exp(-t/T_1)</math>.</p>
Time response	The time response of a system or element is the output as a function of time following the application of a prescribed input under specified operating conditions.
Total overshoot	The total overshoot is the maximum negative value of the system error for a specified positive stimulus.
Transfer function	<p>The transfer function of a system or element is the relationship between the output and the input under specified conditions.</p> <p>In a linear system the transfer function is the ratio of the transform of the output to the transform of its input under the conditions of zero initial energy storage. It is a complete description of the dynamic properties of a system and may be represented as a mathematical expression, the frequency response, or the time response, to a specified input.</p>
Transfer locus	The transfer locus of a system or element is a complex plot of its frequency response.
Transient error	The transient error is the difference between the system error at any time and the steady-state system error for a specified positive stimulus.
Transient overshoot	The transient overshoot is the maximum negative value of the transient error.
Transient response	See time response.
Ultimately controlled variable	The ultimately controlled variable is a general term which refers to the indirectly controlled variable, or in the absence of such a quantity corresponds to the controlled variable.

**Alternate Symbols Commonly Used**

Table 3 gives a comparison of the preferred nomenclature and some of the alternate symbols and nomenclature that may be encountered. See Ref. 3.

TABLE 3. A GLOSSARY OF ALTERNATE TERMS AND SYMBOLS <sup>a</sup>

Preferred Term and Symbol	Other Terms and Symbols Used
Command, $v$	Input, desired value, set point, control point: $\theta_i, d, \theta, i$ as subscript
Reference input elements, $a$	Conversion elements, primary element, sensing elements
Reference input, $r$	Input, reference standard, desired value, set point: $\theta, v, \theta_i, d, i$
Actuating signal, $e$	Error, unbalance, actuating error correction, deviation: $\epsilon, \theta, \theta_e, u$
Control elements, $g_1$	Amplifier, controller, servo amplifier, relay, error corrector: $y, kg, h, \mu, a$
Manipulated variable, $m$	...
Disturbance, $u$	Load disturbance, upset, noise, drift: $l, d, n$
Controlled system, $g_2$	Process, plant, load: $y, kg, h, \mu, a$
Controlled variable, $c$	Output, regulated variable, measured variable: $\theta_o, \theta, r, o$ as subscript
Indirectly controlled system, $z$	Process, plant, load
Indirectly controlled variable, $q$	Output, regulated variable: $\theta_o, \theta, r, o$ as subscript
Feedback elements, $h$	Feedback transfer function: $y, kg, g, \beta$
Primary feedback, $b$	Monitoring feedback, feedback: $\theta_f$
Summing point	Error detector, error measuring means, discriminator
Idealized system, $g_i$	Ideal system, desired system, preferred system, reference system
Ideal value, $i$	Desired value
System error, $y_e$	Error, deviation
System deviation, $y_d$	Error, deviation

<sup>a</sup> From A.I.E.E. Subcommittee Report (Ref. 3, p. 908).

### 3. FEEDBACK CONTROL SYSTEM DESIGN CONSIDERATIONS

#### Usefulness of Feedback

A feedback control is a *closed loop* control in contrast to an *open loop* control in which no information is fed back from the output. The closed loop type of control potentially has the advantages of providing:

1. Less dependence upon system component characteristics.
2. Less sensitivity to load disturbances.
3. Faster response to command signals.
4. Extreme modification of system behavior.

Feedback controls are thus more accurate and the performance is more predictable than open loop controls. Feedback controls have found wide application where these characteristics are desired. Open loop controls, however, are simpler and usually less expensive and are the choice in applications where reduced accuracy and performance are not deleterious and also where a feedback signal is not available; e.g., in a ballistic missile.

#### Major Steps in Feedback Control Design

Design itself is a feedback process. The engineer will use an iterative process in which the following steps are repeated several times. Each time the effect of design modifications on the inputs to the design are observed, and the resulting refinement converges on the final design.

**1. Determine Process Characteristics.** The characteristics of the process to be controlled must be adequately understood and described either by literal expressions or graphical means. Both the dynamic and static characteristics must be represented. It may require extensive analytical investigation and/or experimental tests to define properly the process characteristics.

**2. Specify System Performance.** The desired system performance must be specified or defined before synthesis can proceed. The performance must be interpreted in terms that can be used in synthesizing the controller. Often the customer's requirements will not be in a directly usable form and must be translated into a more convenient one.

Frequently used specifications are indicated in Table 4. The transient specifications must be prescribed for a particular input signal. Because of the practical difficulties, the signal representation is usually limited to simple aperiodic or combinations of aperiodic signals for analytical investigation. In cases where this is not sufficient, analog or digital computers are normally used to determine performance.

The frequency response and transient specifications are related uniquely in a *linear system*; however, the relations are difficult to manipulate and

as a result the specifications are often redundant. No explicit relations exist in nonlinear systems, and it is necessary to specify in detail all significant characteristics.

**3. Selection of Power Elements.** The task to be performed by an automatic control will require certain output power, velocities, and accelerations, and an average power that is dependent upon the duty cycle. This information is quite often not known in detail, and values are assumed. These values will determine the selection of the power element. The choice of electrical, hydraulic, pneumatic elements must be made on the basis of weight, cost, performance, and the ability to be integrated with the overall system. Within the limits of the state of the art, the power element will set the major lags of the controller and, for amplifying control systems such as a power positioning system, it will constitute the major time lags of the system. For this reason the final selection is often deferred until further study of the control performance to determine if a lower performance and less expensive (in cost and/or weight, size, etc.) unit can be used. The environmental conditions to which the equipment will be subjected must be considered as well as the performance requirements. Under extreme ambients or corrosive atmospheres performance may have to be sacrificed in order to obtain the necessary operating life.

**4. Synthesize Controller.** Within the context of the defined process and the preliminary selection of the power elements the actual control design is made. This design must be based upon practical compensating networks and feedbacks and the availability of the necessary components to meet the system performance requirements.

The optimum design of control systems is primarily a matter of engineering judgment. Although various criteria have been advanced and received some application, the complex and diverse factors which influence the final selection are not amenable to simple mathematical representation. No adequate criterion for use as an *index* is available.

In selection of the system compensation, consideration must be given to:

*a. Extraneous Signals.* Satisfying the system requirements is straightforward if the input contains only correct information. Unfortunately, most input signals will contain noise or other extraneous signals. The feedback control must be designed to be selective to the correct signal, and in so far as possible it must reject the extraneous signals. The extraneous signals may occur at points in the system other than the input such as load disturbances. Inclusion of these effects may radically influence the system design.

*b. Nonlinearities.* No system is free from nonlinearities, and these effects must be considered before the design is complete. The effect of nonlinearities *may* be to reduce performance or even to cause instability.

TABLE 4. COMMON PERFORMANCE SPECIFICATIONS

Type	Specified	Definition	General Remarks
1. Transient overshoot	Transient response	Usually taken as ratio of peak of transient to final value for a step command or disturbance.	Convenient when transient solution is available. Can be estimated from root locus or frequency response. Useful for nonlinear systems. System must be underdamped. Used for regulators, meters, position servo-mechanisms which are normally excited by step inputs and are underdamped.
2. Settling time	Transient response	Defined as time to reach and remain within a specified percentage of final value (often as 5% or 2%) after a step command or disturbance.	See 1 above. Used for systems which require rapid synchronization, e.g., fire control system.
3. Steady-state error	Transient response	Final error existing between desired and actual output.	See 1 above. Easily calculated from static characteristics or final value theorem. Useful when input is simple aperiodic function. Can include frequency components which arise in nonlinear systems.
4. Rise time	Transient response	Defined as (a) time to $\frac{1}{2}$ the final value, or (b) slope at $\frac{1}{2}$ the final value, or (c) time between 10% and 90% of final value after a step command.	Easily estimated from frequency response or root locus and is indicative of band pass of system. Used for overdamped systems. Has found application in process controls where characteristics (1) or (2) may not be easily recognized.
5. Dead time	Transient response	Defined as (a) time for output action to be initiated, or (b) for output to reach a given level (10% or 50%), or (c) time to the intersection of the slope	See 1 above. Easily estimated from frequency response and is indicative of phase shift near gain crossover in systems. Useful when delay times exist in system. Used for overdamped systems. Both rise and delay time derive from filter theory.

6. Absolute damping, decrement factor	Transient response or root locus	of the transient at $\frac{1}{2}$ the final value and the initial value after a step input. Defined as the real part of the roots of a quadratic system and as such determines the rate of decay of transient.	Convenient method of interpreting more complex systems in terms of quadratic systems. Valuable in combination with relative damping in work with root locus analysis. Has had extensive use in systems demanding prescribed transient performance, particularly when the time decay is important, e.g., in autopilots.
7. Damping ratio	Transient response or root locus	Damping ratio is defined as $\zeta$ in the quadratic $s^2 + 2\omega_0\zeta s + \omega_0^2$ and indicates the decay per cycle of the natural frequency.	Useful because it is a parameter in nondimensional plot of quadratic response. Used in combination with 6 above in root locus analysis. Used when number and size of overshoot are important. In combination with 6 above defines decay of oscillatory component of transient.
8. Load sensitivity	Transient or frequency response	The sensitivity of the output to disturbance occurring at the output or other extraneous signals within the system. Often expressed as the ratio of the output for a unit disturbance to the output for a unit signal input. Can define steady-state and transient sensitivity.	Useful when analog solutions are available. Steady-state specifications can be readily determined from system equations. If disturbances can be expressed as frequencies then frequency response allows simple interpretation. Used in fire control, autopilot, voltage regulators, etc., where load disturbances (firing torques, wind gusts, load changes) are important considerations.

TABLE 4. COMMON PERFORMANCE SPECIFICATIONS (*Continued*)

Type	Specified	Definition	General Remarks
9. Minimum error criterion	Transient or frequency response	The response of the system is adjusted to minimize a function of the total error that results from both signal and noise or extraneous signals. The criterion may take several forms, e.g., min. squared error, min. absolute error times time.	Used to optimize system response to reject unwanted noise, and pass the true signal. Used to specify <i>performance index</i> when just signal is considered. Within basic assumptions frequency response analysis is very useful. Used on systems which operate on random or noisy data, e.g., missile radar guidance and fire control. Analog computers can be used to apply criterion to nonlinear systems.
10. Phase margin	Frequency response	Defined as $180^\circ +$ phaseshift at unity gain of the open loop frequency response.	Used as a rule of thumb in frequency response analysis to indicate stability and performance. Easy to use and to obtain directly from frequency response diagram.
11. Gain margin	Frequency response	Gain margin is ratio of maximum stable gain to actual gain, i.e., gain at phase crossover.	Same as 10. Indicates relative sensitivity of system to gain variations. Can be calculated by Routh's criterion. Not as good a criterion for performance as 10. Little used.
12. $M_m$ peak	Frequency response	Ratio of maximum of closed loop frequency response to a low frequency value.	Used with Nyquist and frequency response analysis. Rules of thumb relate $M_m$ and transient overshoot. Easy to calculate from frequency response diagram.
13. Band width	Frequency response	Defined variously ( <i>a</i> ) usually as frequency where closed loop response falls to $\sqrt{\frac{1}{2}}$ or 3 db of its low frequency value, or ( <i>b</i> ) sometimes as the frequency at the significant peak $M_m$ ,	Used with frequency response analysis and is related to speed of response of system. Used also when definite frequency bandpass is needed for fidelity. $M_m$ , bandpass, and the phase shift at these values give a good indication of the closed loop response and are often used when a number of closed loops are

14. Static error coefficient	Frequency response	<p>or (c) the crossover of the open loop response.</p> <p>Defined as the final error resulting from a continuous input of position, or velocity, or acceleration, etc. The magnitude of the input and the maximum tolerable error must be specified.</p>	<p>operated in tandem as system.</p> <p>Used to set low-frequency gain of open loop frequency response. Useful where steady inputs are encountered.</p>
15. Dynamic error coefficients (or steady-state error coefficients)	Frequency response and root locus	<p>Defined as the steady-state error resulting from the derivatives of the input function. The time function and/or its derivatives must be specified as well as the maximum tolerable error.</p>	<p>Relates system gain and time constants to errors arising from higher derivatives of input. Used to estimate error resulting from varying input to given system and conversely to determine closed loops pole-zero location to give desired error. Accurate where input varies at slow rate compared to bandpass. Becomes poorer as input varies more rapidly because of transient effects. Used in analysis of fire controls, machine controls, etc., where input varies in an expected manner.</p>
16. Maximum system error	Transient response	<p>Defined as the maximum tolerable system error, <math>y_e</math>. The input function and operating conditions must be specified.</p>	<p>Distinguished from steady-state error because maximum error under dynamic conditions is specified. Normally used to define performance with a varying input, e.g., automatic milling machine control. Not usually used with simple aperiodic inputs. Used in conjunction with minimum error criterion (9) to place absolute bound on error.</p>

TABLE 4. COMMON PERFORMANCE SPECIFICATIONS (*Continued*)

Type	Specified	Definition	General Remarks
17. Resolution	Low level characteristics	Defined as the maximum tolerable change in the input without a change occurring in the output. Input and operating conditions must be specified.	Can appear in various forms, i.e., maximum position input change required to obtain output change, or minimum velocity at which a servomechanism will track with tolerable velocity error.
18. Duty cycle	Power element rating	Defined variously, depending upon application. Intent is to define the average power requirement.	Objective of specification is to allow more efficient selection and/or design of the power element. Specification can take the form of an rms power requirement or where an average does not adequately describe the situation a time distribution and level may be given. Used extensively when large power drives are involved.
19. Maximum operating conditions	Power element rating	Included to indicate the wide variety of maximum performance requirements sometimes specified, e.g., maximum velocity, maximum load torque.	Many of these limits are implied by other performance requirements. Often necessary to define implicitly load requirements separately, e.g., load running torques or power (independent of accelerating torques).

*c. Practical Aspects.* The ultimate cost and manufacturability must be considered during the synthesis. This, of course, implies ascertaining the physical realizability of the controller and assuring that practical tolerances are maintained. Reliability and ease of servicing must also receive proper consideration. Environmental conditions and customer requirements on component packaging must also be factored into the mechanical design and may affect the performance.

**5. Test and Evaluation of Equipment.** In most cases unpredicted and secondary effects will require final adjustment to be made after the actual equipment is assembled. This is often the more economical way to reach a final design when a wide range of adjustment can be included in the design or preliminary models can be built and tested relatively fast. This would be the case for many types of instrument servos. On the other hand, it would be horrendous to attempt this approach with an elaborate, expensive missile system which is expended at each test firing. In such cases the extensive use of analysis and computer facilities to minimize the testing is justified.

#### 4. SELECTION OF METHOD OF SYNTHESIS FOR FEEDBACK CONTROLS

The major analytical methods available to aid in the synthesis of feedback control systems are summarized in Table 5. No general rules are available for the selection of the proper method, and the designer should be familiar with all methods in order to select the one best suited to his problem. It is often desirable to carry root locus and frequency response diagrams in parallel. The root locus supplies time domain information, and the frequency response provides the simplest method of estimating the method of compensation.

#### System Optimization

None of these techniques allows a completely systematic design approach. The major difficulty is in defining and specifying optimum performance and determining what performance index to use for evaluation. See Chap. 24. Although criteria have been proposed, the mathematical labor involved in the more sophisticated is prohibitive. Actually the accuracy and extent of the available data usually warrant the use of only the more simple criteria. These criteria do not encompass the entire problem and therefore must be used carefully. The material in the following chapters presents the available useful design criteria.

TABLE 5. SUMMARY OF MAJOR ANALYTICAL TECHNIQUES FOR FEEDBACK CONTROL SYSTEM ANALYSIS

Type	Usefulness	De- scribed in
1. Differential equations	Classical solutions of differential equations are generally too involved for practical use in synthesis. Nondimensional performance charts help on second order systems. Significance of individual system component values difficult to ascertain.	Chap. 20
2. Routh Hurwitz criterion	Used to determine the limiting stability conditions. Can be extended to include damping factor only with difficulty. Limited usefulness.	Chap. 21
3. Root locus	The best solution to the problem of directly synthesizing the time response. Particularly useful when the performance specifications are in terms of the time response. Construction of the diagrams can be time-consuming and the performance can be sensitive to small relative changes of locus in low-frequency region.	Chaps. 21, 23
4. Frequency response	The most used approach presently available. The locus can be plotted in the form of a Nyquist, log magnitude-angle diagram, or the log magnitude and phase diagram. The latter has the advantages of easy construction by templates and of easy introduction of compensating characteristics. Easy to include experimental data in frequency response analysis. The difficulty of relating transient and frequency response is a limitation.	Chap. 23
5. Describing functions	An extension of the frequency response techniques to nonlinear systems. Good performance criterion not available. Method can treat higher order systems.	Chap. 25
6. Closed loop pole-zero location	Requires determining realizable and practical components <i>after</i> the definition of the system response. Not in wide use as yet but possesses the good feature of working directly from the desired closed loop response.	Chaps. 22, 23

### The Use of Computers

This has supplanted much of the paper design study. This approach allows rapid and complete (often visual) evaluation of the expected system performance. At the present state of the art, however, it is not possible to obtain a complete design from the computer without interpretation at various steps by the design engineer. The ultimate use of the computers will occur when a complete systematic design can be programmed; but this cannot be done until mathematical expressions can be equated to the decisions now based upon "engineering judgment."

Availability of computers has not eliminated the need for a thorough knowledge of the standard feedback control techniques for analysis. Although, when an analog computer facility is available, the conventional analytical techniques are used principally for preliminary, order-of-magnitude estimates and for verifying computer solutions, experience has shown that a thorough knowledge of alternate techniques will enhance the usefulness of the computers.

#### REFERENCES

1. *Letter Symbols for Feedback Control Systems*, ASA Y10.13-1955, American Standards Association, New York, July 1955, Sponsored by American Society of Mechanical Engineers.
2. I.R.E. Standards on Terminology for Feedback Control Systems, 1955 *Proc. I.R.E.*, 44, No. 1 (1956).
3. Proposed Symbols and Terms for Feedback Control Systems, A.S.E.E. Subcommittee Rept., *Elec. Eng.*, October 1951.



## Fundamentals of System Analysis

S. J. Jennings and A. A. Winkeljohann

1. Representation of Physical Systems	20-01
2. Classical Methods of Analysis	20-28
3. Block Diagrams	20-56
4. System Types	20-66
5. Error Coefficients	20-70
6. Analysis of A-C Servos: Carrier Systems	20-79
References	20-84

### 1. REPRESENTATION OF PHYSICAL SYSTEMS

#### Methods of System Analysis

In order to study the performance of a physical system, equations must be written from the physics of the situation to describe the excursion of all variables. To describe the operation of a physical system in mathematical form, its *differential equations* may be written which, in general, will be nonlinear in character. In many cases it is possible, by restricting the region for which results are valid, to write linear differential equations with constant coefficients for the system. The solution of the linear differential equation then yields the complete steady-state and transient response of the system for a given input. The *transient response* indicates the system stability while the *steady-state response* to a sinusoidal input is very useful in system synthesis.

TABLE 1. TYPICAL COMPONENT EQUATIONS (Ref. 10)

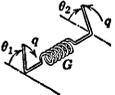
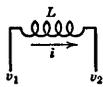
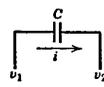
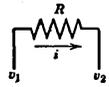
Parameter	Equation	Description
Translation systems: Mass 	$f_1 - f_2 = M \frac{d^2x}{dt^2};$ $\frac{dx}{dt} = \frac{1}{M} \int (f_1 - f_2) dt$	The net force acting on a body is equal to its mass times its acceleration with respect to an arbitrary fixed reference.
Spring 	$f = Kx; \frac{dx}{dt} = \frac{1}{K} \frac{df}{dt}$	The force which must be applied to each end of a spring to deflect it a distance $x$ is equal to the spring constant $K$ times $x$ .
Dashpot (viscous damping) 	$f = D \left( \frac{dx_1}{dt} - \frac{dx_2}{dt} \right)$	The force which must be applied to each end of a dashpot to produce a relative motion of its two ends is equal to the viscous damping coefficient $D$ times the relative velocity.
Rotational systems: Inertia 	$q_1 - q_2 = J \frac{d^2\theta}{dt^2};$ $\frac{d\theta}{dt} = \frac{1}{J} \int (q_1 - q_2) dt$	The net torque acting on a body is equal to its inertia times its angular acceleration with respect to an arbitrary fixed reference.
Torsional spring 	$q = G(\theta_1 - \theta_2)$	The torque which must be applied to each end of a torsional spring to produce a relative angular deformation $\theta_1 - \theta_2$ of its two ends is equal to the rotational spring constant times the angular deformation.
Rotational dashpot 	$q = B \frac{d\theta}{dt}$	The torque which must be applied to a rotational dashpot to cause it to rotate with an angular velocity is equal to the rotational viscous damping coefficient times the angular velocity.

TABLE 1. TYPICAL COMPONENT EQUATIONS (Ref. 10) (Continued)

Parameter	Equation	Description
Electrical systems:		
Inductance	$v_1 - v_2 = L \frac{di}{dt};$ $i = \frac{1}{L} \int (v_1 - v_2) dt$ 	The voltage drop caused by current flowing in an inductance is equal to the inductance times the rate of change of the net current flowing in the direction of the drop.
Capacitance	$v_1 - v_2 = \frac{1}{C} \int i dt;$ $i = C \frac{d}{dt} (v_1 - v_2)$ 	The voltage drop caused by current flowing through a capacitance is equal to the integral of the net current flowing through the capacitance divided by its capacitance.
Resistance	$v_1 - v_2 = Ri;$ $i = \frac{1}{R} (v_1 - v_2)$ 	The voltage drop caused by current flowing through a resistance is equal to the net current flowing through the resistance multiplied by the resistance.

English Gravitational Units		Electrical Units	
<i>t</i>	time, seconds	<i>q</i>	torque, lb-ft
<i>x</i>	distance, feet	<i>J</i>	inertia, slug-ft <sup>2</sup>
<i>M</i>	mass, slugs	<i>θ</i>	angle, radians
<i>K</i>	spring constant, lb/ft	<i>G</i>	torsional spring constant, lb-ft/rad
<i>D</i>	damping coefficient, lb/ft/sec	<i>B</i>	rotational damping coefficient, lb-ft/rad/sec
		<i>t</i>	time, seconds
		<i>v</i>	voltage, volts
		<i>i</i>	current, amperes
		<i>L</i>	inductance, henrys
		<i>C</i>	capacitance, farads
		<i>R</i>	resistance, ohms

The solution to differential equations by either the direct method or by Laplace transformations is useful primarily in the *analysis* of a given system with all parameters prescribed. This approach is less useful in the design or synthesis of a control since the effect of the variations of parameters on the exponential time function exponents is difficult to visualize. For more complex systems the problem of factoring the high order polynomial characteristic equation becomes quite laborious.

For *synthesis* the root locus, frequency response, and closed loop pole zero location methods are recommended. (Chaps. 21, 22, and 23.)

Even for analysis the classical time solution has been largely supplanted by the wide usage and availability of *analog computers*. As a result the

classical techniques of solution are used primarily as checks on analog computer results or as aids in visualizing the basic performance. The charts included in Sect. 2 are useful in this case.

Although the solution of differential equations is no longer of paramount importance, the correct description of the system or component dynamic performance by *differential equations* is basic to all methods of analysis and synthesis. It is most important that the control designer understand differential equations and their application to his field of endeavor.

A suggested approach for obtaining these physical equations is:

1. Understand the system well enough to draw a schematic diagram showing the relationship of all variables, including all pertinent components as well as the load.
2. Replace the schematic with equivalent circuits or analogies.
3. Rearrange this diagram into convenient noninteracting sections or blocks.
4. Write the characteristic equation of each section from the functional relationship.
5. Obtain the transfer function from these equations.
6. Simplify this block diagram and obtain the complete system characteristic equation by algebraic manipulation.

This sequence is an analysis approach; synthesis of a system reverses this method after starting with known requirements to obtain a system equation.

**Physical Laws.** To write the equations which mathematically describe the system or component performance, it is necessary to understand the basic operation of the device and the physical laws governing the various processes involved. The wide field of application of feedback control theory makes it prohibitive to list all the fundamental laws that might be required. The following partial list of textbooks in the particular field of interest for these physical laws and Table 1 are useful.

1. Physics: Erich Hausmann and E. P. Slack, *Physics*, Van Nostrand, Princeton, N. J., 1948.
2. Electrical: W. L. Everitt, *Communication Engineering*, McGraw-Hill, New York, 1937.
3. Thermodynamics: P. J. Kiefer and M. C. Stuart, *Principles of Engineering Thermodynamics*, Wiley, New York, 1954.
4. Fluid Mechanics: R. C. Binder, *Fluid Mechanics*, Prentice-Hall, New York, 1949.
5. Kinematics: J. L. Synge and B. A. Griffith, *Principles of Mechanics*, McGraw-Hill, New York, 1949.
6. Circuit Analysis: E. A. Guillemin, *Mathematics of Circuit Analysis*, Wiley, New York, 1949.
7. Materials: Stephan Timoshenko, *Strength of Materials*, McGraw-Hill, New York, 1953.

8. Hydrodynamics: H. Lamb, *Hydrodynamics*, The University Press, Cambridge, England, 1932.

9. Mechanics: F. B. Seely, *Analytical Mechanics for Engineers*, Wiley, New York, 1952.

EXAMPLES. The following examples illustrate the use of basic physical laws and Table 1 in obtaining the equations describing the system performance. Whenever possible, simplifying initial conditions are chosen.

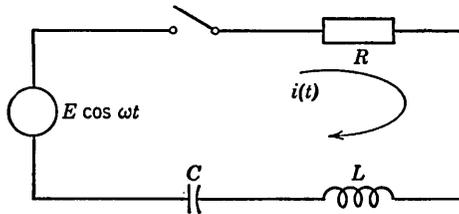


FIG. 1. Electric circuit.

1. An electric circuit such as Fig. 1 requires:

KIRCHHOFF'S LAW. *The summation of voltage drops in a closed loop is equal to zero.*

$$(1) \quad E \cos \omega t = Ri(t) + L \frac{di(t)}{dt} + \frac{1}{C} \int_0^t i(t) dt.$$

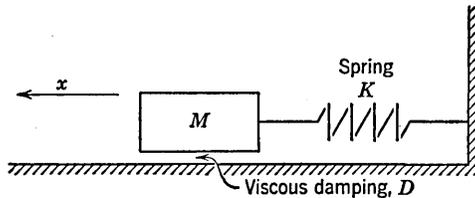


FIG. 2. Damped spring mass system.

2. A spring mass system such as Fig. 2 requires:

NEWTON'S LAW. *The summation of forces acting on a body equals the change in momentum.*

$$(2) \quad M \frac{d^2x}{dt^2} = -D \frac{dx}{dt} - Kx \quad \text{or} \quad (Ms^2 + Ds + K)x = 0.$$

3. A combined electrical and rotational mechanical system is a d-c motor with fixed field excitation (ignoring armature inductance) and a pure inertia load is shown in Fig. 3.

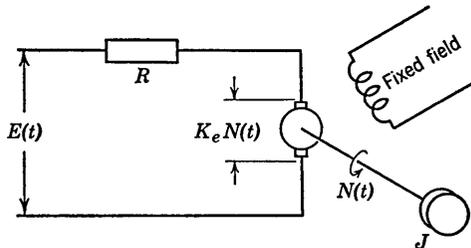


FIG. 3. D-c motor with inertia load.

Summing voltage drops:

$$(3) \quad E(t) = Ri(t) + K_eN(t),$$

where  $K_e$  = motor voltage constant,  
 $N(t)$  = motor speed.

Summing torques:

$$(4) \quad J \frac{dN(t)}{dt} = K_t i(t),$$

where  $K_t$  = motor torque constant,  
 $J$  = motor inertia.

Eliminating  $i(t)$  from eqs. (3) and (4) results in the transfer function of output speed to input voltage:

$$(5) \quad \frac{N(t)}{E(t)} = \frac{1}{K_e \left( \frac{RJ}{K_e K_t} s + 1 \right)} = \frac{1}{K_e (T_m s + 1)},$$

where  $s = d/dt$ ,

$$T_m = RJ/K_e K_t = \text{time constant.}$$

4. A common electromechanical system is a synchro with a pure inertia

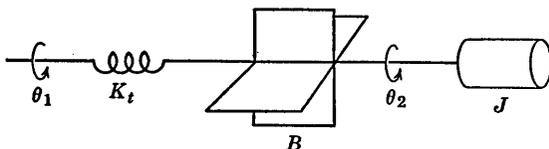


FIG. 4. Schematic of synchro system.

load and with viscous damping as shown schematically in Fig. 4, where  $K_t$  is the torque gradient.

Summing torques:

$$(6) \quad J \frac{d^2\theta_2}{dt^2} + B \frac{d\theta_2}{dt} = K_t(\theta_1 - \theta_2) \quad \text{or} \quad (Js^2 + Bs + K_t)\theta_2 = K_t\theta_1.$$

Further examples are used throughout this section.

### Circuit Simplification Techniques

*Analogies* are useful in setting up physical systems and interpreting their boundary conditions since this approach compares known systems with the unknown. Often thermal, mechanical, hydraulic, etc., systems are converted to an electrical equivalent since electric circuit analysis methods have been developed to a high degree. *Examples* of conversions of physical systems to electrical equivalents are given in Ref. 2.

**ANALOGIES.** The following equations show the analogies among three systems:

<i>Equations</i>	<i>System</i>
(7) $M \frac{d^2x}{dt^2} + D \frac{dx}{dt} + K_s x = f(t)$	Mechanical translatory system
(8) $J \frac{d^2\theta}{dt^2} + F \frac{d\theta}{dt} + K_t \theta = f(t)$	Mechanical rotation system
(9) $L \frac{d^2q}{dt^2} + R \frac{dq}{dt} + \frac{q}{C} = f(t)$	Electric circuit

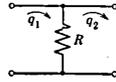
where  $M$  = mass, slugs,  
 $D$  = damping, lb/ft/sec,  
 $K_s$  = spring gradient, lb/ft,  
 $x$  = distance, ft,  
 $K_t$  = torque gradient, lb/ft,  
 $\theta$  = angular displacement, rad,  
 $L$  = inductance, henrys,  
 $R$  = resistance, ohms,  
 $J$  = inertia, slug-ft<sup>2</sup>,  
 $F$  = friction, lb/rad/sec,  
 $C$  = capacitance, farad,  
 $q$  = charge, coulombs.

Analogous elements are listed in Table 2.

TABLE 2. ANALOGOUS ELEMENTS (Ref. 2)

Electrical elements

Electrical resistor



$$E_R = R \frac{d(q_1 - q_2)}{dt}$$

$R$  = resistance  
 $q$  = charge

Electrical capacitor



$$E_C = \frac{1}{C} (q_1 - q_2)$$

$C$  = capacitance

Electrical inductor



$$E_L = L \frac{d^2q}{dt^2}$$

$L$  = inductance

Mechanical elements (translational)

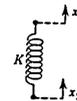
Viscous damper



$$f = D \frac{d(x_1 - x_2)}{dt}$$

$x$  = displacement  
 $D$  = damping coefficient

Spring



$$f = K(x_1 - x_2)$$

$K$  = spring constant

Inertia

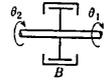


$$f = M \frac{d^2x}{dt^2}$$

$M$  = inertia

Mechanical elements (rotational)

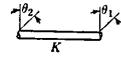
Torsional damper



$$T = B \frac{d(\theta_1 - \theta_2)}{dt}$$

$B$  = damping coefficient

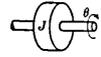
Shaft stiffness



$$T = K(\theta_1 - \theta_2)$$

$K$  = stiffness coefficient

Inertia

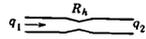


$$T = J \frac{d^2\theta}{dt^2}$$

$J$  = moment of inertia

Hydraulic elements

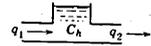
Fluid resistance



$$P = R_h q = R_h \frac{dQ}{dt}$$

$R_h$  = hydraulic resistance  
 $q$  = rate of flow  
 $Q$  = quantity of flow  
 ( $Q_2 = 0$  in electrical analog)

Fluid capacity



$$P = \frac{Q_1 - Q_2}{C_h}$$

$C_h$  = hydraulic capacity  
 $Q_1$  = quantity of inflow  
 $Q_2$  = quantity of outflow  
 $P$  = pressure

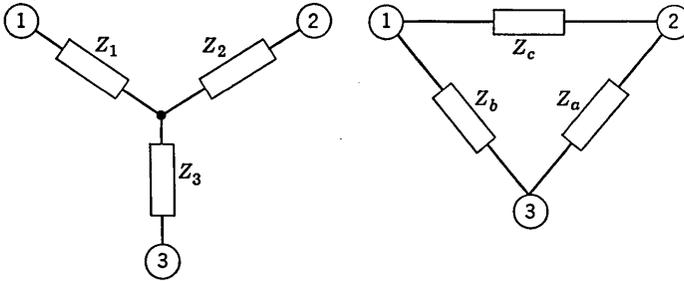


FIG. 5. Wye-delta transformation.

**Aids for Circuit Simplification.** The following techniques are useful in reducing the system equations to simpler form:

**WYE-DELTA TRANSFORMATION.** *The circuits of Fig. 5 are equivalent if the following relations are satisfied:*

$$(10) \quad Z_1 = \frac{Z_b Z_c}{Z_a + Z_b + Z_c}$$

$$(13) \quad Z_a = \frac{Z_1 Z_2 + Z_2 Z_3 + Z_3 Z_1}{Z_1}$$

$$(11) \quad Z_2 = \frac{Z_a Z_c}{Z_a + Z_b + Z_c}$$

$$(14) \quad Z_b = \frac{Z_1 Z_2 + Z_2 Z_3 + Z_3 Z_1}{Z_2}$$

$$(12) \quad Z_3 = \frac{Z_a Z_b}{Z_a + Z_b + Z_c}$$

$$(15) \quad Z_c = \frac{Z_1 Z_2 + Z_2 Z_3 + Z_3 Z_1}{Z_3}$$

**SUPERPOSITION.** *If a system is linear the system response to several inputs will be the sum of the response to each input separately (refer to Fig. 6).*

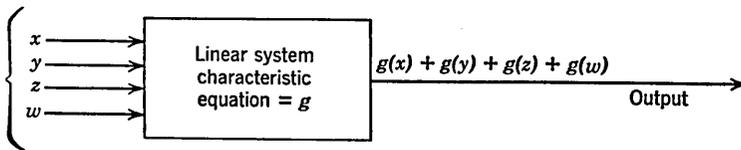


FIG. 6. Superposition.

**THEVENIN'S THEOREM.** *The effect of any impedance element in a circuit may be determined by replacing all the voltage sources by a single equivalent voltage source and all other impedances by a signal impedance in series with the impedance of interest. For Fig. 7, the equivalent voltage  $E_{ab}$  is equal to the open circuit voltage that is present across  $a-b$  with the circuit broken at  $a-b$ .*

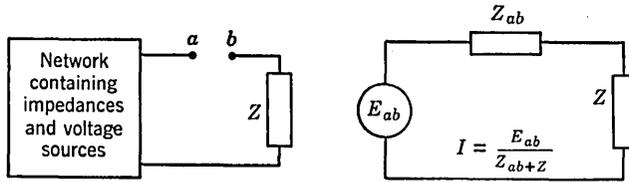


FIG. 7. Thevenin's theorem.

**NORTON'S THEOREM.** *The current in any impedance  $Z_R$ , connected to two terminals of a network, is the same as if  $Z_R$  were connected to a constant-current generator whose generated current ( $I_{SC}$ ) is equal to the current which flows through the two terminals when these terminals are short-circuited, the constant-current generator being in shunt with an impedance equal to the impedance of*

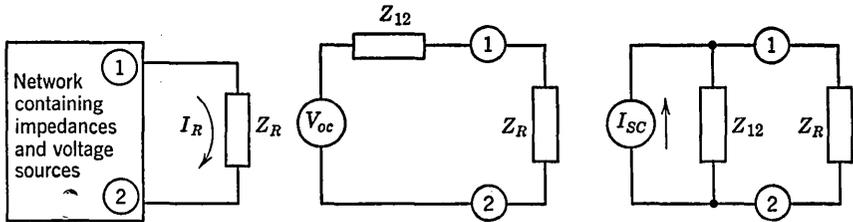


FIG. 8. Equivalent circuits using Norton's theorem.

*the network looking back from the terminals in question.* This theorem is similar in many respects to Thevenin's theorem. It is illustrated by Fig. 8.

**Nodal and Mesh Analysis.** A general approach to circuit analysis is illustrated by Fig. 9a, b and eqs. (16) through (21). In the *nodal analysis*

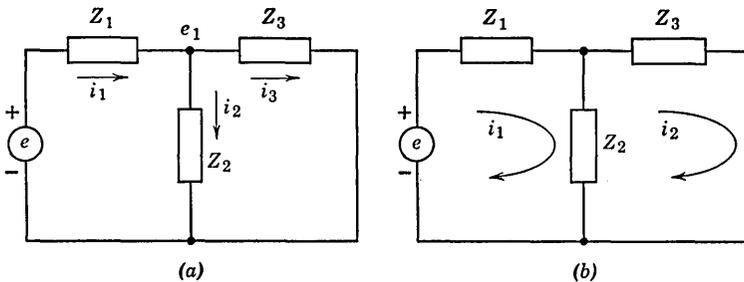


FIG. 9. (a) Nodal approach; (b) mesh or loop approach.

the summation of currents at a junction or node is equal to zero. This is useful in solving for an unknown voltage, given driving voltages and im-

pedances. From Fig. 9a,

$$(16) \quad i_1 - i_2 - i_3 = 0,$$

$$(17) \quad \frac{e - e_1}{Z_1} - \frac{e_1}{Z_2} - \frac{e_1}{Z_3} = 0,$$

$$(18) \quad e_1 = \frac{eZ_2Z_3}{Z_2Z_3 + Z_1Z_2 + Z_1Z_3}.$$

Use of the *mesh analysis* uses voltage summations about the closed loops. Usually the unknown current, such as  $i_2$  of Fig. 9b is found in terms of the known voltages and impedances. From Fig. 9b,

$$(19) \quad e = Z_1i_1 + Z_2(i_1 - i_2) = (Z_1 + Z_2)i_1 - Z_2i_2,$$

$$(20) \quad 0 = Z_2(i_2 - i_1) + Z_3i_2 = -Z_2i_1 + (Z_2 + Z_3)i_2$$

to solve for  $i_2$  by using determinates:

$$(21) \quad i_2 = \frac{\begin{vmatrix} Z_1 + Z_2 & e \\ -Z_2 & 0 \end{vmatrix}}{\begin{vmatrix} Z_1 + Z_2 & -Z_2 \\ -Z_2 & Z_2 + Z_3 \end{vmatrix}} = \frac{eZ_2}{Z_1Z_2 + Z_2Z_3 + Z_1Z_3}.$$

**Tables of Typical Transfer Functions.** *The transfer function of a system or element is the ratio of the transform of the output to the transform of its input under the conditions of zero initial energy storage.* It is a complete description of the dynamic properties of a system and may be represented as a mathematical expression of the frequency response, or the time response to a specified input.

In Tables 3 to 5 are summarized typical transfer functions in Laplace transform form for typical mechanical, electrical, and hydraulic control elements. For a more complete tabulation of transfer functions of *RC* networks see Chap. 23, Sect. 2.

Tables 6 to 8 consist of three sections of a morphological table of servo components appearing in Ref. 11.

Further material on the subject of transfer functions may be found in Refs. 3, 11, 12.

TABLE 3. SUMMARY OF TRANSFER FUNCTIONS FOR REPRESENTATIVE MECHANICAL ELEMENTS (Ref. 3a)

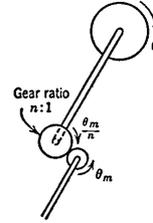
Mechanical Elements: Transfer Functions

Nomenclature

Rotation

Spring mass damper 
$$\frac{\Theta_l(s)}{\Theta_m(s)} = \frac{1/n}{(J/K_s)s^2 + (B/K_s)s + 1}$$

Rotation



- $\theta_l$  = load angular position, radians,
- $\theta_m$  = motor position, radians,
- $J$  = moment of inertia, pound-foot-seconds/second,
- $n$  = gear ratio,
- $K_s$  = shaft spring constant, pound-foot/radian,
- $B$  = damping torque coefficient, pound-feet/radian/second.

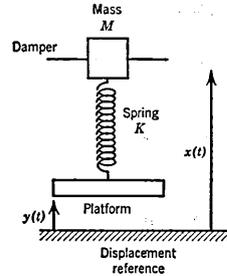
Translation

Spring mass damper 
$$\frac{X(s)}{Y(s)} = \frac{1}{(M/K)s^2 + (D/K)s + 1}$$

Spring-dashpot (lag) 
$$\frac{X(s)}{Y(s)} = \frac{1}{(D/K)s + 1}$$

Spring-dashpot (lead) 
$$\frac{X(s)}{Y(s)} = \frac{(D/K)s}{(D/K)s + 1}$$

Translation



- $x$  = mass displacement, feet,
- $y$  = platform displacement, feet,
- $M$  = mass, pound-seconds/second/foot,
- $D$  = damping coefficient, pounds/foot/second,
- $K, K_s$  = spring constant, pounds/foot.

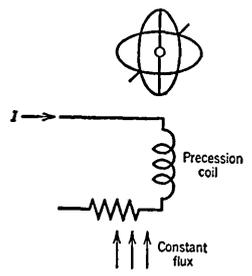
TABLE 4. SUMMARY OF TRANSFER FUNCTIONS FOR REPRESENTATIVE ELECTRIC ELEMENTS (Ref. 3a)

Electric Elements: Transfer Functions	D-C Motor	Nomenclature
D-C motor For speed control		$N$ = velocity of motor, radians/second, $\theta$ = position of motor, radians, $V_a$ = applied voltage, volts, $K_e$ = voltage constant of motor, volts/radians/second, $T_m$ = motor time constant, seconds.
For position control		
$\frac{N(s)}{V_a(s)} = \frac{1}{K_e(T_m s + 1)}$		
$\frac{\Theta(s)}{V_a(s)} = \frac{1}{K_e s(T_m s + 1)}$		
D-C generator and motor For position control		$K_g$ = generator voltage constant, volts/field ampere, $R$ = series resistance of motor and generator armature circuit, ohms, $T_f$ = generator field time constant, seconds, $E_c$ = voltage applied to generator control field, volts, $V_a$ = voltage across drive motor terminals, volts.
$\frac{\Theta(s)}{E_c(s)} = \frac{K_g/K_e R}{s(T_f s + 1)(T_m s + 1)}$		
Galvanometer		$J$ = moment of inertia of galvanometer element, pound-foot-seconds/second, $K_1$ = torque coefficient, pound-feet/ampere, $T_1$ = time constant of galvanometer coil circuit, seconds, $\theta$ = position of galvanometer element, radians, $I$ = signal current, amperes.
$\frac{\Theta(s)}{I(s)} = \frac{K_1}{J s^2(T_1 s + 1)}$		

Gyroscope

$$\frac{\Omega(s)}{I(s)} = \frac{K_2}{Js(T_2s + 1)}$$

Gyroscope



- $J$  = moment of inertia of gyroscope, pound-foot-seconds/second,
- $\Omega$  = angular velocity of gyroscope, radians/second,
- $J\Omega$  = angular momentum of gyroscope, pound-foot-second,
- $K_2$  = torque coefficient, pound-foot/ampere,
- $T_2$  = time constant of gyroscope precession coil circuit, seconds,
- $I$  = signal current, amperes.

Stabilizing networks

For rate signals (phase lead)

$$\frac{E_o(s)}{E_{in}(s)} = \frac{Ts}{Ts + 1}$$

For integral signals (phase lag)

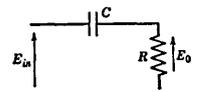
$$\frac{E_o(s)}{E_{in}(s)} = \frac{1}{Ts + 1}$$

For rate and integral (lead-lag)

$$\frac{E_o(s)}{E_{in}(s)} = \frac{T_1T_2s^2 + (T_1 + T_2)s + 1}{T_1T_2s^2 + (T_1 + T_2 + T_{12})s + 1},$$

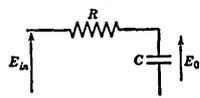
$$T_{12} \gg T_1 + T_2$$

Phase Lead



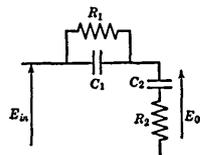
- $E_{in}$  = input voltage, volts,
- $E_o$  = output voltage, volts,
- $T = RC$ , time constant, seconds.

Phase Lag



$T = RC$ , time constant, seconds.

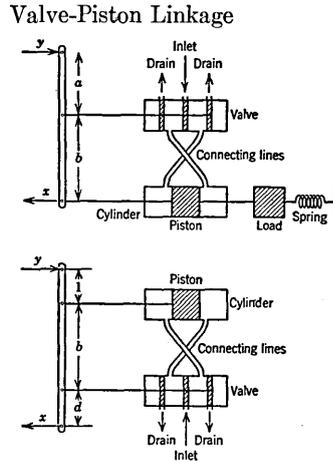
Lead-Lag



- $T_1 = R_1C_1$
- $T_2 = R_2C_2$
- $T_{12} = R_1C_2$
- } = time constants, seconds.

TABLE 5. SUMMARY OF TRANSFER FUNCTIONS OF REPRESENTATIVE HYDRAULIC ELEMENTS (Ref. 3a)

Hydraulic Elements	
Valve-piston	
Load reaction negligible	$\frac{X(s)}{Y(s)} = \frac{C_1}{s}$
Spring load dominant	$\frac{X(s)}{Y(s)} = C_2$
Valve-piston linkage	
For phase lag	$\frac{X(s)}{Y(s)} = \frac{b/a}{T_v s + 1}$
For phase lead	$\frac{X(s)}{Y(s)} = \frac{d}{1 + b} \left[ \frac{T_3 s + 1}{T_v s + 1} \right]; T_3 > T_v$



Nomenclature

$x$  = piston displacement from neutral, feet,  
 $y$  = input displacement from neutral, feet,  
 $C_1$  = piston velocity per valve displacement, second<sup>-1</sup>,  
 $C_2$  = piston travel per valve displacement.

For phase lag

$a, b$  = linkage distances, feet,  
 $T_v = \frac{a + b}{aC_1}$ , valve effective time constant, seconds.

For phase lead

$1, b, d$  = linkage distances, feet,  
 $T_v = \frac{1}{(1 + b)C}$ , valve time constant, seconds,  
 $T_3 = \frac{T_v(1 + b)(b + d)}{d}$ , lead time constant, seconds.

### Hydraulic motor

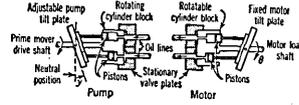
With compressibility

$$\frac{\Theta(s)}{Y(s)} = \frac{S_p/d_m}{s \left[ \frac{VJ}{Bd_m^2} s^2 + \frac{LJ}{d_m^2} s + 1 \right]}$$

With negligible compressibility

$$\frac{\Theta(s)}{Y(s)} = \frac{S_p/d_m}{s \left[ \frac{LJ}{d_m^2} s + 1 \right]}$$

### Hydraulic Motor



- $\theta$  = motor position, radians,
- $y$  = displacement of pump stroke from neutral, feet,
- $S_p$  = flow, cubic feet/second, from pump per unit displacement,  $y$ , feet,
- $d_m$  = motor displacement, cubic feet,
- $J$  = moment of inertia, pound-foot-seconds/second,
- $L$  = leakage coefficient, cubic feet/second/pound/square foot,
- $V$  = total oil volume under compression, cubic feet,
- $B$  = oil bulk modulus, pounds/square foot.

TABLE 6. ERROR DETECTORS (REF. 11)

No.	(a) Type (b) Main Application	Operation	Possible Modifications	Operating Features	Accuracy Limited by	Features Determining Energy Required to Vary Reference Quantity Measurement	Frequently Used with This Device: (a) Table 7 Amplifier (b) Table 8 Error Corrector
1	(a) D-c or a-c resistance bridge (b) Position control	Error voltage, $x$ , appears when the position of the moving arms of the potentiometers $A$ and $B$ are not matched. The power source, $E$ , is applied across both potentiometers. $A$ measures reference position as voltage and $B$ regulated position as voltage, their difference being $x$ .	Potentiometer can be wound on a helix to get more than $360^\circ$ of rotation.	$A$ and $B$ can be remote. Continuous rotation not possible.	Potentiometer winding.	Contact arm and bushing friction.	(a) 2, 3, 4, 5 (b) 1, 2, 3
2	(a) D-c tachometer bridge (b) Speed control	Error voltage, $x$ , appears when speeds of tachometers $A$ and $B$ vary. $A$ measures reference speed as a voltage and $B$ regulated speed as a voltage. The difference between these voltages is $x$ .	$A$ can be replaced by a battery as the reference.	$A$ and $B$ can be remote. Top speed limited by commutator.	Tachometer accuracy. Commutator resistance.	Brush and bearing friction.	(a) 2, 3, 4 (b) 1, 2, 3
3	(a) A-c magnetic bridge (b) Position control, particularly for gyro pickups where very small forces prevail	Error voltage, $x$ , appears when relative positions of rotor $A$ and stator $B$ do not match. Rotor $A$ measures reference position magnetically and stator $B$ regulated position magnetically. Voltage $E$ , across exciting coil, $L$ , provides energy. When rotor covers unequal areas of each exposed stator pole (unbalanced magnetic bridge) pickup coils $M$ and $N$ have unequal voltages induced. Voltage difference is $x$ .	Four poles instead of three can be used with two having exciting windings and two pickup coils connected bucking.	Limited rotation. Air gap usually small.	Machining tolerance, magnetic fringing, and voltage phase shift.	Load taken from $x$ . Bearing friction.	(a) 2, 4 (b) 1, 2, 3
4	(a) A-c synchro-system (b) Position control where continuous rotation is desired	Error voltage, $x$ , appears whenever the relative positions of the rotors of synchro-generator, $A$ , and synchro-control transformer, $B$ , are not matched. The reference position is measured by $A$ as a magnetic flux pattern which is transmitted to the synchro-control transformer through the interconnected stator windings. If the rotor of $B$ is not exactly $90^\circ$ from the transmitted flux pattern, $x$ is produced.	A dual system can be used whereby the unity synchro-system sets the approximate position and the high-speed or vernier system sets the accurate position.	Unlimited rotation. The synchro-generator and control transformer can be remote.	Machining tolerance, accuracy of winding distribution.	Distributed or non-distributed winding of control transformer rotor. Load taken from $x$ . Bearing and slip ring friction.	(a) 2, 4 (b) 1, 2, 3

- 5 (a) Frequency bridge  
(b) Frequency control

Error voltage,  $x$ , appears when reference and regulated frequencies differ. Tube channel  $A$  produces a filtered sawtooth wave that gives a d-c voltage inversely proportional to the reference frequency. Tube channel  $B$  produces a similar voltage as a measure of the regulated frequency. The difference of these d-c voltages is  $x$ .

May be used as a speed regulator if  $B$  is made an a-c tachometer.

$A$  and  $B$  can be remote. Tubes can be either gas or vacuum. A wide range of frequencies can be covered. Vacuum tubes should be used for high frequencies.

Temperature and aging effects on tube and circuit elements.

Tube input impedance.

- (a) 4  
(b) 1, 3

- 6 (a) Millivolt bridge  
(b) Temperature control

Error voltage,  $x$ , appears whenever the regulated temperature differs from the reference temperature. The regulated temperature is measured as a voltage by the thermoelectric effect of two dissimilar metals,  $B$ . The reference temperature is represented as a voltage from the battery-potentiometer source  $A$ . The difference in these voltages is  $x$ .

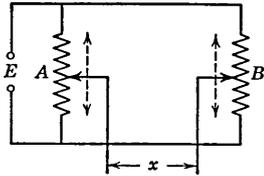
An electronic voltage source or another thermocouple can be substituted for  $A$ .

$A$  and  $B$  can be remote. A wide range of temperature can be covered.

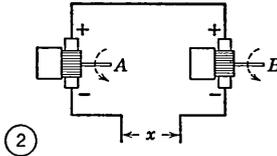
Ability to detect very low millivolt signals.

Contact arm and bushing friction. If electronic voltage source  $A$  is used, tube input impedance.

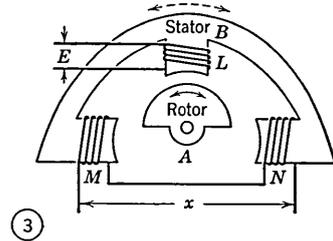
- (a) 2, 4  
(b) 1, 6



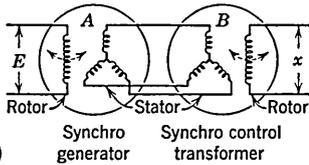
1



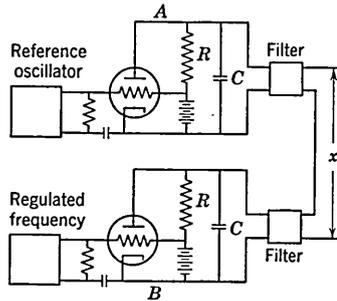
2



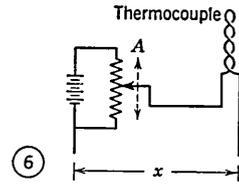
3



4



5



6

TABLE 6. ERROR DETECTORS (Continued)

No.	(a) Type (b) Main Application	Operation	Possible Modifications	Operating Features	Accuracy Limited by	Features Determining Energy Required to Vary Reference Quantity Measurement	Frequently Used with This Device: (a) Table 7 Amplifier (b) Table 8 Error Corrector
7	(a) Phototube bridge (b) Position control by intercepting a light beam	Error voltage, $x$ , appears when movable shutter is in other than desired position. Light reaching phototube, $B$ , measures shutter position. This light is measured as a voltage by the phototube current variation. A reference position of the shutter is represented by the battery-potentiometer voltage. The difference of these voltages is $x$ .	An electronic voltage source or another light source and phototube can be substituted for $A$ .	$A$ and $B$ can be remote. Glass surfaces through which light travels must be kept clean.	Continued accuracy of light source and phototube.	Contact arm and bushing friction. If electronic voltage source $A$ is used, tube input impedance.	(a) 2, 4 (b) 1
8	(a) Mechanical differential (b) Position control and speed control	Displacement $x$ appears whenever the relative reference and regulated positions change. Reference position is measured as an angle by one side of the differential $A$ and regulated position as an angle by the other side of the differential $B$ . The difference in the two positions rotates the middle member of the differential giving displacement $x$ .	Spur-gear differential.	Since $A$ and $B$ must be located together, synchro ties or their equivalent can be used to transmit remote positions to $A$ and $B$ . Continuous rotation possible with speed limited by gears.	Gearing backlash.	Power taken from $x$ . Bearing friction. Pitch of gears.	(a) 1, 6, 7, 8 (b) 1, 2, 3, 4, 5
9	(a) Beam balance (b) Voltage control, speed control, and tension control	Displacement $x$ appears whenever the variable force is different from the reference force. The variable force, $B$ , and the reference spring force, $A$ , are measured as moments. The difference in these moments produces displacement $x$ .	Any variable force other than a spring can be used.	For remote operation $B$ can be a transmitted force. $x$ movement limited. By changing springs a wide force range can be covered.	Load taken from $x$ . Bearing friction.	Magnitude of forces. Screw pitch and friction.	(a) 1, 6, 7 (b) 1, 3, 5
10	(a) Modified beam balance (b) Speed control (flyball governors)	Displacement $x$ appears when regulated speed, $\omega$ , differs from reference speed. This is represented by spring force, $A$ , about fulcrum, $O$ , the regulated speed by centrifugal force of mass, $B$ , about $O$ . Difference in moments of forces about $O$ produces displacement $x$ .	Any variable force other than a spring can be used.	A wide speed range can be covered. $x$ movement limited.	Load taken from $x$ . Friction.	Magnitude of forces. Screw pitch and friction.	(a) 1, 7, 8 (b) 1, 5

- 11 (a) Bimetal Displacement  $x$  appears whenever the surrounding temperature and the reference temperature are different. The reference temperature is represented by the position of the adjustable reference point,  $A$ . The surrounding temperature is measured by the position of the bimetal strip,  $B$ . The difference in these positions produces displacement  $x$ .
- (b) Temperature control
- 12 (a) Float Displacement  $x$  appears when regulated and reference liquid levels differ. Point  $A$  is reference. The liquid level is measured as a position by the float  $B$ . The difference produces displacement  $x$ .
- (b) Liquid level control

Bimetal can be made snap acting at some standard temperature.

Wide temperature range possible by selection of proper bimetal.

Load taken from  $x$ . Ability to measure accurately small  $x$  deflection. Time lag and hysteresis of bimetal.

Mounting of reference point.

(a) 1, 6  
(b) 1, 6

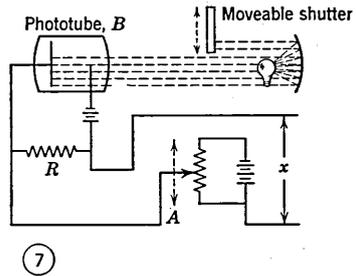
A float controlling a pulley system can be used rather than a lever.

With the proper mechanical arrangement a wide variation in liquid height can be controlled.

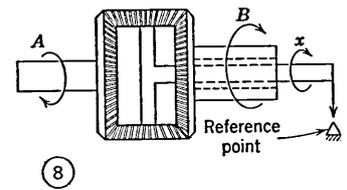
Load taken from  $x$ . Variable density of the liquid. Friction.

Mounting of reference point.

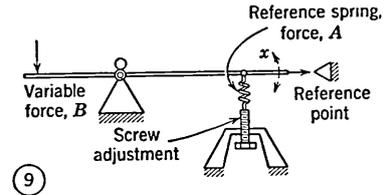
(a) 2, 7  
(b) 3, 4



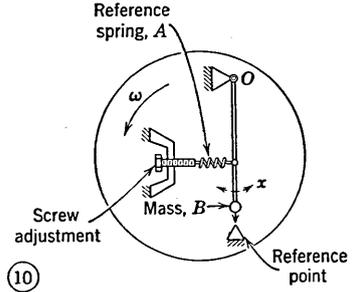
7



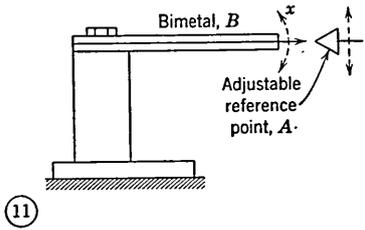
8



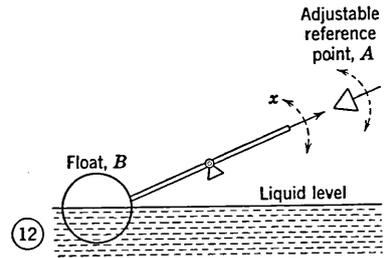
9



10



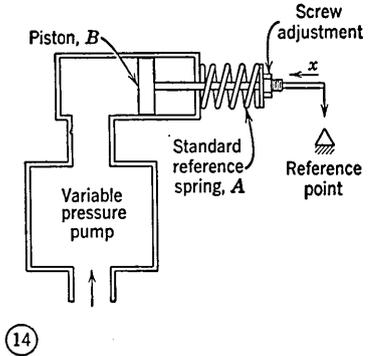
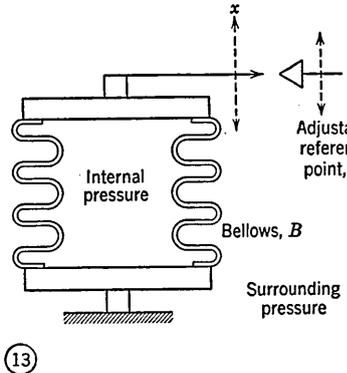
11



12

TABLE 6. ERROR DETECTORS (Continued)

No.	(a) Type (b) Main Application	Operation	Possible Modifications	Operating Features	Accuracy Limited by	Features Determining Energy Required to Vary Reference Quantity Measurement	Frequently Used with This Device: (a) Table 7 Amplifier (b) Table 8 Error Corrector
13	(a) Bellows (b) Pressure control and temperature control	Displacement $x$ appears when surrounding and reference pressures differ. Reference pressure is represented as position by adjustable point, $A$ . Surrounding pressure is measured by the bellows as a position. Difference in these positions produces displacement $x$ .	Spring can be added in addition to bellows spring.	Limited $x$ travel.	Load taken from $x$ . Hysteresis of bellows spring.	Mounting of reference point.	(a) 1, 4, 5, 6 (b) 1, 6
14	(a) Piston (b) Pressure control	Displacement $x$ appears when regulated pressure outputs of pump and reference differ. Reference pressure is a force on the piston by spring, $A$ . Regulated pressure is a force on the piston by the fluid. Their difference produces displacement $x$ .	A standard pressure source can be substituted for the spring.	$A$ and $B$ can be remote. Limited $x$ travel.	Friction. Load taken from $x$ .	Piston forces involved. Screw pitch and friction.	(a) 7, 8 (b) 5



### Methods of Evaluating Transfer Functions Experimentally

In the analysis or design of feedback controls and when equipment is available, it may be desirable to measure experimentally the actual transfer function or functions. Several techniques for obtaining this information are available. The choice of technique depends upon the complexity of the system, available instrumentation, available evaluation equipment such as high response drives, and the accuracy desired. In general the accuracy and response of the instrumentation should be several times better than the desired accuracy of the measured transfer function.

**Use of Steady-State Data.** Complex systems are more difficult to measure, and it is sometimes desirable to perform some analytical work before beginning experimental measurements. For *example*, an  $n$ th order linear system can, in general, be broken up into  $n$  first order equations. The individual equations can then be evaluated by obtaining steady-state and transient data for each. In breaking the original equation one must be careful to obtain physically significant combinations so that the steady-state and transient response can actually be measured. This technique is particularly valuable when the system has numerous inputs. For example, an  $n$ th order linear system with  $m$  inputs can be simplified to  $n$  first order linear partial differential equations. To obtain the  $m$  times  $n$  transfer functions would then require  $m$  times  $n$  pieces of steady-state data but still only  $n$  pieces of transient data. For a more detailed treatment see Ref. 4.

The assumption that the output to input relationship is the same during transients as the known steady-state relationship is called the *quasi-static assumption*. This is particularly useful for highly nonlinear phenomena such as turbojet compressor maps or aerodynamic shock systems. See Ref. 9 and Chap. 25, Sect. 2.

**Transient Data.** When suitable evaluation equipment such as sinusoidal input generators and phase measuring equipment is lacking, the system can be excited by any available time varying signal such as a unit impulse, step function, ramp, or combination of these and the output compared with analytically derived response of many systems to the same input. *Examples* of this method are given in Sect. 2, Figs. 16 to 19 and Tables 12 to 14. Other *examples* are in Refs. 5 and 6. Methods of obtaining the frequency response directly from the transient response to a step input are given in Chap. 22.

A graphical method to estimate system time constants from step response transient data requires plotting the response ( $x$ ) on semilog paper with time on the linear scale and magnitude on the log scale normalized so that the response  $x \rightarrow 0$  as  $t \rightarrow \infty$ . A system with a single time constant

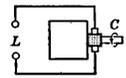
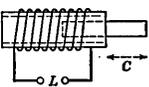
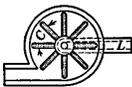
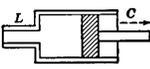
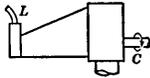
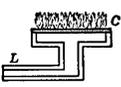
TABLE 7. POWER

No.	Schematic Representation	Type	"Gate" Element
1		Contact	Contact
2		Relay	Contact
3		Generator	Field
4		Electronic tube	Grid
5		Saturable reactor	D-c coil
6		Silverstat	Contacts
7		Valve	Valve gate
8		Throttle	Throttle valve
9		Clutch	Clutch disk

AMPLIFIERS (Ref. 11)

Typical Input Units	Typical Output Units	Approximate Power Amplification Factor	Devices Represented by Load, $L$	Power Control
Ounces	Watts	$1 \times 10^7 \times t$	Relay motor, generator field, impedance, solenoid	On-off
Watts	Watts or kilowatts	$1 \times 10^3$	Relay motor, generator field, impedance, solenoid	On-off
Watts	Watts or kilowatts	50	Motor, impedance	Continuous
Micro-watts	Watts	$1 \times 10^5$	Relay motor, generator field, impedance	Continuous
Milli-watts	Watts	$3 \times 10^2$	Generator field, impedance	Continuous
Grams	Watts	$1 \times 10^7 \times t$	Generator field, impedance	Stepped
Inch-pound	Horsepower	$1 \times 10^7 \times t$	Turret, press, heat absorption	Continuous
Inch-pound	Horsepower	$1 \times 10^7 \times t$	Propeller, vehicle, generator, mill	Continuous
Inch-pound	Horsepower	$5 \times 10^4 \times t$	Vehicle, mill	On-off

TABLE 8. ERROR CORRECTORS (Ref. 11)

No.	Schematic Representation	Type	Input Energy	Output Energy	Approximate Output Power Range
1		Electric motor	Electric	Mechanical rotation	$1 \times 10^{-2}$ to $4 \times 10^4$ hp
2		Solenoid	Electric	Mechanical translation	$1 \times 10^{-3}$ to 15 hp
3		Hydraulic motor	Hydraulic	Mechanical rotation	$1 \times 10^{-2}$ to $11 \times 10^4$ hp
4		Piston	Hydraulic	Mechanical translation	$1 \times 10^{-3}$ to $1 \times 10^3$ hp
5		Steam or gas prime mover	Heat or chemical (fuel)	Mechanical rotation	5 to $1.65 \times 10^5$ hp
6		Burner	Chemical (fuel)	Heat	$1 \times 10^2$ to $1:5 \times 10^8$ Btu/hr

$T$  will give a linear response and the slope of this line is the reciprocal of the time constant, i.e.,

$$\frac{\Delta \log x}{\Delta t} = \frac{k}{T}.$$

*Note.* For  $\log_e$ ,  $k = 1$ , for  $\log_{10}$ ,  $k = 0.434$ . A plot of the response of a system with two time constants ( $T_1 > T_2$ ) will start flat at  $\log x_0$  ( $x_0 =$  initial value) and then asymptotically approach the slope of the reciprocal of the larger time constant ( $k/T_1$ ). Drawing the asymptote will give an intersection

$$\log \left[ x_0 \frac{T_1}{T_1 + T_2} \right]$$

on the response axis. See Ref. 6.

**Frequency Response.** The most generally useful method is to excite the system or element under test with a sinusoidal signal. The *frequency response* is obtained by making a comparison of the amplitude and phase relations of the input and output over the frequency range of interest.

The *phase* and *amplitude relations* can be obtained in a number of ways, e.g., from (a) direct oscillograph or recorder readings of the variables, (b) Lissajous patterns on a long persistence oscilloscope, and (c) special test equipment that gives a direct reading of the phase and amplitude ratios.

An analytical expression approximating the transfer function can be obtained by *curve matching* techniques. Sufficiently good results are often obtained by a simple trial and error approximation of the frequency response obtained by the use of the straight line asymptote defined in Chap. 21. Straight lines with slopes which are multiples of 20 db per decade are first drawn so as to approximate the experimental data. The exact frequency response corresponding to the estimated straight line response is then calculated by use of the graphs of Chap. 21. The agreement between the calculated and measured response is checked and the process is repeated if necessary. With a little experience one or two iterations are usually sufficient. The intersections of the straight lines are the poles and zeros of the transfer function. More elaborate approximation methods are available if needed. See Refs. 6 and 8.

**Correlation Technique.** The autocorrelation function of white noise is an impulse. Therefore the cross correlation of the input and output of the system is simply the impulse response of the system when the input is white noise. An experimental setup similar to Fig. 10 can therefore be

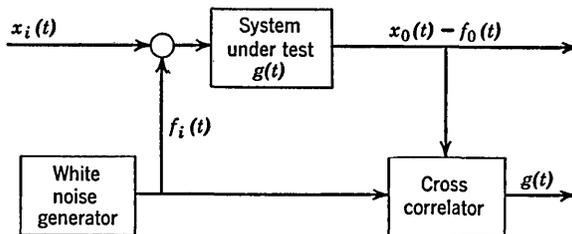


FIG. 10. Test configuration to obtain system response by correlation technique.

used to evaluate the transfer function of one element or system. Because the cross correlation filters all signals not correlated with the input white noise, the technique has the potential advantage of allowing the normal system operation to continue while the test is being conducted. See Ref. 8. The practical difficulties of mechanizing a satisfactory cross-correlator has limited the usefulness of this method.

## 2. CLASSICAL METHODS OF ANALYSIS

### System Equations

In Methods of System Analysis in Sect. 1, the correct description of the system or component dynamic performance by differential equations was stated to be basic to all methods of analysis and synthesis.

**General Linear Differential Equations.** A common type of system equation, the general linear integro-differential equation with constant coefficients may be written in terms of the input  $x(t)$  and the driving function  $y(t)$  (Ref. 13) as

$$(22) \quad a_0 \frac{d^n x(t)}{dt^n} + a_1 \frac{d^{n-1} x(t)}{dt^{n-1}} + \cdots + a_{n-1} \frac{dx(t)}{dt} + a_n x(t) \\ + a_{n+1} \int x(t) dt + \cdots + a_{n+q} \int^q x(t) dt^q = y(t).$$

As a class, the homogeneous equation resulting from reducing the right-hand side of the equation to zero has as its general solution a linear combination of solutions of the exponential form  $e^{p_n t}$ , where  $p_n$  may be real or complex.

**Characteristic Equation.** The operator  $p = d/dt$  together with  $1/p = \int dt$  may be substituted into the reduced homogeneous equation. The resulting operational equation may be handled by the rules of algebra as explained in Chap. 8, Sect. 1. Factoring out the operational part of this equation yields the characteristic equation (Ref. 13):

$$(23) \quad a_0 p^{n+q} + a_1 p^{n+q-1} + \cdots + \\ a_{n-1} p^{1+q} + a_n p^q + a_{n+1} p^{q-1} + \cdots + a_{n+q} = 0.$$

### General Solution to Linear Differential Equations

The complementary solution to eq. (22) is

$$(24) \quad x_t = A_{n+q} e^{(p_{n+q})t} + A_{n+q-1} e^{(p_{n+q-1})t} + \cdots + A_1 e^{(p_1)t},$$

where  $p_n$  are the roots of the characteristic eq. (23). The complete solution is (Ref. 13)

$$(25) \quad x(t) = x_t + x_s,$$

where  $x_s$  is the particular solution to eq. (22). The particular solution is obtained by substituting an assumed solution and solving for the coefficients. (See Part A, General Mathematics.)

**Absolute Stability Defined.** (See also Chap. 21.) The *stability* of a system may be broadly defined as that property which insures that it will remain in operating equilibrium through normal conditions (Ref. 14). A system is said to be on the verge of stability when it is hunting, that is, subject to sustained oscillations; if the oscillations grow, the system is unstable; if they decay, the system is stable. Nonoscillatory instability is also possible, such as the exponential growth of a system variable in response to a disturbance. Tables 14 and 15 give *examples* of stable and unstable performance and the dependence of stability upon the nature of the roots of the characteristic equation (or exponents of the complementary solution). When system gain is increased to provide desired accuracy, instability is frequently encountered. This is the situation which is attacked with equalization (or stabilization) methods designed to provide a margin of stability without compromising system accuracy. A margin of stability is nearly always desired from the hunting condition. It is implied that the system is linear or may be linearized in the neighborhood of the operating point for the purpose of analyzing stability (for linearization of nonlinear systems, see Linearization, Chap. 25, Sect. 2).

**EXAMPLE.** *Second Order System (Motor Synchronizing on a Fixed Signal).* In Fig. 11 a motor drives a load to which it is coupled directly from an

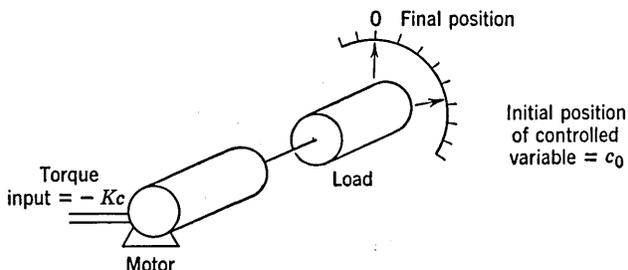


FIG. 11. Motor driving load from an initial position  $c_0$  to correspondence at position 0. Combined inertia =  $J$  lb-ft-sec<sup>2</sup>; damping =  $D$  lb-ft/rad/sec; stiffness =  $K$  lb-ft/rad (Ref. 3a).

initial rest position  $c_0$  to correspondence at position 0 (Ref. 13). The input to the motor produces a torque that is proportional to the difference between the controlled load position and the reference input position. Thus motor torque equals  $-Kc$  since the desired final position is zero.

Present in the load and the motor are mechanical friction and electrical damping torques both of which are proportional to the motor speed; friction and damping torque equal  $D(dc/dt)$ . Static friction forces are negligible. There is also a torque due to the combined inertia  $J$  and this torque equals  $J(d^2c/dt^2)$ .

The complete torque equation can now be written as

$$(26) \quad J \frac{d^2c}{dt^2} + D \frac{dc}{dt} + Kc = 0.$$

The steady-state displacement is zero, that is, the corresponding position, so that the transient response is the entire motion. By writing the characteristic equation as

$$(27) \quad p^2 + \frac{D}{J}p + \frac{K}{J} = 0,$$

a further modification will be made in the interest of obtaining a simpler form of the solution. Let

$$(28) \quad \sqrt{\frac{K}{J}} = \omega_0 = \text{undamped natural frequency,}$$

and

$$(29) \quad \frac{D}{2\sqrt{KJ}} = \zeta = \text{damping factor.}$$

If these substitutions are made, the characteristic equation may now be written (in nondimensional form) as

$$(30) \quad p^2 + 2\zeta\omega_0p + \omega_0^2 = 0,$$

in which the two roots are

$$(31) \quad p_1 = -[\zeta - \sqrt{\zeta^2 - 1}]\omega_0,$$

$$(32) \quad p_2 = -[\zeta + \sqrt{\zeta^2 - 1}]\omega_0.$$

The effect of  $\zeta$  upon the form of the transient solution of a second order system is treated in the next section.

### Use of Laplace Transform

The work involved in using the classical approach to the solution of linear differential equations may be simplified to a routine process through the use of the Laplace transform and its inverse, which uses the same approach to obtain both transient and steady-state solutions (Refs. 15, 16). The Laplace transform has the advantage of handling initial conditions and discontinuous inputs directly. For a complete presentation of this method see Ref. 17.

The *Laplace transform* is defined as

$$\mathcal{L}[f(t)] = F(s) = \int_0^{\infty} f(t)e^{-st} dt.$$

The *inverse Laplace transform* is defined as

$$\mathcal{L}^{-1}[F(s)] = f(t) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} F(s)e^{ts} ds, \quad t \geq 0.$$

In these definitions  $s$  is the complex operator  $\sigma + j\omega$ . The *abscissa of absolute convergence*, denoted by  $\sigma_0$ , is at  $\sigma_0 > 0$ .

The Laplace transform and inverse Laplace transform are referred to as a *transform pair*.

**Laplace Transform Applied to Feedback Control System.** A closed loop linear control system may be represented in terms of the complex operator  $s$  by the eqs. (33), (35), and (36) as follows:

$$(33) \quad C(s) = G(s)E(s),$$

where  $C(s)$  = transform of the controlled variable,  $c(t)$ ,

$E(s)$  = transform of the actuating error,  $e(t)$ ,

$G(s)$  = transform of the transfer function of the forward control elements and may be given the factored form:

$$(34) \quad G(s) = \frac{K(s - s_a)(s - s_b) \cdots}{s^n(s - s_1)(s - s_2) \cdots},$$

$$(35) \quad E(s) = R(s) - B(s),$$

where  $R(s)$  = transform of the reference input,  $r(t)$ ,

$B(s)$  = transform of the feedback,  $b(t)$ .

$$(36) \quad B(s) = H(s)C(s),$$

where  $H(s)$  is the transform of the transfer function of the feedback elements and may be similar in form to that given in eq. (34).

The block diagram for the above system of equations is given in Fig. 25b. The transform of the closed loop transfer function (see Fig. 25c for the block diagram) for this control system is

$$(37) \quad \frac{C(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)}.$$

Expressing  $C(s)$  in terms of polynomials in  $s$ ,

$$(38) \quad C(s) = \frac{N(s)}{D(s)} = \frac{a_v s^v + a_{v-1} s^{v-1} + \cdots + a_1 s + a_0}{s^q + b_{q-1} s^{q-1} + \cdots + b_1 s + b_0},$$

where, because of the nature of the functions  $G(s)$ ,  $H(s)$ , and  $R(s)$ , in general  $q \geq v$ .

TABLE 9. SOME USEFUL LAPLACE TRANSFORM PAIRS (Ref. 15)

No.	$F(s)$	$f(t), \quad 0 \leq t$
1	$\frac{1}{s}$	1 or $u(t)$ , unit step at $t = 0$
2	$\frac{1}{s} e^{-as}$	$u(t - a)$
3	$\frac{1}{s} (e^{-as} - e^{-bs})$	$u(t - a) - u(t - b), \quad a < b$
4	1	$\lim_{a \rightarrow 0} \frac{u(t) - u(t - a)}{a}$ , unit impulse at $t = 0$
5	$\frac{1}{s^2}$	$t$ , unit ramp at $t = 0$
6	$\frac{1}{s^n}$	$\frac{1}{(n - 1)!} t^{n-1}$
7	$\frac{1}{s + \alpha}$	$e^{-\alpha t}$
8	$\frac{1}{(s + \alpha)^2}$	$te^{-\alpha t}$
9	$\frac{1}{(s + \alpha)^n}$	$\frac{1}{(n - 1)!} t^{n-1} e^{-\alpha t}$
10	$\frac{1}{s(s + \alpha)}$	$\frac{1}{\alpha} (1 - e^{-\alpha t})$
11	$\frac{1}{s^2(s + \alpha)}$	$\frac{e^{-\alpha t} + \alpha t - 1}{\alpha^2}$
12	$\frac{1}{s(s + \alpha)^2}$	$\frac{1}{\alpha^2} (1 - e^{-\alpha t} - \alpha t e^{-\alpha t})$
13	$\frac{1}{(s + \alpha)(s + \gamma)}$	$\frac{e^{-\alpha t} - e^{-\gamma t}}{\gamma - \alpha}$
14	$\frac{1}{s(s + \alpha)(s + \gamma)}$	$\frac{1}{\alpha\gamma} + \frac{\gamma e^{-\alpha t} - \alpha e^{-\gamma t}}{\alpha\gamma(\alpha - \gamma)}$
15	$\frac{1}{s^2 + \beta^2}$	$\frac{1}{\beta} \sin \beta t$
16	$\frac{s}{s^2 + \beta^2}$	$\cos \beta t$
17	$\frac{a_1 s + a_0}{s^2 + \beta^2}$	$A \cos(\beta t + \psi)$ $A \triangleq [a_1^2 + a_0^2/\beta^2]^{1/2}$ $\psi \triangleq \tan^{-1} \frac{-a_0/\beta}{a_1}$

TABLE 9. SOME USEFUL LAPLACE TRANSFORM PAIRS (Ref. 15) (Continued)

No.	$F(s)$	$f(t), \quad 0 \leq t$
18	$\frac{1}{(s^2 + \beta^2)^2}$	$\frac{1}{2\beta^3}(\sin \beta t - \beta t \cos \beta t)$
19	$\frac{s}{(s^2 + \beta^2)^2}$	$\frac{t}{2\beta} \sin \beta t$
20	$\frac{s^2}{(s^2 + \beta^2)^2}$	$\frac{1}{2\beta}(\sin \beta t + \beta t \cos \beta t)$
21	$\frac{s^2 - \beta^2}{(s^2 + \beta^2)^2}$	$t \cos \beta t$
22	$\frac{1}{(s + \alpha)^2 + \beta^2}$	$\frac{1}{\beta} e^{-\alpha t} \sin \beta t$
23	$\frac{s + \alpha}{(s + \alpha)^2 + \beta^2}$	$e^{-\alpha t} \cos \beta t$
24	$\frac{a_1 s + a_0}{(s + \alpha)^2 + \beta^2}$	$A e^{-\alpha t} \cos(\beta t + \psi)$
		$A \triangleq [a_1^2 + (a_1 \alpha - a_0)^2 / \beta^2]^{1/2}$
		$\psi \triangleq \tan^{-1} \frac{(a_1 \alpha - a_0) / \beta}{a_1}$
25	$\frac{1}{s^2 - \beta^2}$	$\frac{1}{\beta} \sinh \beta t$
26	$\frac{s}{s^2 - \beta^2}$	$\cosh \beta t$
27	$sF(s) - f(0+)$	$\frac{df(t)}{dt}$
28	$s^2 F(s) - sf(0+) - \frac{df(t)}{dt} (0+)$	$\frac{d^2 f(t)}{dt^2}$
29	$\frac{F(s)}{s} + \frac{f^{(-1)}(0+)}{s}$	$\int f(t) dt$
30	$\frac{F(s)}{s^2} + \frac{f^{(-1)}(0+)}{s^2} + \frac{f^{(-2)}(0+)}{s}$	$\int \left[ \int f(t) dt \right] dt$
31	$aF(s)$	$af(t)$
32	$F_1(s) \pm F_2(s)$	$f_1(t) \pm f_2(t)$
33	$aF(as)$	$f\left(\frac{t}{a}\right)$
34	$F(s + a)$	$e^{-at} f(t)$
35	$F(s - a)$	$e^{at} f(t)$
36	$e^{\mp as} F(s)$	$f(t \mp a), \quad a > 0$

where  $f(t - a) = 0, \quad 0 < t < a$

$f(t + a) = 0, \quad -a < t < 0$

The *inverse transformation*  $\mathcal{L}^{-1}[C(s)]$ , which is the solution  $c(t)$ , may be obtained for the simpler cases by direct reference to transform Tables 9, 10, and 11 or Tables 1 and 2, Chap. 9 (see also Ref. 15). However, in general it is necessary to express  $C(s)$  as a sum of partial fractions with constant coefficients, whose inverse transforms may then be readily obtained from the table of transforms and summed to yield the complete solution,  $c(t)$ .

TABLE 10. LAPLACE TRANSFORM PAIRS

FOR FIRST ORDER SYSTEM  $X(s) = \frac{1}{Ts + 1} F(s)$

Input as $f(t)$	Input as $F(s)$	Response Function as $X(s)$	Time Response as $x(t)$
Impulse, $A \left[ \lim_{a \rightarrow 0} \frac{u(t) - u(t-a)}{a} \right]$ ,			
whence $A = \int_{-\infty}^{\infty} f(t) dt$	$A$	$\frac{A}{Ts + 1}$	$\frac{A}{T} e^{-t/T}$
Step, $Au(t)$	$\frac{A}{s}$	$\frac{A}{s(Ts + 1)}$	$A(1 - e^{-t/T})$
Ramp, $At$	$\frac{A}{s^2}$	$\frac{A}{s^2(Ts + 1)}$	$At - AT(1 - e^{-t/T})$

**Partial Fraction Expansions of the Laplace Transform.** For expansion into partial fractions the denominator of the rational proper fraction  $C(s)$  is first factored, yielding (Ref. 16):

$$(39) \quad C(s) = \frac{N(s)}{D(s)} = \frac{N(s)}{(s - s_1)(s - s_2) \cdots (s - s_k) \cdots (s - s_n)}$$

Expansion into partial fractions with first order poles only (Ref. 13)

$$(40) \quad \frac{N(s)}{D(s)} = \frac{C_1}{s - s_1} + \frac{C_2}{s - s_2} + \cdots + \frac{C_k}{s - s_k} + \cdots + \frac{C_n}{s - s_n}$$

leads to the solution (see transform pair No. 0.11 of Ref. 15)

$$(41) \quad c(t) = \sum_{k=1}^n \frac{N(s_k)}{D'(s_k)} e^{s_k t},$$

where

$$(42) \quad D'(s_k) \triangleq \left[ \frac{d}{ds} D(s) \right]_{s=s_k} \quad (\triangleq \text{ means equal by definition}).$$

TABLE 11. TIME RESPONSES OF SECOND ORDER SYSTEM  $C(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0s + \omega_0^2} R(s)$

Type of Input $\mathcal{L}[r(t)] = R(s)$	$\zeta < 1$	$\zeta = 1$	$\zeta > 1$
Unit impulse, $\mathcal{L}\left[\lim_{a \rightarrow 0} \frac{u(t) - u(t-a)}{a}\right] = 1$	$\frac{\beta_0^2}{\beta} e^{-\alpha t} \sin \beta t$	$\alpha^2 t e^{-\alpha t}$	$\frac{\alpha\gamma}{\gamma - \alpha} (e^{-\alpha t} - e^{-\gamma t})$
Unit step, $\mathcal{L}[u(t)] = \frac{1}{s}$	$1 + \frac{\beta_0}{\beta} e^{-\alpha t} \sin\left(\beta t - \tan^{-1} \frac{\beta}{-\alpha}\right)$	$1 - (1 + \alpha t)e^{-\alpha t}$	$1 + \frac{\gamma e^{-\alpha t} - \alpha e^{-\gamma t}}{\alpha - \gamma}$
Unit ramp, $\mathcal{L}[t] = \frac{1}{s^2}$	$t - \frac{2\alpha}{\beta_0^2} + \frac{1}{\beta} e^{-\alpha t} \sin\left(\beta t - 2 \tan^{-1} \frac{\beta}{-\alpha}\right)$	$t - \frac{2}{\alpha} + t e^{-\alpha t} + \frac{2}{\alpha} e^{-\alpha t}$	$t - \frac{\gamma + \alpha}{\alpha\gamma} + \frac{\gamma^2 e^{-\alpha t} - \alpha^2 e^{-\gamma t}}{\alpha\gamma(\gamma - \alpha)}$
Specific form of $C(s)/R(s)$	$\frac{\beta_0^2}{(s + \alpha)^2 + \beta^2}$ where $\alpha = \zeta\omega_0$ and $\beta_0^2 = \alpha^2 + \beta^2 = \omega_0^2$	$\frac{\alpha^2}{(s + \alpha)^2}$ where $\alpha = \omega_0$	$\frac{\alpha\gamma}{(s + \alpha)(s + \gamma)}$ where $\alpha + \gamma = 2\zeta\omega_0$ and $\alpha\gamma = \omega_0^2$

Note. Laplace transform,  $C(s)$ , of each time response in this table is simply the product of the transform of the input,  $R(s)$ , by the system transform  $C(s)/R(s)$  in the table.

The general case of expansion into partial fractions with higher order poles yields the solution (see transform pair No. 0.21 of Ref. 15)

$$(43) \quad c(t) = \sum_{k=1}^n \sum_{i=1}^{m_k} \frac{K_{ki}}{(m_k - i)!} t^{m_k - i} e^{s_k t}, \quad m_1 + m_2 + \cdots + m_n = q,$$

where

$$(44) \quad K_{ki} \triangleq \frac{1}{(i-1)!} \left[ \frac{d^{i-1}}{ds^{i-1}} \frac{(s - s_k)^{m_k} N(s)}{D(s)} \right]_{s=s_k},$$

and

$$(45) \quad D(s) \triangleq (s - s_1)^{m_1} (s - s_2)^{m_2} \cdots (s - s_k)^{m_k} \cdots (s - s_n)^{m_n}.$$

**Order of System Responses as Seen from Partial Fraction Expansions.** Any complex linear system can be represented as a combination of first and second order systems. This may be seen from the partial fraction expansion of a system response function such as that of eq. (38) into partial fractions

$$(46) \quad C(s) = \frac{N(s)}{D(s)} \\ = \frac{C_1}{s - s_1} + \frac{C_2}{s - s_2} + \cdots + \frac{C_k}{s - s_k} + \frac{a_1 s + a_0}{s^2 + 2\zeta\omega_0 s + \omega_0^2} + \cdots,$$

where two conjugate complex first order poles have been combined into a single term.

The response of the complex system can, therefore, be considered as the sum of the responses of first and second order systems. The responses of first and second order systems are thus of considerable importance and are given for various inputs in the following.

**First Order System Responses.** A first order system is characterized by a single energy storage. An *example* of a first order system is the simple hydraulic servo of Table 5 for phase lag. The system equation is thus (using  $T$  for  $T_v$ ),

$$(47) \quad T \frac{dx(t)}{dt} + x(t) = (b/a)y(t) = f(t).$$

The transform of the equation may be written:

$$(48) \quad [X(s)] = \left[ \frac{1}{Ts + 1} \right] [F(s)]$$

which is of the form (see Ref. 16)

$$(\text{Response function}) = (\text{System function})(\text{Excitation function}).$$

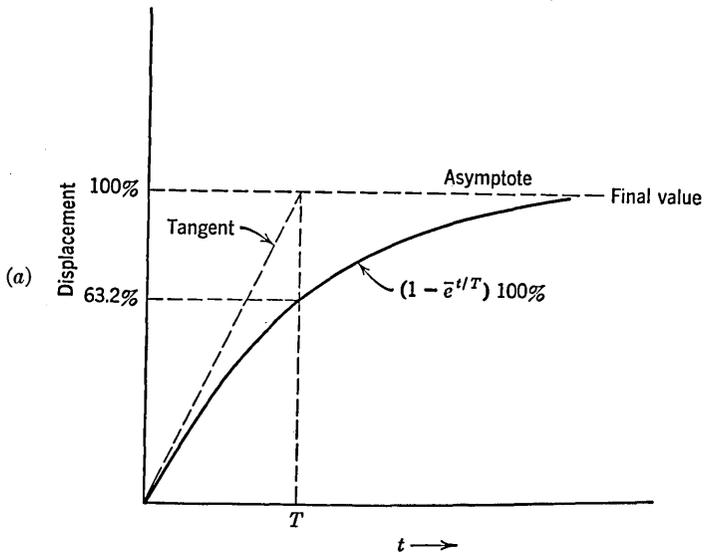


FIG. 12a. Response of first order system to step showing time constant relationships.

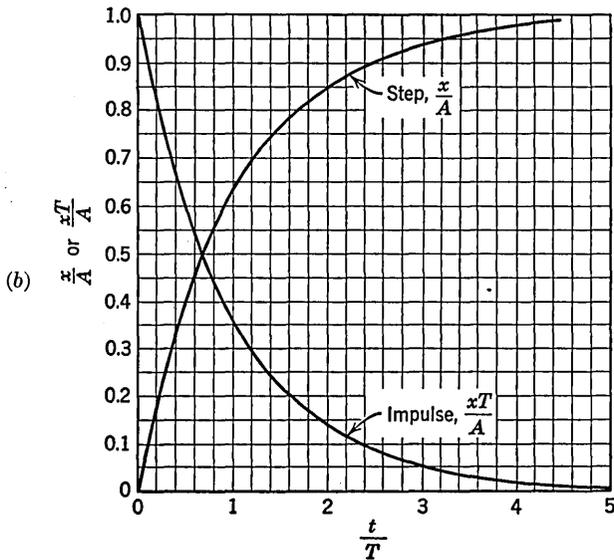


FIG. 12b. Step and impulse response of first order system:

$$T \frac{dx}{dt} + x = f(t).$$

$$\text{Impulse: } \frac{xT}{A} = e^{-t/T}, \quad A = \int_{-\infty}^{\infty} (\text{impulse function}) dt.$$

$$\text{Step: } \frac{x}{A} = 1 - e^{-t/T}, \quad A = \text{magnitude of step.}$$

The characteristic equation is

$$(49) \quad Ts + 1 = 0,$$

and the transient solution is

$$(50) \quad x_t = A_1 e^{-t/T}.$$

The performance can be characterized by the quantity  $T$ , called the time constant of the system (see Ref. 13). Physically, the time constant is the time to complete  $1 - e^{-1} = 63.2\%$  of the change after either a step

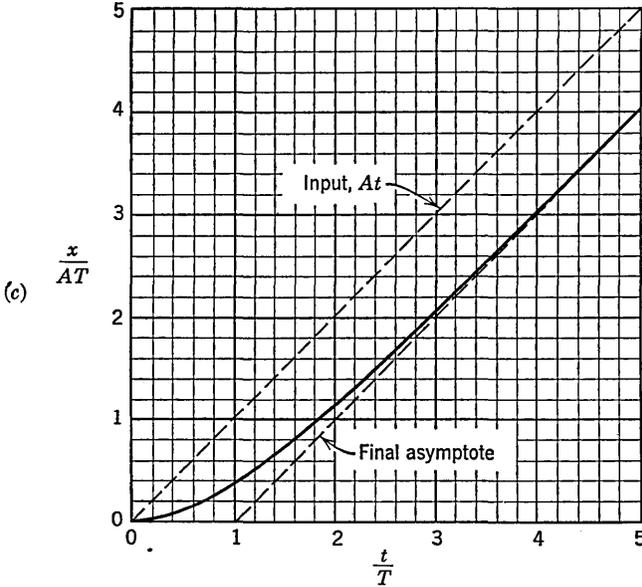


FIG. 12c. Ramp response of first order system:

$$T \frac{dx}{dt} + x = f(t) = At,$$

$$\frac{x}{AT} = \frac{t}{T} - (1 - e^{-t/T}).$$

or impulse input. Also it is the time given by the intersection of the tangent to the transient at  $t = 0$  with the asymptote to the final value when a step or impulse is applied at  $t = 0$  (see Fig. 12a).

Table 10 lists three types of input  $f(t)$ , the corresponding excitation function  $F(s)$ , the response function  $X(s)$  for the system function  $1/(Ts + 1)$  and the inverse transform  $x(t)$  of the response function,  $X(s)$  and  $x(t)$  forming a Laplace transform pair. In Fig. 12b are plotted the step and impulse response of a first order system obtained from the solutions

appearing in Table 10. In Fig. 12c is plotted the ramp response of a first order system from the solution appearing in the same table.

**Second Order System Responses.** The solutions for unit impulse, step and ramp inputs to the second order system of eq. (26) generalized by setting the right-hand side equal to  $Kr(t)$ , namely

$$(51) \quad J \frac{d^2c}{dt^2} + D \frac{dc}{dt} + Kc = Kr(t)$$

are illustrated respectively in Figs. 13, 14, and 15a, b, c. The transform

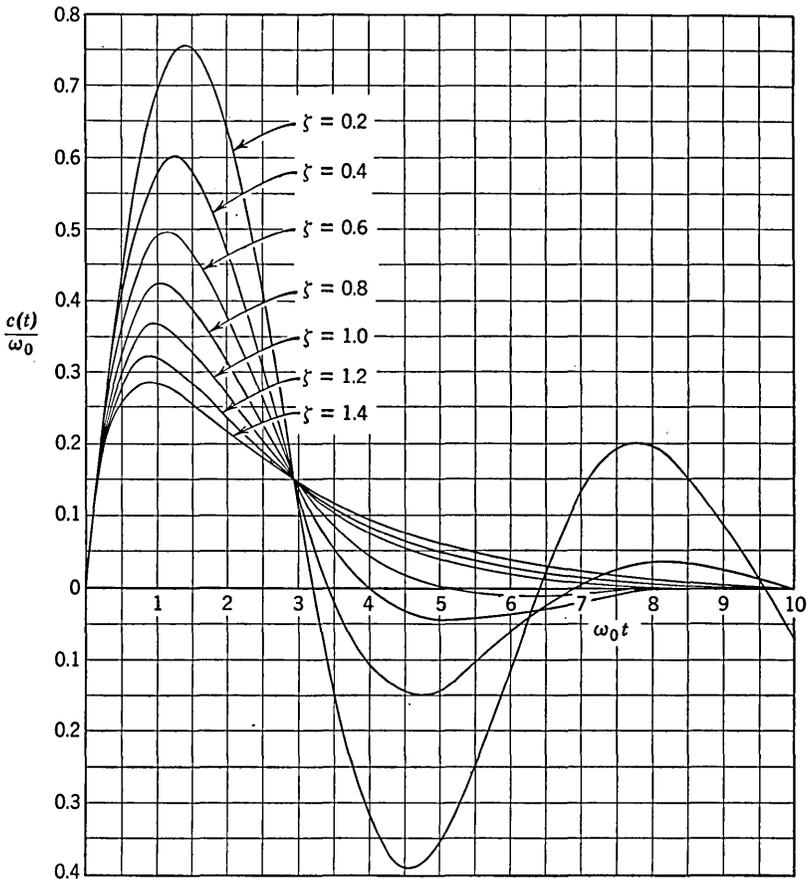


FIG. 13. Response of second order system:

$$C(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0s + \omega_0^2} R(s)$$

to unit impulse in  $r(t)$  for various values of  $\zeta$ .

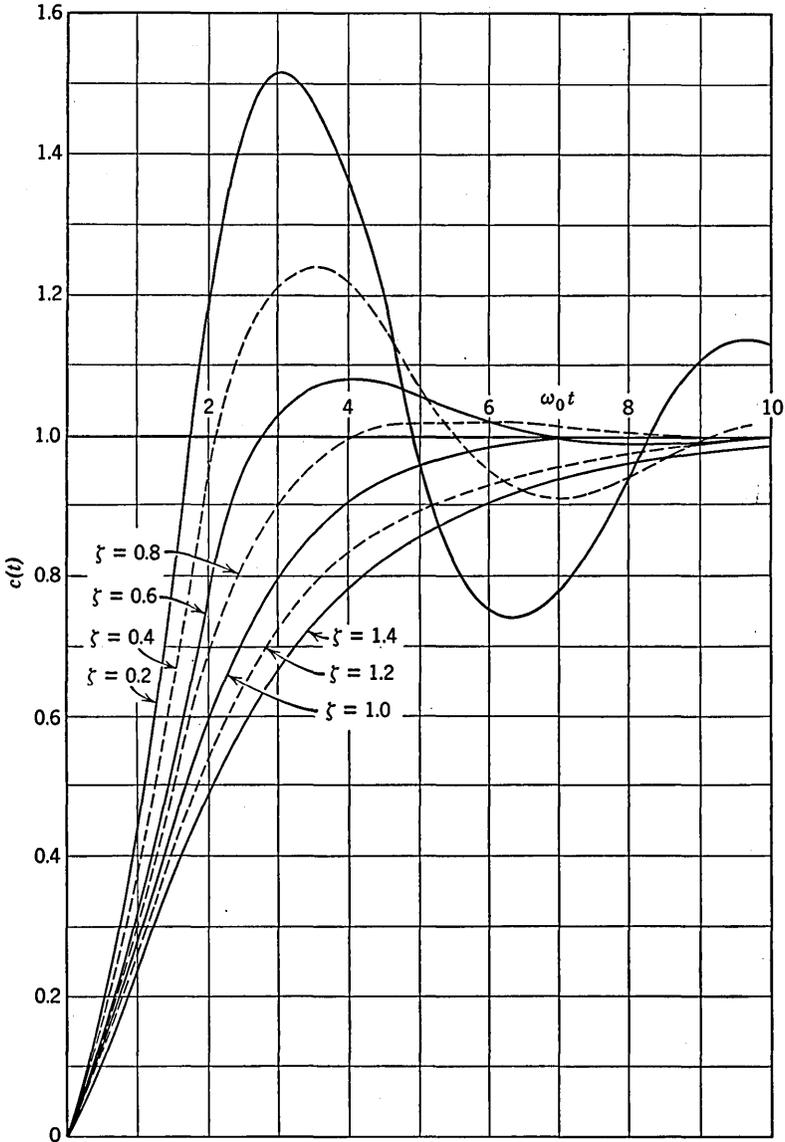


FIG. 14. Response of second order system:

$$C(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2} R(s)$$

to unit step in  $r(t)$  for various values of  $\zeta$ .

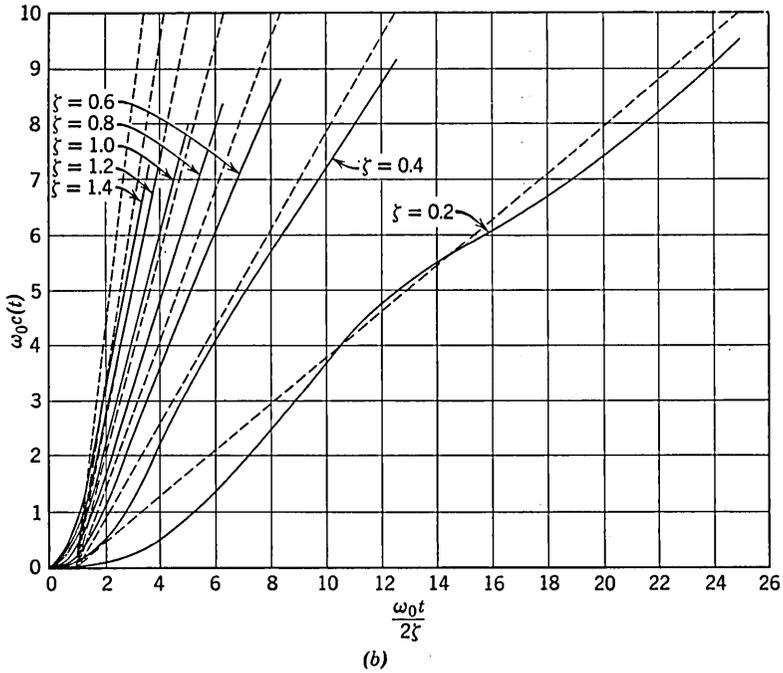
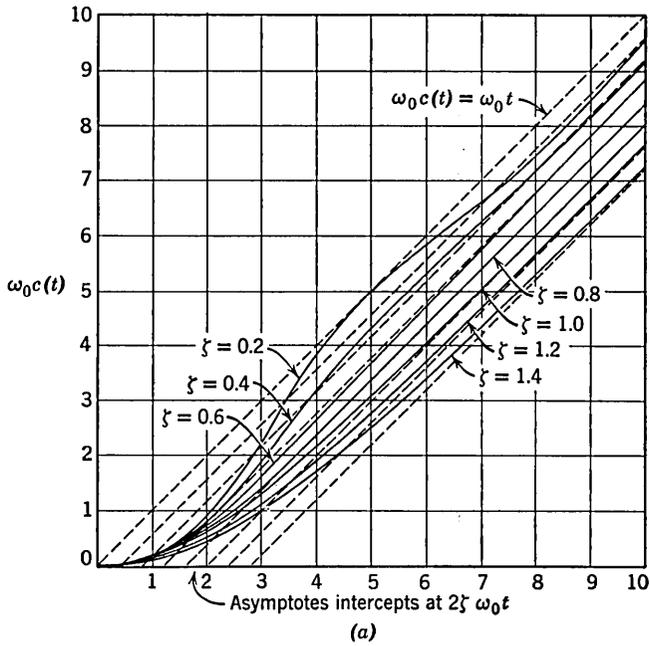


FIG. 15a, b. Response of second order system:  $C(s) = [(\omega_0^2)/(s^2 + 2\zeta\omega_0s + \omega_0^2)]R(s)$  to unit ramp  $r(t) = t^2$  for various values of  $\zeta$ .

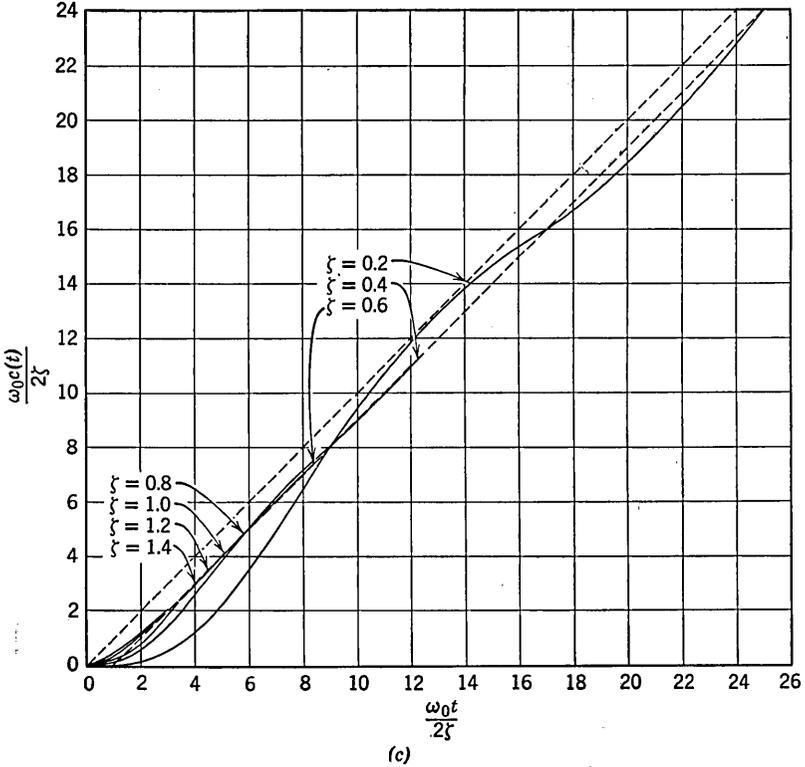


FIG. 15c. Response of second order system:

$$C(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0s + \omega_0^2} R(s)$$

to unit ramp  $r(t) = t$  for various values of  $\zeta$ .

of this equation may be written in a nondimensional form (as for the first order system of eq. 47).

$$(\text{Response function}) = (\text{System function})(\text{Excitation function})$$

(see Ref. 16), which in this case is

$$(52) \quad C(s) = \left[ \frac{\omega_0^2}{s^2 + 2\zeta\omega_0s + \omega_0^2} \right] [R(s)].$$

The time response may then be obtained by looking up the inverse transform pair in a table of Laplace transforms (see Ref. 15). The form of the response will be oscillatory, critically damped, or overdamped as

the damping factor  $\zeta < 1, = 1,$  or  $> 1$ . Table 11 is a chart of the time responses illustrated in Figs. 13 to 15 (see Refs. 6, 13, 18, and 19).

Tables 12 and 13 and Figs. 16 to 19, from Ref. 20, are for the determination of equation coefficients and system parameters for second order systems. Table 14 illustrates time responses. Table 15 treats stability.

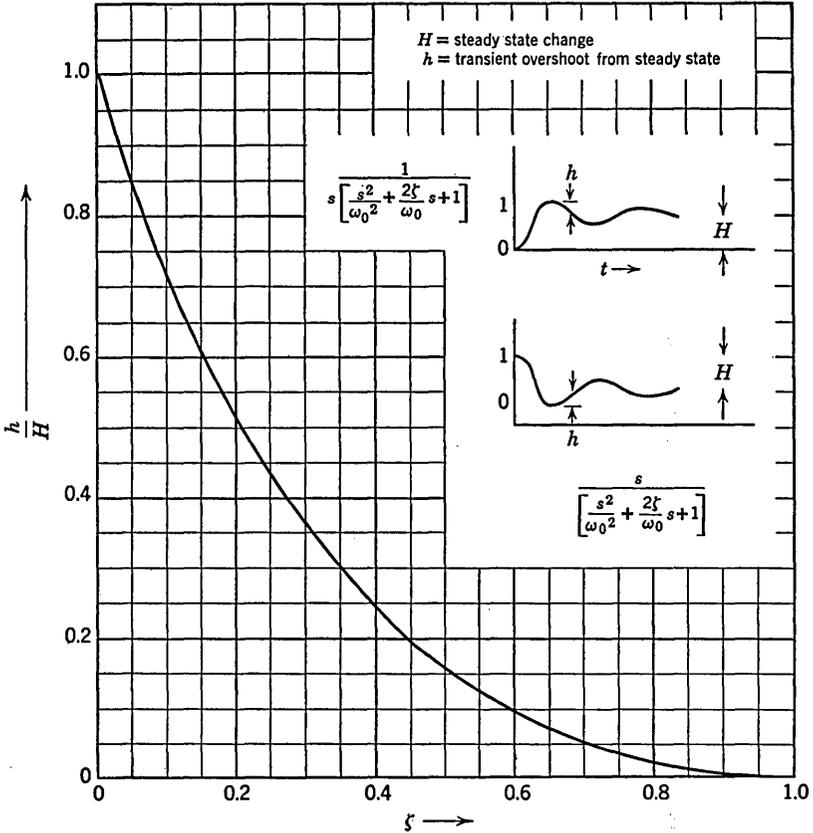


FIG. 16. Determination of equation coefficients for second order systems from response curves (Ref. 20).

TABLE 12. RELATIONSHIP AMONG SYSTEM PARAMETERS AND EQUATION COEFFICIENTS FOR SECOND ORDER SYSTEM (Ref. 20)

$$a_2[(d^2x)/(dt^2)] + a_1[(dx)/(dt)] + a_0x = y(t)$$

Parameter	Symbol	Definition in Terms of Equation Coefficients	Equivalent Expressions
Damping ratio	$\zeta$	$\frac{a_1}{2\sqrt{a_0a_2}}$	$\frac{1}{\omega_0 T_c}, \frac{T_0}{2\pi T_c}, \frac{T_1 + T_2}{2\sqrt{T_1 T_2}}, \frac{\nu + 1}{2\sqrt{\nu}}$
Undamped angular natural frequency	$\omega_0$	$\sqrt{\frac{a_0}{a_2}}$	$\frac{\omega}{\sqrt{1 - \zeta^2}}, 2\pi f_0, \frac{2\pi}{T_0}, \frac{1}{\zeta T_c}, \frac{1}{\sqrt{T_1 T_2}}$
Undamped natural frequency	$f_0$	$\frac{1}{2\pi\sqrt{\frac{a_2}{a_0}}}$	$\frac{\omega_0}{2\pi}, \frac{\omega}{2\pi\sqrt{1 - \zeta^2}}, \frac{1}{T_0}, \frac{1}{2\pi\zeta T_c}$
Undamped natural period	$T_0$	$2\pi\sqrt{\frac{a_2}{a_0}}$	$\frac{2\pi}{\omega_0}, T\sqrt{1 - \zeta^2}, 2\pi\zeta T_c, \frac{1}{f_0}$
Angular natural frequency	$\omega$	$\sqrt{\frac{a_0}{a_2} - \left(\frac{a_1}{2a_2}\right)^2}$	$\omega_0\sqrt{1 - \zeta^2}, 2\pi f, \frac{2\pi}{T}, \frac{\sqrt{1 - \zeta^2}}{\zeta T_c}, \frac{2\pi\sqrt{1 - \zeta^2}}{T_0}$
Natural frequency	$f$	$\frac{1}{2\pi}\sqrt{\frac{a_0}{a_2} - \left(\frac{a_1}{2a_2}\right)^2}$	$\frac{\omega_0\sqrt{1 - \zeta^2}}{2\pi}, \frac{\omega}{2\pi}, \frac{1}{T}, \frac{\sqrt{1 - \zeta^2}}{2\pi\zeta T_c}$

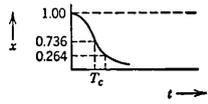
Natural period	$T$	$\frac{2\pi}{\sqrt{\frac{a_0}{a_2} - \left(\frac{a_1}{2a_2}\right)^2}}$	$\frac{2\pi}{\omega}, \frac{2\pi}{\omega_0\sqrt{1-\zeta^2}}, \frac{2\pi\zeta T_c}{\sqrt{1-\zeta^2}}, \frac{1}{f}, \frac{T_0}{\sqrt{1-\zeta^2}}$
Critical time constant	$T_c$	$\frac{2a_2}{a_1}$	$\frac{1}{\zeta\omega_0}, \frac{T_0}{2\pi\zeta}, \frac{T\sqrt{1-\zeta^2}}{2\pi\zeta}, \frac{2T_1T_2}{T_1+T_2}, \frac{2T_1}{\nu+1}, \frac{T_t}{2\zeta^2}$
Large time constant ( $\zeta > 1$ )	$T_1$	$\frac{1}{\frac{a_1}{2a_2} - \sqrt{\left(\frac{a_1}{2a_2}\right)^2 - \frac{a_0}{a_2}}}$	$\frac{1}{\omega_0[\zeta - \sqrt{\zeta^2 - 1}]}, \frac{\sqrt{\nu}}{\omega_0}, \frac{T_c\sqrt{\nu}}{2\pi}, \sqrt{\nu}\zeta T_c, \nu T_2, \frac{\nu T_t}{\nu+1}$
Small time constant ( $\zeta > 1$ )	$T_2$	$\frac{1}{\frac{a_1}{2a_2} + \sqrt{\left(\frac{a_1}{2a_2}\right)^2 - \frac{a_0}{a_2}}}$	$\frac{1}{\omega_0[\zeta + \sqrt{\zeta^2 - 1}]}, \frac{1}{\sqrt{\nu}\omega_0}, \frac{T_0}{2\pi\sqrt{\nu}}, \frac{\zeta T_c}{\sqrt{\nu}}, \frac{T_1}{\nu}$
Overcritical time constant	$T_t$	$\frac{a_1}{a_0}$	$\frac{2\zeta}{\omega_0}, T_1 + T_2, \frac{\nu+1}{\nu} T_1, 2\zeta^2 T_c, \frac{2\zeta}{\sqrt{\nu}} T_c$
Time parameter ratio ( $\zeta > 1$ )	$\nu$	$\frac{\frac{a_1}{2a_2} + \sqrt{\left(\frac{a_1}{2a_2}\right)^2 - \frac{a_0}{a_2}}}{\frac{a_1}{2a_2} - \sqrt{\left(\frac{a_1}{2a_2}\right)^2 - \frac{a_0}{a_2}}}$	$\frac{\zeta + \sqrt{\zeta^2 - 1}}{\zeta - \sqrt{\zeta^2 - 1}}, \frac{T_1}{T_2}$

TABLE 13. DETERMINATION OF EQUATION COEFFICIENTS FOR SECOND ORDER SYSTEMS FROM RESPONSE CURVES (Ref. 20)

Transfer Function	Type of Response	Response Curve	Equation Parameters Used	Method Used to Find Equation Parameters	Equation Coefficients in Terms of $a_0$ and Equation Parameters	
					$a_1$	$a_2$
$\frac{s}{\left[\left(\frac{s}{\omega_0}\right)^2 + \frac{2\zeta}{\omega_0}s + 1\right]}$	Oscillatory $0 < \zeta < 0.5$		$\zeta, T$	Measure $T$ , $x_0, x_1, x_2, x_3, \dots$	$\frac{a_0 T'}{\pi} \zeta \sqrt{1 - \zeta^2}$	$\frac{a_0 T'^2}{4\pi^2} (1 - \zeta)$
$\frac{1}{\left[\left(\frac{s}{\omega_0}\right)^2 + \frac{2\zeta}{\omega_0}s + 1\right]}$				Form ratios $\frac{x_1}{x_0}, \frac{x_2}{x_0}, \frac{x_3}{x_0}, \frac{x_2}{x_1}, \frac{x_3}{x_1}, \frac{x_3}{x_2}, \dots$		
$\frac{1}{s \left[\left(\frac{s}{\omega_0}\right)^2 + \frac{2\zeta}{\omega_0}s + 1\right]}$				Find $\zeta$ from Fig. 17		
$\frac{s}{\left[\left(\frac{s}{\omega_0}\right)^2 + \frac{2\zeta}{\omega_0}s + 1\right]}$	Near critically aperiodic $0.5 < \zeta < 2.0$		$\zeta, \omega_0$	Measure $t_1, t_2, t_3$ Form ratios $\frac{t_2}{t_1}, \frac{t_3}{t_1}, \frac{t_3 - t_2}{t_2 - t_1}$	$\frac{2a_0\zeta}{\omega_0}$	$\frac{a_0}{\omega_0^2}$
$\frac{1}{s \left[\left(\frac{s}{\omega_0}\right)^2 + \frac{2\zeta}{\omega_0}s + 1\right]}$				Find $\zeta, \omega_0 t_1, \omega_0 t_2, \omega_0 t_3$ from Fig. 18 Compute value of $\zeta = \zeta_{AV}$ and $\omega_0 = \omega_{0AV}$		

$$\frac{s}{\left[ \left( \frac{s}{\omega_0} \right)^2 + \frac{2\zeta}{\omega_0} s + 1 \right]}$$

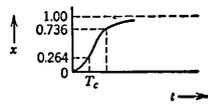
Critically aperiodic  
 $\zeta = 1.0$



$T_c$  Measure  $T_c$  on Fig. 18 for  $\zeta = 1.0$ ,  
 $T_c = t_1 = t_2 - t_1 = t_3 - t_2$

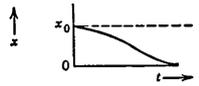
$$2a_0T_c$$

$$s \left[ \left( \frac{s}{\omega_0} \right) + \frac{2\zeta}{\omega_0} s + 1 \right]$$



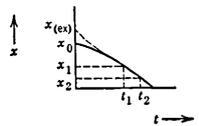
$$\frac{s}{\left[ \left( \frac{s}{\omega_0} \right)^2 + \frac{2\zeta}{\omega_0} s + 1 \right]}$$

Nonoscillatory  
 $\zeta > 1.0$



$\nu_1, T_1$  Plot response curve on semilog paper. Extrapolate straight line portion of plot to  $t = 0$ . Measure  $\ln x_1$  and  $\ln x_2$  at  $t_1$  and  $t_2$  respectively. Measure  $x_0$  and  $x_{(ex)}$ . Compute  $T_1$  from

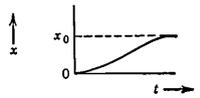
$$\frac{a_0T_1(\nu + 1)}{\nu}$$



Semilog plot of response curve

$$T_1 = \frac{t_2 - t_1}{\ln x_1 - \ln x_2}$$

$$s \left[ \left( \frac{s}{\omega_0} \right)^2 + \frac{2\zeta}{\omega_0} s + 1 \right]$$



Invert response curve to agree with above plot, then plot on semilog paper.

Compute  $\nu$  from

$$\nu = \frac{x_{(ex)}}{x_{(ex)} - x_0}$$

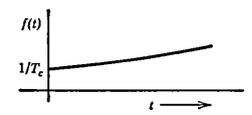
TABLE 14. TIME RESPONSES OF SOME COMMON TRANSIENT MODES (Ref. 20)

	$F(s)$	$f(t)$	Time Response
Step function position	① $\frac{1}{s}$	1	
Step function velocity	② $\frac{1}{s^2}$	$t$	
Step function acceleration	③ $\frac{1}{s^3}$	$\frac{1}{2}t^2$	
First order lag converging	④ $\frac{1}{T_c s + 1}$	$\frac{1}{T_c} e^{-t/T_c}$	
	⑤ $\frac{1}{s(T_c s + 1)}$	$1 - e^{-t/T_c}$	

First order lag  
diverging

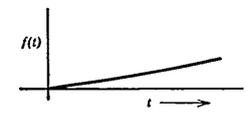
⑥ 
$$\frac{1}{-T_c s + 1}$$

$$\frac{1}{T_c} e^{t/T_c}$$



⑦ 
$$\frac{1}{s(-T_c s + 1)}$$

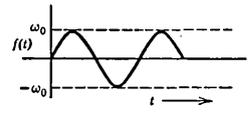
$$-1 + e^{t/T_c}$$



Undamped  
second order  
 $\zeta = 0$

⑧ 
$$\frac{1}{(s^2/\omega_0^2) + 1}, \quad \zeta = 0$$

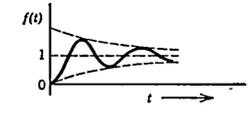
$$\omega_0 \sin \omega_0 t$$



Converging  
second order  
 $0 < \zeta \leq 1$

⑨ 
$$\frac{1}{s \left( \frac{s^2}{\omega_0^2} + 2 \frac{\zeta}{\omega_0} s + 1 \right)}$$

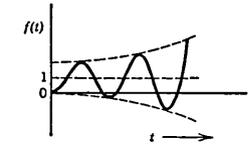
$$1 - \frac{e^{-\zeta \omega_0 t}}{\sqrt{1 - \zeta^2}} \sin \left( \omega_0 \sqrt{1 - \zeta^2} t + \tan^{-1} \frac{\sqrt{1 - \zeta^2}}{-\zeta} \right)$$



Diverging  
second order  
 $\zeta < 0$   
 $\zeta' = |\zeta|$

⑩ 
$$\frac{1}{s \left( \frac{s^2}{\omega_0^2} + 2 \frac{\zeta'}{\omega_0} s + 1 \right)}$$
  
$$= \frac{1}{s \left( \frac{s^2}{\omega_0^2} - 2 \frac{\zeta'}{\omega_0} s + 1 \right)}$$

$$1 - \frac{e^{\zeta' \omega_0 t}}{\sqrt{1 - \zeta'^2}} \sin \left( \omega_0 \sqrt{1 - \zeta'^2} t + \tan^{-1} \frac{\sqrt{1 - \zeta'^2}}{\zeta'} \right)$$



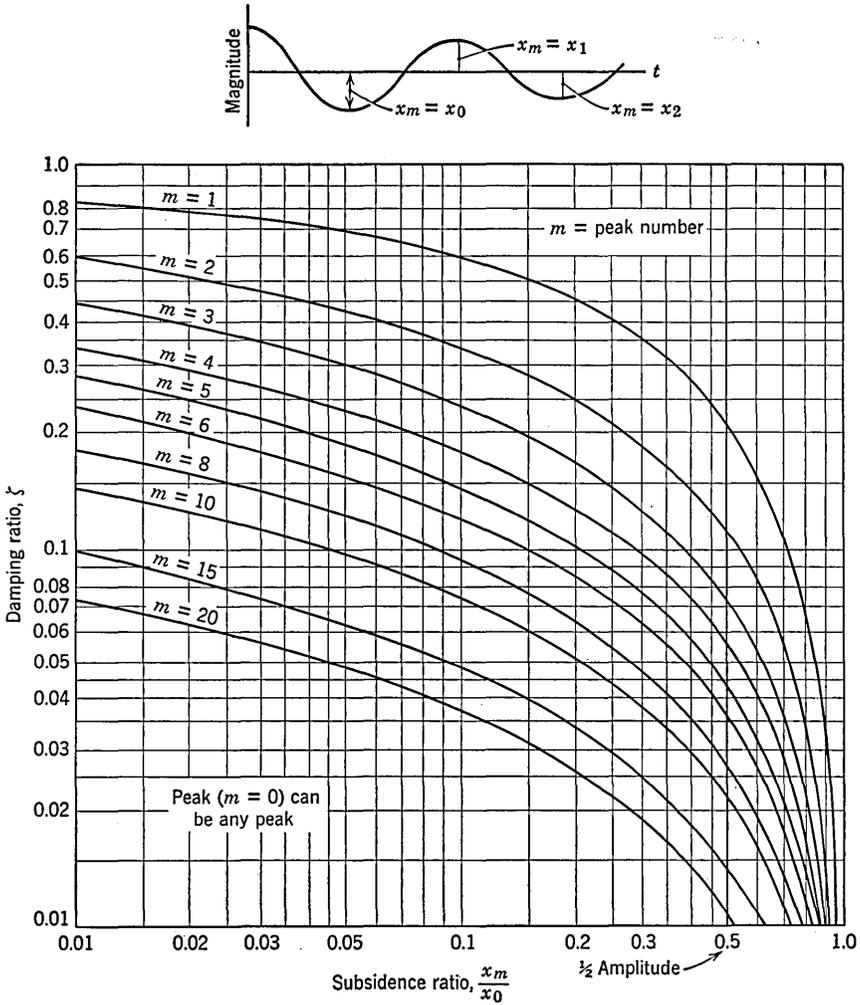


Fig. 17. Damping ratio of oscillatory transients as a function of subsidence ratio for second order system (Ref. 20) of the type:

$$\frac{s}{-\frac{s^2}{\omega_0^2} + \frac{2\zeta}{\omega_0}s + 1}, \quad \frac{1}{-\frac{s^2}{\omega_0^2} + \frac{2\zeta}{\omega_0}s + 1}, \quad \frac{1}{s \left[ \frac{s^2}{\omega_0^2} + \frac{2\zeta}{\omega_0}s + 1 \right]}$$

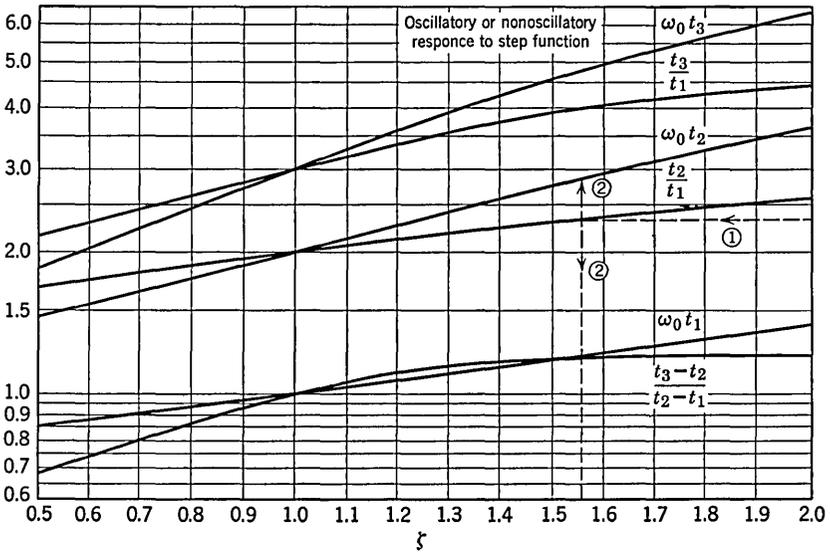
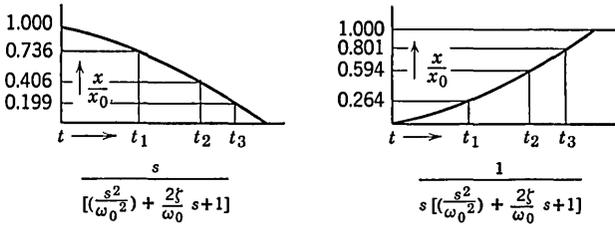
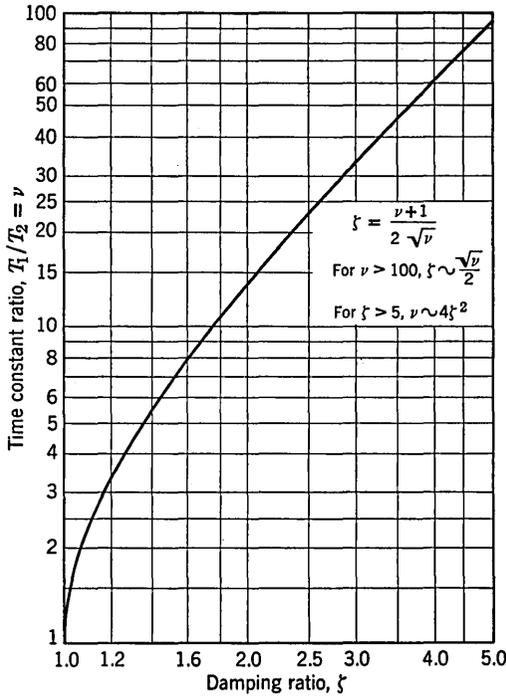


FIG. 18. Chart for determining  $\zeta$  and  $\omega_0$  for second order system when  $\zeta$  is close to 1 (Ref. 20). (The calculations on which these curves are based were made by Prof. John D. Trimmer of the University of Tennessee.)



$$\frac{s}{T_1 T_2 s^2 + (T_1 + T_2)s + 1} = \frac{s}{\frac{T_1^2}{\nu} s^2 + \frac{T_1}{\nu} (\nu + 1)s + 1}$$

$$\frac{1}{T_1 T_2 s^2 + (T_1 + T_2)s + 1} = \frac{1}{\frac{T_1^2}{\nu} s^2 + \frac{T_1}{\nu} (\nu + 1)s + 1}$$

$$\frac{1}{s[T_1 T_2 s^2 + (T_1 + T_2)s + 1]} = \frac{1}{s \left[ \frac{T_1^2}{\nu} s^2 + \frac{T_1}{\nu} (\nu + 1)s + 1 \right]}$$

FIG. 19. Time constant ratio  $T_1/T_2$  as a function of damping ratio for overdamped second order system (Ref. 20).

TABLE 15. STABILITY AS A FUNCTION OF THE NATURE OF THE ROOTS OF THE CHARACTERISTIC EQUATION (Ref. 20)

Type of Stability of System	Nature of Roots of Characteristic Equation (or Exponents of Complementary Solution)	No. of Example of Performance Given in Table 14	
		Nonoscillatory	Oscillatory
Stable	All roots have negative real parts.	4	
Stable	A single zero root; all other roots, if any, have negative real parts.	1	9
Verge of stability; underdamped oscillatory time response	Conjugate imaginary roots all different, in addition to roots for stable systems above, if any.	5	8
Unstable	Roots with positive real parts, in addition to other types of roots, if any.	6	10
Unstable	Repeated zero or conjugate imaginary roots, in addition to other types of roots, if any.	7	
		2	
		3	

**Application of Convolution Integral**

A convenient method of calculating the time response of a system to any arbitrary input makes use of the *convolution integral* (see Ref. 21), which may be written as

$$(53) \quad c(t) = \int_{-\infty}^t f(\tau)g(t - \tau) d\tau,$$

where  $c(t)$  is the time response,  $f(t)$  is the input, and  $g(t)$  is the weighting function or characteristic time response to a unit impulse (see Weighting Function in Chap. 9). To evaluate this equation the arbitrary input is approximated by a series of impulses as shown in Fig. 20. If the im-

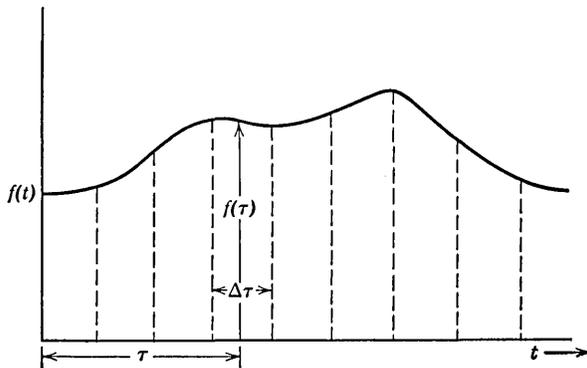


FIG. 20. Approximation of a function,  $f(t)$ , by a series of impulses.

pulsive response,  $g(t)$ , is known, the sum of these responses to the impulses approximating the input signal constitutes the total time response, as illustrated by Figs. 1.47 and 1.48 of Ref. 8. Of course, a theoretical impulse function has zero width in time; in the practical case, if its width is much smaller than the response time of the system being considered, the results obtained will be valid. The quantity  $f(\tau)$  is the average height of the rectangular approximation of an impulse;  $\Delta\tau$  is the width; and  $\tau$  is the time to the center of the rectangle as illustrated in Fig. 20.

The value of the time response at time  $t_1$  may be expressed as

$$(54) \quad c(t_1) = \sum_{\tau=\tau_1, \tau_2, \dots, t_1} f(\tau) \Delta\tau g(t_1 - \tau).$$

This indicates that  $c(t_1)$  is the sum of responses to impulse inputs, all evaluated at  $t_1$ .

This same method may be used with the transfer function of the system, since it is simply the Laplace transform of the weighting function.

### Steady-State Solution of System Equations

Although the complete solution of a linear differential equation for a system subjected to some driving function contains both transient and steady-state portions, the steady-state part can be obtained independently of the transient.

**Sinusoidal Driving Functions.** The general form of the steady-state response of linear systems to sinusoidal excitation is sinusoidal and of the same frequency as the driving function. An *example* of the steady-state response  $c_s(t)$  of the second order system of eq. (51) to a sinusoidal input is given in Fig. 21.

When steady-state excitation with sinusoidal driving forces is considered, the Laplace transform is intimately related to the impedance concept. For the Laplace transform it will be found that  $s$  may be replaced by  $j\omega$  to obtain the steady-state response to a sinusoidal driving function (see Ref. 22). In Table 16 are given typical terms of an integro-differential equation showing use of the operator  $j\omega$  to obtain the electrical and "motional" impedances of analogous electrical and mechanical forms.

The justification for this substitution of  $j\omega$  for  $s$  is given in Ref. 22. Application of this technique to the differential eq. (52) yields:

$$(55) \quad C(j\omega) = \frac{\omega_0^2}{(j\omega)^2 + 2\zeta\omega_0(j\omega) + \omega_0^2} R(j\omega).$$

**Complex Plane Plot.** The steady-state response of a system as a function of frequency is very useful in servomechanism and regulator design

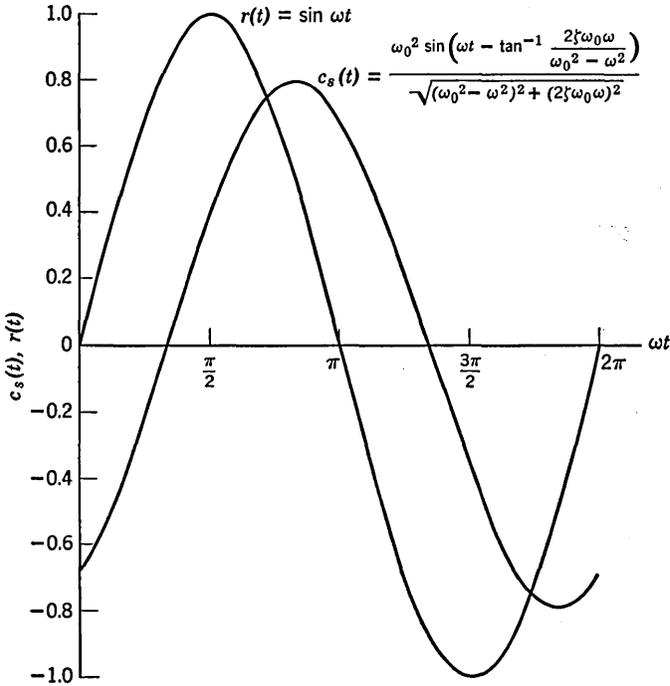


FIG. 21. Example of steady-state response,  $c_s(t)$  to unit amplitude sine wave input,  $v(t)$ , for second order system:

$$C(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0s + \omega_0^2} \text{ for } \zeta = 1 \text{ and } \omega = 0.5\omega_0.$$

TABLE 16. SUMMARY OF EQUATION TERMS AND COMPLEX QUANTITIES (Ref. 3a)

Physical System	Derivative Form	Transform Form	Complex Form	Complex Impedance
Electrical	$L \frac{di(t)}{dt}$	$LsI(s)$	$j\omega LI(j\omega)$	$j\omega L = jX_L$
	$Ri(t)$	$RI(s)$	$RI(j\omega)$	$R$
	$\frac{1}{C} \int i(t) dt$	$\frac{1}{C} \frac{I(s)}{s}$	$\frac{I(j\omega)}{j\omega C}$	$\frac{-j}{\omega C} = -jX_C$
Mechanical	$M \frac{dv(t)}{dt}$	$MsV(s)$	$j\omega MV(j\omega)$	$j\omega M$
	$Dv(t)$	$DV(s)$	$DV(j\omega)$	$D$
	$K \int v(t) dt$	$\frac{KV(s)}{s}$	$\frac{K}{j\omega} V(j\omega)$	$-j \frac{K}{\omega}$

(Ref. 23). Use is made of complex plane diagrams in which the magnitude and the angle of the output to input ratio are shown by a single line on the complex plane as in Fig. 22. The complex output-input ratio  $C/R$  is obtained by substituting  $j\omega$  for  $s$  in the transform eq. (52).

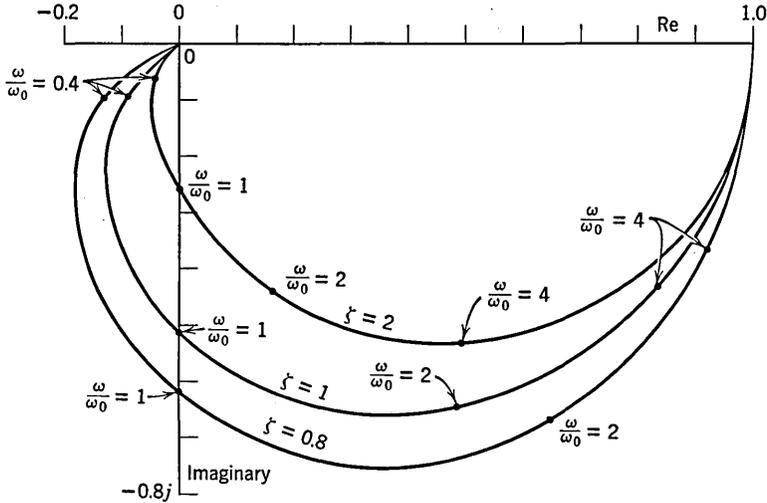


FIG. 22. Complex plane plot of

$$\frac{C}{R} = \frac{\omega_0^2}{(j\omega)^2 + 2\zeta\omega_0(j\omega) + \omega_0^2}$$

for control system of Fig. 11.

**Logarithmic Plots.** Instead of plotting vector loci of the transfer function as in Fig. 22, the contours can be plotted to a logarithmic scale (see Refs. 24, 25, and 26). To exploit certain manipulative advantages, the attenuation and phase angle graphs are made separately. The attenuation is plotted in decibels, or  $20 \log_{10} |\text{atten.}|$  versus the  $\log_{10} \omega$ ; the phase angle is also plotted versus  $\log_{10} \omega$ . In Fig. 23 the complex transfer function of eq. (55) has its attenuation and phase angle plotted against the  $\log_{10}(\omega/\omega_0)$ , giving a nondimensional chart for the frequency response of second order systems over a range of values of damping factor  $\zeta$ .

### 3. BLOCK DIAGRAMS

#### Definition of Terms

A block diagram is a simplified method of presenting the interconnections of significant variables. It displays the functional relationships rather

than the physical and thus gives a clear insight to the problem. The physical system and interrelationship determine the block diagram arrangement, each block is a logical step in the flow or signal process. Block diagrams are built up by algebraic combinations of individual blocks where each block is a transfer function. An example is shown by Fig. 24 where the transfer function of the controlled system is  $G_3 = C/M_5$ . Most block diagrams only show the desired inputs and outputs; however, in many physical systems there are loading and regulating effects. These effects must be considered and can be handled as separate input effects.

The recommended nomenclature (Ref. 29) for symbols in the block diagram is illustrated by Fig. 24 where:

- $V$  = desired value,
- $R$  = reference input,
- $E$  = actuating error,
- $M$  = manipulated variable,
- $U$  = disturbance function,
- $C$  = controlled variable,
- $Q$  = indirectly controlled variable,
- $B$  = feedback.

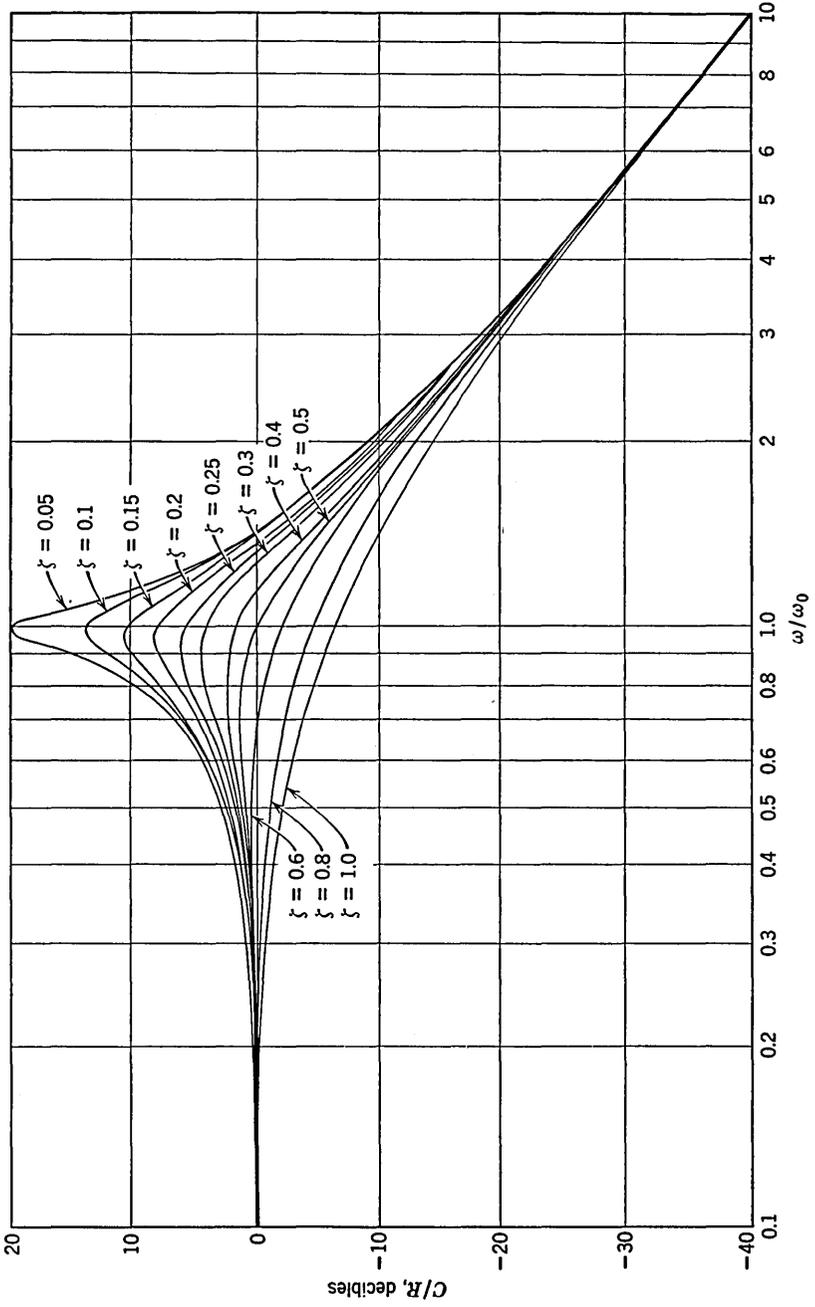
The symbols used, such as  $C$ , may be in Laplace, operational, or sinusoidal form and can be indicated as  $C(s)$ ,  $C(p)$ , or  $C(j\omega)$ . Lower-case letters are used to indicate time functions ( $r, v, e, m, u, c, b$ ). Generally the parentheses ( $s$ ), ( $p$ ), or ( $j\omega$ ) are dropped unless a particular form of representation is required.

The transfer functions are labeled as follows:  $A$  for reference input,  $G$  for forward elements, i.e., from error to output,  $N$  for disturbance input,  $Z$  for indirectly controlled system, and  $H$  for feedback. Numerical subscripts are used to identify individual elements. References 29 to 31 are the standards of the American Institute of Electrical Engineers and the Institute of Radio Engineers. The important point is that a consistent system be used.

**Construction and Signal Flow**

As illustrated by Fig. 24 the arrows connecting the blocks indicate the unidirectional signal flow. The circular junction point with appropriate plus or minus signs is used to indicate summing or differencing points respectively, i.e.,





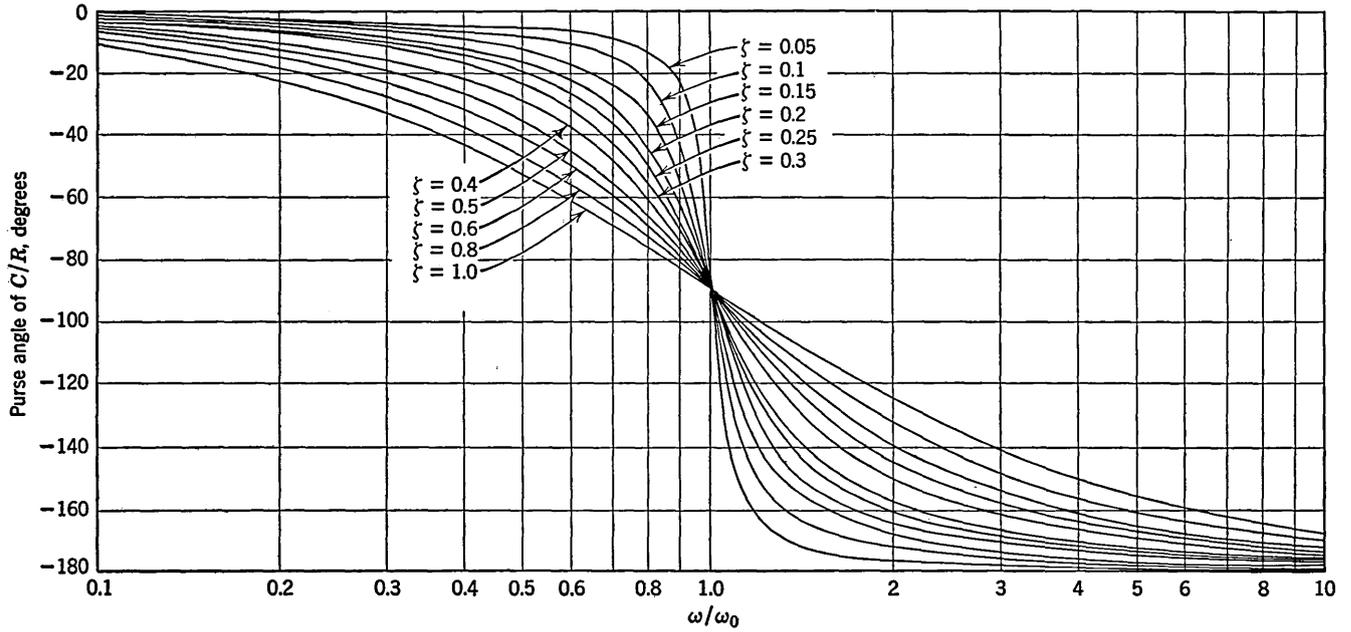


FIG. 23. Magnitude and phase shift of  $C/R$  versus frequency ratio  $\omega/\omega_0$  for various values of  $\zeta$ .

*Note.* For  $\zeta > 1$  plots are simply those for two unequal time lags (Ref. 3a).

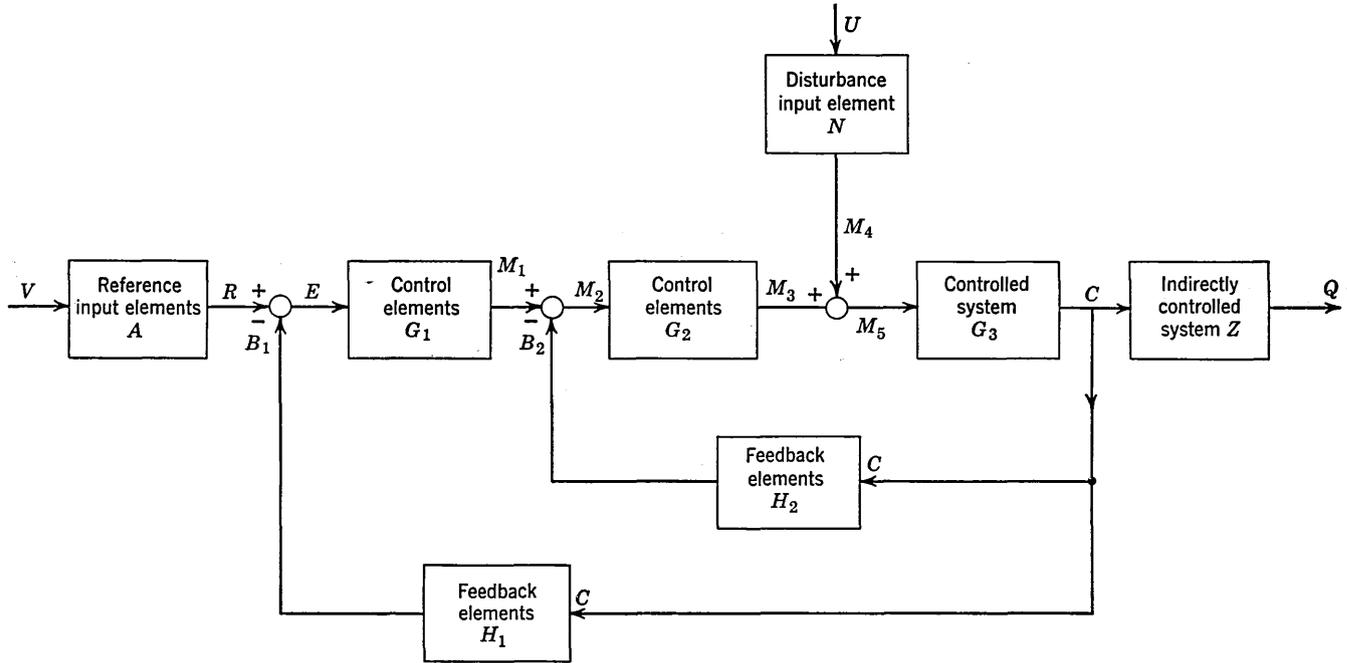
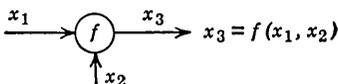


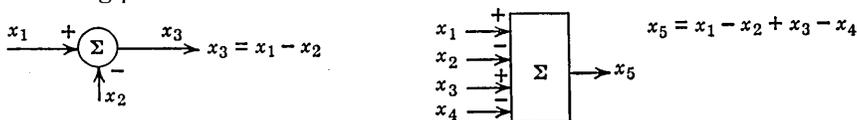
FIG. 24. Block diagram of representative closed loop system.

The I.R.E. standard graphical symbols may also be used (Refs. 30 and 31).

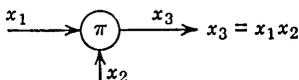
Mixing point:



Summing point:



Multiplication point:



### Algebra of Block Diagrams

A complex block diagram can be rearranged or reduced by combining blocks algebraically. When all the loops are concentric, the indicated manipulations can be carried out directly by successively applying the relation  $C/R = G/(1 + GH)$  to the innermost loops. When the inner loops are not concentric or even intertwining, the block diagrams can usually be reduced to concentric loops by the following *rules* and by reference to Table 17.

1. Data takeoff channels can be moved forward (in the direction of arrows) or backward in the system at will except that the takeoff point cannot pass a summing point. Whenever a data takeoff branch is moved forward past a function  $G$ , the function  $1/G$ , must be added in series with the branch. Whenever a data takeoff branch is moved backward past a function  $G$ , the function  $G$ , must be added in series with the branch.

2. A channel feeding into a summing point can be moved forward and backward in the system at will except that it cannot pass a data takeoff point. As this feed channel is moved forward in the system past a function  $G$ , the function  $G$  must be added in series with the channel. As it is moved backward past a function  $G$ , the function  $1/G$  must be inserted in the channel.

3. In some cases, it will be found necessary to move a takeoff point past a summing point or a summing point past a data takeoff point in order to reduce the system block diagram to simple concentric loops or parallel paths, which can be handled by methods (1) and (2). This can be done by removing a troublesome feedback point or data takeoff by closing an

TABLE 17. THEOREMS FOR THE TRANSFORMATION AND REDUCTION OF BLOCK DIAGRAM NETWORKS (Ref. 17)

Theorem	Original Network	Equivalent Network
1. Interchange of elements		
2. Interchange of summing points		
3. Rearrangement of summing points		
4. Interchange of takeoff points		
5. Moving a summing point ahead of an element		
6. Moving a summing point beyond an element		
7. Moving a takeoff point ahead of an element		
8. Moving a takeoff point beyond an element		
9. Moving a takeoff point ahead of a summing point		
10. Moving a takeoff point beyond a summing point		
11. Combining cascade elements		

TABLE 17. THEOREMS FOR THE TRANSFORMATION AND REDUCTION OF BLOCK DIAGRAM NETWORKS (Continued)

<p>12. Removing an element from a forward loop</p>		
<p>13. Inserting an element in a forward loop</p>		
<p>14. Eliminating a forward loop</p>		
<p>15. Removing an element from a feedback loop</p>		
<p>16. Inserting an element in a feedback loop</p>		
<p>17. Eliminating a feedback loop</p>		
<p>18. Special form of 17</p>		
<p>19. Special form of 17</p>		
<p>20. Inserting a feedback loop to replace an element</p>		
<p>21. Different form of 20</p>		

internal loop, thus replacing a loop by a closed loop transfer function, which has no takeoff or feedback points.

**Examples to Illustrate Transformation Rules.**

EXAMPLE 1. The forward elements  $G_1$ ,  $G_2$ , and  $G_3$ , in Fig. 25a may be combined by multiplication as shown by eq. (56) and Fig. 25b:

$$(56) \quad \frac{C}{E} = \frac{M_1}{E} \cdot \frac{M_2}{M_1} \cdot \frac{C}{M_2} = G_1 G_2 G_3 = G.$$

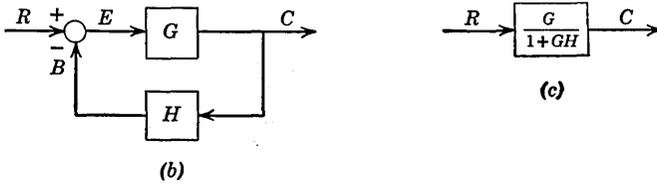
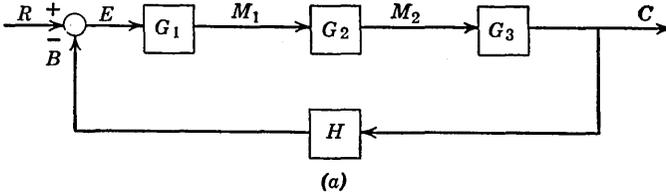


FIG. 25. (a) Simple closed loop system; (b) combinations of forward transfer functions; (c) system in simplest form.

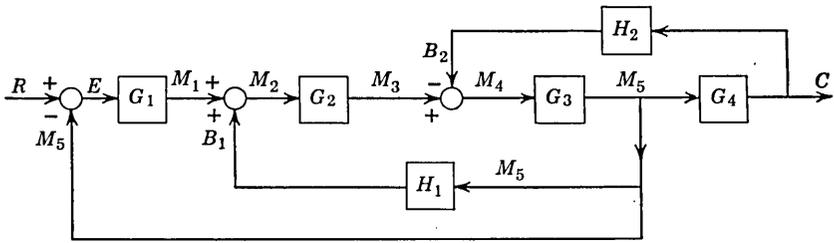
In practice, loading of one element by another must be considered. Figure 25b is further reduced to Fig. 25c by use of eq. (57):

$$(57) \quad \frac{C}{R} = \frac{G}{1 + GH}.$$

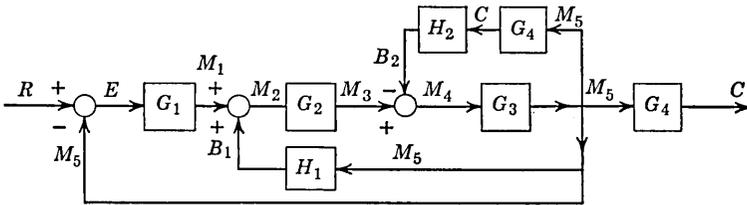
Direct or unity feedback is also common and represents a particular case of eq. (57), where  $H = 1$ .

EXAMPLE 2. Reduction of a complex diagram is shown in Fig. 26. Note in the first reduction, Fig. 26b that the block diagram is altered to include an additional  $G_4$  element for mathematical simplicity, although the signal flow and algebra is identical. Also note in Fig. 26d that a positive feedback is accomplished by using eq. (58), except that the  $H$  has a negative sign.

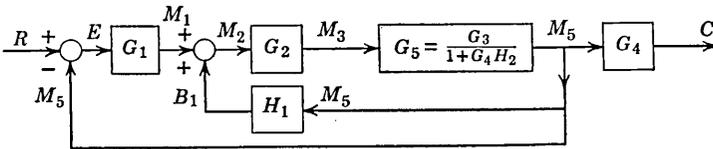
$$(58) \quad G_6 = \frac{G_2 G_5}{1 - G_2 G_5 H_1}.$$



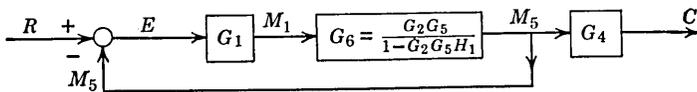
(a) Original system



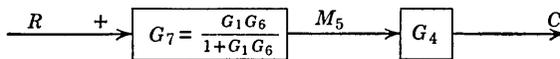
(b) First reduction



(c) Second reduction



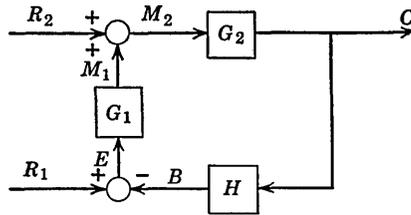
(d) Third reduction



(e) Final reduction

FIG. 26. (Ref. 3a).

EXAMPLE 3. For systems with multiple input and/or disturbances, the superposition theorem is used. The example of Fig. 27 is used to show the response  $C$  as a function of two inputs.



(a) System with multiple inputs

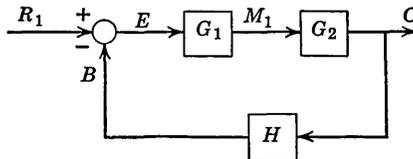
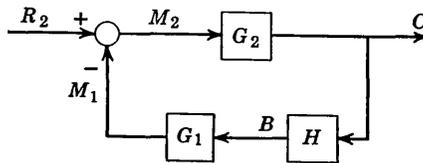
(b) With  $R_2 = 0$ (c) With  $R_1 = 0$ 

FIG. 27.

Let  $R_2 = 0$  as in Fig. 27b:

$$(59) \quad \frac{C}{R_1} = \frac{G_1 G_2}{1 + G_1 G_2 H}$$

Let  $R_1 = 0$  as in Fig. 27c:

$$(60) \quad \frac{C}{R_2} = \frac{G_2}{1 + G_1 G_2 H}$$

Combining inputs:

$$(61) \quad C = \frac{G_1 G_2 R_1 + G_2 R_2}{1 + G_1 G_2 H}$$

#### 4. SYSTEM TYPES

##### Definition of System Types

The idea of the functional similarity of seemingly different transfer functions is strengthened by classification into types. *Three common types*

are ones in which the following conditions are obtained after the transient has subsided:

*Type 0.* A constant value of the controlled variable requires a constant actuating error signal.

*Type 1.* A constant rate of change of the controlled variable requires a constant actuating error signal.

*Type 2.* A constant acceleration of the controlled variable requires a constant actuating error signal.

These characteristics may be identified in terms of the transfer function. For a simple closed loop system with direct feedback the error signal

$$(62) \quad E(s) = R(s) - C(s),$$

where  $R$ , the reference input signal, is compared with  $C$ , the output signal. The forward transfer function

$$(63) \quad G(s) = \frac{C(s)}{E(s)}$$

is of the general form

$$(64) \quad G(s) = \frac{K(1 + a_1s + a_2 + \dots)}{s^n(1 + b_1s + b_2s^2 + b_3s^3 + \dots)}$$

The value of the integer  $n$  in eq. (64) is equal numerically to the type of the system.

Complex plane plots may be obtained by replacing  $s$  by  $j\omega$  in eq. (64). The nature of the plots as  $\omega \rightarrow 0$  will be representative of the type of servomechanism studied, as illustrated in the following section subdivision, Typical Complex Plane Plots.

**Typical Complex Plane Plots**

A *type 0 servomechanism representative plot* is given in Fig. 28 (Ref. 34). At  $\omega = 0$ , the transfer function  $G(j\omega)$  is on the positive real axis and has a

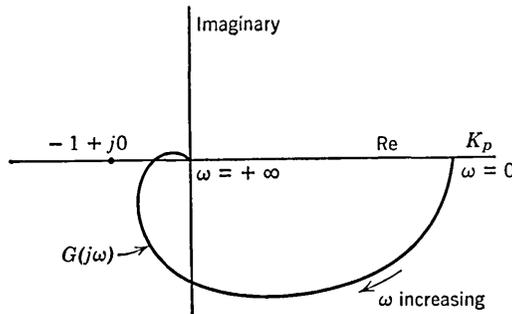


FIG 28. Representative complex plane plot for type 0 servomechanism system (Ref. 34).

finite value  $K_p$ . Generally as  $\omega \rightarrow \infty$ ,  $G(j\omega)$  traverses the fourth and then the third quadrants and approaches the origin.

A *type 1 servomechanism representative plot* is given in Fig. 29 (Ref. 34). For this plot as  $\omega \rightarrow 0$ , the polar plot of  $G(j\omega)$  approaches minus infinity on the imaginary axis. Generally as  $\omega$  increases toward  $+\infty$ ,  $G(j\omega)$  enters

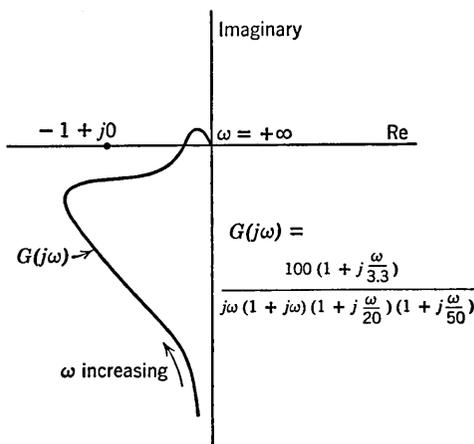


FIG. 29. Representative complex plane plot for type 1 servomechanism system (Ref. 3).

the third and then the second quadrant. The type 1 servomechanism when used for a position control system may also be called a “zero displacement-error system” meaning that the output has the desired value of displacement, in contrast to the type 0 servomechanism, where an error proportional to the desired amount of displacement is necessitated.

A *type 2 servomechanism representative plot* is presented in Fig. 30 (Ref. 34). For this plot as  $\omega \rightarrow 0$ , the plot of  $G(j\omega)$  approaches minus infinity

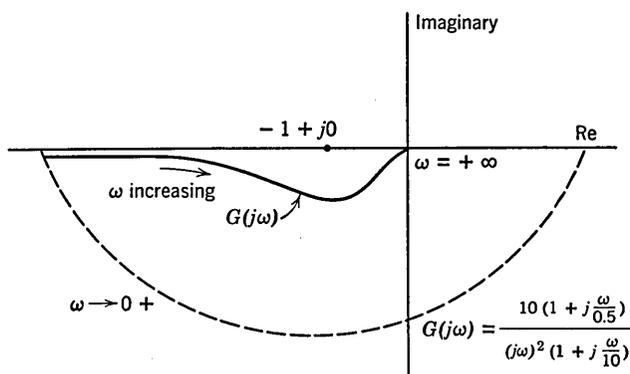


FIG. 30. Representative complex plane plot for type 2 servomechanism system (Ref. 3).

on the real axis. The plot may be closed from  $\omega = 0+$  to  $\omega = 0-$  by a circle of infinite radius traversed in a counterclockwise direction, as indicated by the dotted line. The type 2 servomechanism has a "zero-velocity error" characteristic since it is able to maintain a constant output speed with no actuating error. It is also capable, like the type 1 servomechanism, of maintaining a constant output position without actuating error.

**Typical Application**

*Examples of type 0 servomechanisms* are speed regulators for d-c motors and jet engines and other forms of regulators controlling voltage, current, or temperature, where proportional controllers are employed.

*Examples of type 1 servomechanisms* are position control systems with such integral controllers as d-c motors, hydraulic motors, and hydraulic valve-piston linkages. Other examples of type 1 servomechanisms are speed control systems such as for a jet engine with proportional and integral control.

*Examples of type 2 servomechanisms* are position control systems in which a pilot motor is employed to drive a control element, whose position controls the speed of the main drive motor that supplies power to the load being positioned and torque motors with series compensation.

Block diagrams of each of these servomechanism types are given in Ref. 34. The following paragraph is in substance from *Servomechanism Analysis* by G. J. Thaler and R. G. Brown (Ref. 35).

TABLE 18. CHARACTERISTICS AND APPLICATIONS OF TYPES 0, 1 AND 2 SERVOMECHANISMS (Ref. 35)

Type System	Locus Characteristic	Error Characteristic	Application
0	Closed.	Position error at all times.	Static positioning systems where high accuracy is not important. Some regulator systems.
1	Open. The low-frequency end of the locus goes to infinity along the negative imaginary axis.	No static error. Lag error when operated at constant velocity.	High-accuracy static and dynamic positioning systems.
2	Open. The low-frequency end of the locus approaches infinity along the negative real axis.	No static error. No position error at constant velocity. Constant error in acceleration.	High-accuracy dynamic positioning systems. Control acceleration errors.

In general, the complexity of equipment, cost, and difficulty in design increase greatly with the more advanced type of systems. Type 1 servomechanisms are therefore more common than any of the others. Occasionally, accuracy requirements will justify the type 2 system, and in other cases where high accuracy is not essential a type 0 system is more economical. Table 18 summarizes characteristics and applications of types 0, 1, and 2 servomechanisms.

## 5. ERROR COEFFICIENTS

One of the important figures of merit of a system is its accuracy under various conditions. By *accuracy* is meant the ability of the system to minimize the error between the actual output and the desired output. The usual types of accuracy specified for a control system are its static accuracy and its dynamic accuracy. *Static accuracy* is the accuracy for the output or one of its specified derivatives in a steady-state condition. *Dynamic accuracy* is the accuracy existing during transient conditions of the output and of its derivatives.

### Static Error Coefficients

A *static error coefficient* may be defined as the ratio of the steady-state constant value of the output or of one of its constant derivatives to a constant applied error. The static error coefficients are then:

*Position error coefficient,*

$$K_p = \frac{\text{Output, } c}{\text{Applied error, } e} = \lim_{s \rightarrow 0} \frac{C(s)}{E(s)}$$

for constant output,  $c$ .

*Velocity error coefficient,*

$$K_v = \frac{\text{Velocity of output, } \dot{c}}{\text{Applied error, } e} = \lim_{s \rightarrow 0} \frac{sC(s)}{E(s)}$$

for constant velocity of output,  $\dot{c} \triangleq \frac{dc}{dt}$ .

*Acceleration error coefficient,*

$$K_a = \frac{\text{Acceleration of output, } \ddot{c}}{\text{Applied error, } e} = \lim_{s \rightarrow 0} \frac{s^2 C(s)}{E(s)}$$

for constant acceleration of output,  $\ddot{c} \triangleq \frac{d^2c}{dt^2}$ .

$K_p$ ,  $K_v$ , and  $K_a$  are respectively the gain constants of type 0, 1, and 2 control systems.

For a sinusoidal applied error  $e$ , the error coefficients for types 1 and 2 control systems may be defined in terms of the maximum velocity and acceleration of the output  $c$  as follows:

Velocity error coefficient (type 1):

$$(65) \quad K'_v = \lim_{\omega \rightarrow 0} \frac{\dot{c}_{\max}}{e|\dot{c}_{\max}}$$

where  $e|\dot{c}_{\max} \triangleq e$  at time of max  $\dot{c}$ .

Acceleration error coefficient (type 2):

$$(66) \quad K'_a = \lim_{\omega \rightarrow 0} \frac{\ddot{c}_{\max}}{e|\ddot{c}_{\max}}$$

where  $e|\ddot{c}_{\max} \triangleq e$  at time of max  $\ddot{c}$ .

In the limit  $K'_v$  and  $K'_a$  are identical in value to the values of  $K_v$  and  $K_a$  obtained for constant  $\dot{c}$  and  $\ddot{c}$ .

Table 19 presents a comparison of errors for various types of controlled motion in which  $c$ ,  $\dot{c}$ , or  $\ddot{c}$  is constant or  $c$  is oscillating sinusoidally at a much lower frequency than that corresponding to the shortest time constant of the control system.

TABLE 19. COMPARISON OF STEADY-STATE ERRORS (Ref. 34)

Type of control system	0	1	2
Limit $G(s)$ , $s \rightarrow 0$	$K_p$	$K_v/s$	$K_a/s^2$
Error for constant output, $c$	$c/K_p$	0	0
Error for constant velocity of output, $\dot{c}$	$\infty$	$\dot{c}/K_v$	0
Error for constant acceleration of output, $\ddot{c}$	$\infty$	$\infty$	$\ddot{c}/K_a$
Maximum $e$ for sinusoidally varying $c = C_{\max} \sin \omega t$	$C_{\max}/K_p$	$\omega C_{\max}/K_v$	$\omega^2 C_{\max}/K_a$

### Dynamic Error Coefficients

A form of *dynamic error coefficient* is defined as the ratio of the input of one of its specified derivatives to the component of the error which may be assigned to it during a dynamic condition. That is, the error may be expanded in a series in terms of the input and the derivatives of the input. The dynamic error coefficients are then the reciprocals of the coefficients of the various derivatives since they indicate proportionality between dynamic error components and input derivatives.

Writing the transform of a system with unity feedback gives

$$(67) \quad \frac{E(s)}{R(s)} = \frac{1}{1 + G(s)}.$$

$E(s)$  may be expanded from the ratio of two functions of  $s$ ,

$$(68) \quad E(s) = \frac{R(s)}{1 + G(s)}$$

into  $R(s)$  operated upon by the Maclaurin series expansion of  $1/[1 + G(s)]$ , or by simply dividing the numerator by the denominator,

$$(69) \quad E(s) = \frac{1}{1 + K_p} R(s) + \frac{1}{K_1} sR(s) + \frac{1}{K_2} s^2R(s) \cdots,$$

which is valid near  $s = 0$  (the steady state). Let  $K_0 = 1 + K_p$ .

$K_0, K_1, K_2, \dots$  are commonly called *dynamic error coefficients* (Refs. 36 and 37) of the system. Note. Strictly speaking,  $K_0, K_1, K_2$ , etc., are not "dynamic" error coefficients since the transient terms were lost upon the series expansion of eq. (68). More correctly, they might be termed *steady-state error coefficients*. The term *dynamic* is common usage.

$K_0, K_1, K_2$  may contain not only the values of the static error coefficients  $K_p, K_v$ , and  $K_a$ , respectively, but also expressions involving the static error coefficients and the time constants of the system. Therefore, high gain alone is not sufficient for accurate dynamic performance, for low system time constants are also important for this purpose.

**EXAMPLE.** *Dynamic error coefficients.* For the position servo of Fig. 11:

$$(70) \quad G(s) = \frac{K_v}{s(1 + T_1s)},$$

where  $K_v = K/D$ ,  
 $T_1 = J/D$ .

From eqs. (68) and (70) for unity feedback

$$(71) \quad E(s) = \frac{R(s)}{1 + G(s)} = \frac{s + T_1s^2}{K_v + s + T_1s^2} R(s),$$

$$(72) \quad E(s) = \frac{1}{K_v} sR(s) + \left[ \frac{T_1}{K_v} - \frac{1}{K_v^2} \right] s^2R(s) + \cdots.$$

The dynamic error coefficients in this case are  $K_v$  and  $K_v^2/(K_vT_1 - 1)$ , showing that here the system time constant produces an acceleration component of error proportional to the time constant.

TABLE 20 SERVO ERROR COEFFICIENTS (Ref. 39)

$$\left[ e(t) = \frac{1}{K_0} r(t) + \frac{1}{K_1} \frac{d}{dt} r(t) + \frac{1}{K_2} \frac{d^2}{dt^2} r(t) + \frac{1}{K_3} \frac{d^3}{dt^3} r(t) + \dots (K_0 = 1 + K_p \text{ and } K_p = \infty \text{ for all servos in table}) \right]$$

Locus Identification	Transfer Function $G(s)$	$\frac{1}{K_1}$	$\frac{1}{K_2}$	$\frac{1}{K_3}$
6-12	$\frac{\omega_1 \omega_2}{s(s + \omega_2)}$	$\frac{1}{\omega_1}$	$\frac{\omega_1 - \omega_2}{\omega_1^2 \omega_2}$	$\frac{\omega_2 - 2\omega_1}{\omega_1^3 \omega_2}$
6-12-18	$\frac{\omega_1 \omega_2 \omega_3}{s(s + \omega_2)(s + \omega_3)}$	$\frac{1}{\omega_1}$	$\frac{\omega_1 \omega_2 + \omega_1 \omega_3 - \omega_3 \omega_2}{\omega_1^2 \omega_2 \omega_3}$	$\frac{\omega_1^2 - 2\omega_1 \omega_2 - 2\omega_1 \omega_3 + \omega_2 \omega_3}{\omega_1^3 \omega_2 \omega_3}$
6-12-6-12	$\frac{\omega_1 \omega_2 (s + \omega_3)}{s(s + \omega_2)(s + \omega_4)}$	$\frac{\omega_4}{\omega_1 \omega_3}$	$\frac{\omega_1 \omega_3 (\omega_2 + \omega_4) - \omega_2 \omega_4 (\omega_1 + \omega_4)}{\omega_1^2 \omega_3^2 \omega_2}$	$\frac{\omega_1 \omega_3 (\omega_1 \omega_3 - 2\omega_4^2 - 2\omega_2 \omega_4 - \omega_1 \omega_4 - \omega_1 \omega_2) + \omega_2 \omega_4 (\omega_1 + \omega_4)^2}{\omega_1^3 \omega_3^3 \omega_2}$
6-12-18-12	$\frac{\omega_1 \omega_2 (\omega_3 / \omega_4) (s + \omega_4)}{s(s + \omega_2)(s + \omega_3)}$	$\frac{1}{\omega_1}$	$\frac{\omega_1 \omega_4 (\omega_2 + \omega_3) - \omega_2 \omega_3 (\omega_1 + \omega_4)}{\omega_1^2 \omega_2 \omega_3 \omega_4}$	$\frac{\omega_1 \omega_4 (\omega_1 \omega_4 - 2\omega_3 \omega_4 - 2\omega_2 \omega_4 - \omega_1 \omega_3 - \omega_1 \omega_2) + \omega_2 \omega_3 (\omega_1 + \omega_4)^2}{\omega_1^3 \omega_4^3 \omega_2}$
12-6-12	$\frac{\omega_1 (s + \omega_2)}{s^2 (s + \omega_3)}$	0	$\frac{\omega_3}{\omega_1 \omega_2}$	$\frac{\omega_2 - \omega_3}{\omega_1 \omega_2^2}$

**Error Calculation for a Given Input.** The error may be calculated for a given input when  $r(t)$ ,  $r'(t)$ ,  $r''(t)$ , etc., are known or calculated. First the transforms  $R(s)$ ,  $sR(s)$ ,  $s^2R(s)$ , etc., are evaluated from the input and its derivatives by the formula for real differentiation:

$$(73) \quad s^n F(s) = \mathfrak{L}[f^{(n)}(t)] + \sum_{k=1}^n f^{(k-1)}(0+) s^{n-k} \xrightarrow{s \rightarrow 0} \mathfrak{L}[f^{(n)}(t)].$$

Then the values of  $s^n R(s)$  are substituted into eq. (69) to obtain the Laplace transform of the error,  $E(s)$ . The inverse Laplace transform taken term by term then yields the error as a function of time,  $e(t)$ . If impulses at  $t = 0$  are neglected,

$$(74) \quad e(t) = \frac{1}{K_0} r(t) + \frac{1}{K_1} r'(t) + \frac{1}{K_2} r''(t) + \dots$$

When an analytical expression is not available, the values of  $r(t)$ ,  $r'(t)$ ,  $r''(t)$ , etc., can be estimated from a graph of the input function.

Table 20 gives the literal values of dynamic error coefficients of eq. (74) for some common systems with unity feedback. The locus identification in the table means the succession of decibel per octave slopes with increasing  $\omega$  of the straight-line approximation to the log magnitude versus log frequency diagram for  $G(j\omega)$ .

An *example* of the evaluation of the dynamic error is given in Table 21 (Ref. 40). Use of eq. (74) is illustrated further in Refs. 36 and 38 where six examples are given. See also Chap. 23.

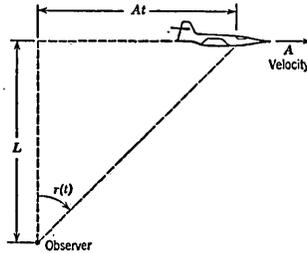
#### Method of Calculating Accuracy of Dynamic Error Coefficients.

As previously stated, the error coefficients of eq. (69) or (74) are evaluated for  $s = 0$  and hence eq. (74) describes the steady-state error. Furthermore, the usefulness or validity of eq. (74) depends upon the nature of  $r(t)$  and also upon the rapidity of convergence of the series. The remainder of this section is in substance according to Arthurs and Martin (Ref. 41). The correction term  $R_{n+1}$  involved in expressing the error  $e(t)$  by a finite number,  $n + 1$ , of coefficients (or terms) of eq. (74) (where  $r$ ,  $r'$ ,  $r''$ ,  $\dots$ ,  $r^n$  are continuous) is

$$(75) \quad R_{n+1} = \int_{-\infty}^t r^{n+1}(\tau) g_{n+1}(t - \tau) d\tau,$$

where by recurrence  $g_{i+1}(t)$  is obtained as  $[-u(t) \int_t^\infty g_i(y) dy]$ ,  $u(t)$  being a unit step and  $g_0(t)$  being the response, such as  $e(t)$ , of the system to a unit impulse.

TABLE 21. DYNAMIC ERROR,  $e(t)$ , FOR SERVO OF FIG. 11 WHEN TRACKING AIRPLANE AT SAME ELEVATION (Ref. 40)



Input Angle  $r(t)$  and Two Derivatives

$r(t)$	$\frac{d^2}{dt^2} [r(t)]$	$\frac{d^2}{dt^2} [r(t)]$
$\tan^{-1} \frac{At}{L}$	$\frac{A/L}{1 + (At/L)^2}$	$\frac{-2(A^3/L^3)t}{[1 + (At/L)^2]^2}$

Values of Error Coefficients  $K_0, K_1, K_2$  (see eqs. 69 and 72)

$\frac{1}{K_0} = \frac{1}{1 + K_-}$	$\frac{1}{K_1}$	$\frac{1}{K_2}$
0	$\frac{1}{K_v}$	$\frac{T_1}{K_v} - \frac{1}{K_v^2}$

Dynamic Error,  $e(t)$  (see eq. 74)

$$e(t) = \left[ \frac{1}{K_v} \right] \left[ \frac{A/L}{1 + (At/L)^2} \right] - \left[ \frac{T_1}{K_v} - \frac{1}{K_v^2} \right] \left[ \frac{2(A^3/L^3)t}{[1 + (At/L)^2]^2} \right] + \dots$$

An upper bound for the correction term  $R_{n+1}$  may be found by replacing  $r^{n+1}(\tau)$  by  $|r^{n+1}(\tau)|_{\max}$  and performing the integration of eq. (75) (a valid procedure for many functions  $g_{n+1}(t - \tau)$ ).

When  $r, r', r'', \dots, r^n$  suffer at most step discontinuities, the response may be expressed by the first  $n + 1$  terms of eq. (74) plus  $R_{n+1}$  of eq. (75) plus the expression:

$$\sum_{i=0}^n \sum_{k=1}^{M_i} \Delta_{ik} g_{i+1}(t - t_{ik}),$$

where  $M_i$  is the number of step discontinuities of  $r^i(t)$ ,  $\Delta_{ik}$  is the magnitude of the discontinuities, and  $t_{ik}$  is the time of the  $k$ th discontinuity. The contribution of impulses can be added in separately.

The response at any time  $t$  may be written in three parts:

1. A finite number of terms of the equivalent of the familiar dynamic error expansion, eq. (74),

2. A corresponding finite set of transient terms which accounts for possible discontinuities in the arbitrary forcing function and its derivatives.

3. A convolution integral which clearly places in evidence the *exact inaccuracy* in the response involved in using a finite number of coefficients (equivalent to the familiar error coefficients).

The above expression results from a closed expansion of the convolution integral to which the response at any time  $t$  may be equated. The usefulness of this expansion of the response into three parts lies in the fact that in many problems  $R_{n+1}$  contributes a small portion of the total response. Consequently  $R_{n+1}$  may either be neglected or crudely approximated without introducing appreciable inaccuracy in the total response. In such cases the process of convolution to obtain the response is replaced by differentiation and summation.

If  $r^{n+1}(\tau)$  in eq. (75) is replaced by its maximum absolute value, then for many functions  $g_{n+1}(t - \tau)$  it may be shown that integration yields an upper bound,

$$|R_{n+1}|_{\max} = |g_{n+2}(0)| |r^{n+1}(\tau)|_{\max} = |c_{n+1}| |r^{n+1}(\tau)|_{\max},$$

where  $c_{n+1}$  is the coefficient of  $r^{n+1}(t)$  in the expansion of the response and of  $s^{n+1}$  in the Maclaurin series for the system function such as  $E(s)/R(s)$ .

EXAMPLE. *Dynamic Error Expressed by Expansion of the Convolution Integral.* Illustrate the expansion of the dynamic error for a case in which there exists a known solution. Let  $E(s)/R(s)$  of eq. (67) be  $B/(s + b)$ , the response characteristics  $g_{i+1}(t)$  being

$$B \left( \frac{1}{-b} \right)^{i+1} e^{-bt} u(t),$$

and the Maclaurin series expansion being

$$\sum_{i=0}^{\infty} \frac{B}{b} \left( \frac{s}{-b} \right)^i$$

(see eq. 69). Then the dynamic error may be expressed by  $n + 1$  terms of the error expansion

$$\sum_{i=0}^n \frac{B}{b} \left( \frac{1}{-b} \right)^i r^i(t)$$

(see eq. 74) plus the transient terms

$$\sum_{i=0}^n \sum_{k=1}^{M_i} \Delta_{ik} B \left( \frac{1}{-b} \right)^{i+1} e^{-b(t-t_{ik})}$$

plus the remainder

$$\int_{-\infty}^t [r^{n+1}(\tau)] \left[ B \left( \frac{1}{-b} \right)^{n+1} e^{-b(t-\tau)} u(t-\tau) \right] d\tau$$

(see eq. 75), which may be bounded by

$$|r^{n+1}(\tau)|_{\max} \left| \frac{B}{b} \left( \frac{1}{-b} \right)^{n+1} \right|,$$

in which is recognized the  $(n + 1)$ th coefficient of the Maclaurin series for  $E(s)/R(s)$ .

Let  $R(s)$  be  $A/(s + a)$ . Then the input and its derivatives  $r^i(t)$  are given by  $(-a)^i A e^{-at} u(t)$ , and for each value of  $i$ ,  $r^i(t)$  has one discontinuity ( $M_i = 1$ ) at time  $t_{i1} = 0$  of magnitude  $\Delta_{i1}$  given by  $(-a)^i A$ . Also  $|r^{n+1}(\tau)|_{\max}$  is given by  $a^{n+1} A$ .

The dynamic error then may be written as

$$\sum_{i=0}^n A \left( \frac{B}{b} \right) \left( \frac{a}{b} \right)^i (e^{-at} - e^{-bt}),$$

with the remainder  $R_{n+1}$  bounded by  $A(B/b)(a/b)^{n+1}$ . This series for the error converges rapidly and is useful for  $a \ll b$ , corresponding to an input which is slow compared with the system response characteristics. As  $n \rightarrow \infty$ , the dynamic error expansion  $\rightarrow AB[(e^{-at} - e^{-bt})/(b - a)]$ , which checks with the inverse Laplace transform of  $AB/(s + a)(s + b)$ .

Another *example* of the expansion of the convolution integral to yield a time response in three parts is given in Ref. 41.

**Relative Usefulness of Error Coefficients.** Both static and dynamic error coefficients are treated in Ref. 34. Dynamic error coefficients are also treated in Refs. 37 and 36, which includes a treatment of the approximate relations, for small overshoot, between the time delay and rise time of the step function response and the dynamic error coefficients.

Dynamic error coefficients  $K_0, K_1, K_2 \dots$  have been defined by eq. (69) and have the same values for all error constants up to and including the one which is nonzero and finite in the classical definition of static error coefficients,  $K_p, K_v, K_a, \dots$  (as given in Static Error Coefficients). This definition of dynamic error coefficients by eq. (69) has the advantage of giving additional information about a system, because the value of any constant is not forced to be zero whenever the preceding constant is nonzero and finite as is the case with static error coefficients  $K_p, K_v, K_a, \dots$ . (These coefficients  $K_0, K_1, K_2 \dots$  are the same in the steady state as the dynamic error coefficients defined for any time  $t$  at the beginning of this section subdivision.)

**Relationship between Dynamic Error Coefficients and Roots of System Equations.** This section is, in substance, from *Automatic Feedback Control System Synthesis* by J. G. Truxal (Ref. 42).

As a result of the relation between  $C/R$  and  $E/R$ , there results the Maclaurin expansion

$$(76) \quad \frac{C(s)}{R(s)} = 1 - \frac{1}{K_0} - \frac{1}{K_1} s - \frac{1}{K_2} s^2 - \dots$$

The relation between the dynamic error coefficients  $K_0$ ,  $K_1$ ,  $K_2$  and the poles and zeros of the closed loop expression  $C/R$  is readily determined if  $C/R$  is written in factored form.

$$(77) \quad \frac{C(s)}{R(s)} = K \frac{(s + z_1)(s + z_2) \cdots (s + z_m)}{(s + p_1)(s + p_2) \cdots (s + p_n)}$$

where the zeros lie at  $-z_i$ , the poles at  $-p_i$ .

The solutions for the dynamic error coefficients are:

$$(78) \quad K_0 = \frac{K \prod_{i=1}^m z_i}{\prod_{i=1}^n p_i - K \prod_{i=1}^m z_i}$$

where  $\prod_{i=1}^m$  indicates the product of all factors from  $i = 1$  to and including  $i = m$  and for cases where  $K_p \rightarrow \infty$  ( $K_0 = 1 + K_p$ ),

$$(79) \quad \frac{1}{K_1} = \sum_{i=1}^n \frac{1}{p_i} - \sum_{i=1}^m \frac{1}{z_i}$$

$$(80) \quad \frac{-2}{K_2} = \frac{1}{K_1^2} + \sum_{i=1}^n \frac{1}{p_i^2} - \sum_{i=1}^m \frac{1}{z_i^2}$$

Equations (79) and (80) are of basic importance in servo synthesis for they represent the correlation between the dynamic error coefficients and the system response characteristics, specifically the time delay and rise time of the response of the system to a step function. In addition the two equations indicate the manner in which lead and integral equalization permit control over  $K_1$  and  $K_2$  without affecting relative stability. Generalized (dynamic) error coefficients are treated in greater detail in Ref. 36, and their relation to closed loop roots is treated at length in Ref. 42.

**Guillemin's Method.** Equations (79) and (80) are of basic importance in feedback control system synthesis by Guillemin's method, which is described in Chap. 23.

In the first step of this method the closed loop transfer function is determined from the specifications for frequency response and transient response. In Chap. 23 use is made of techniques for obtaining with the aid of these two equations the zeros and poles for compensation required to obtain the desired closed loop transfer function.

## 6. ANALYSIS OF A-C SERVOS: CARRIER SYSTEMS

In a d-c servo the signals are directly proportional to the instantaneous amplitude whereas in an a-c system the signals are modulated carrier waves, and the information is carried in the modulation. For instance in an amplitude modulated system the envelope of the modulated wave contains the signal information. Owing to the convenience and simplicity of using a synchro or chopper circuit as a modulator and a two-phase motor as a demodulator, many carrier systems use suppressed-carrier, amplitude modulation. Frequency and phase modulation are not yet in common use although they offer theoretical advantage for null detection.

The value of an a-c system lies in the possible use of sensitive, accurate, low-force level sensors, inexpensive and relatively easily produced as amplifiers, and power elements with low maintenance requirements.

**Basic Types of Elements.** Three types of elements are encountered in carrier systems:

Type 1. Elements in which both input and output are modulated carriers.

Type 2. Modulators which have inputs at signal frequencies and outputs which are modulated carriers.

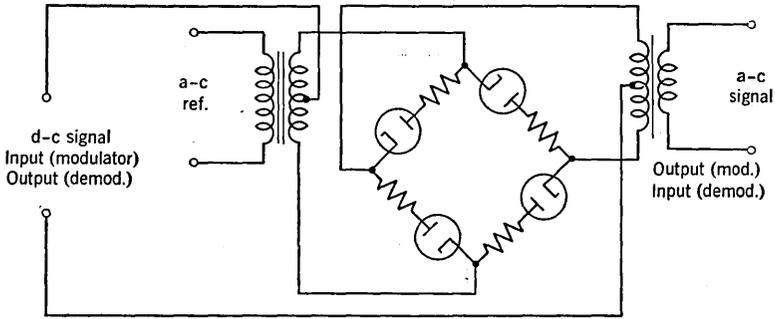
Type 3. Demodulators which have modulated carriers as inputs and have signal frequencies as outputs.

Figure 31 shows several electronic circuits that can be used as modulators and demodulators. In addition the mechanically tuned or electromechanical choppers are used extensively. The a-c servo motor also serves as a demodulator and the a-c tachometer acts as a modulator.

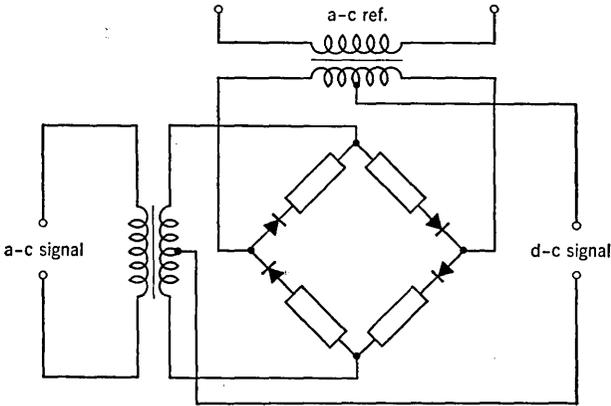
Two-phase servomotors and a-c tachometers are analyzed in Refs. 45-47 and 49-51. An accurate mathematical representation of these elements is complex and simplifying assumptions and analogies are commonly used.

**Simplifying Assumptions.** The majority of work done with a-c feedback systems has been on suppressed carrier systems. For this type of system the following assumptions are normally made:

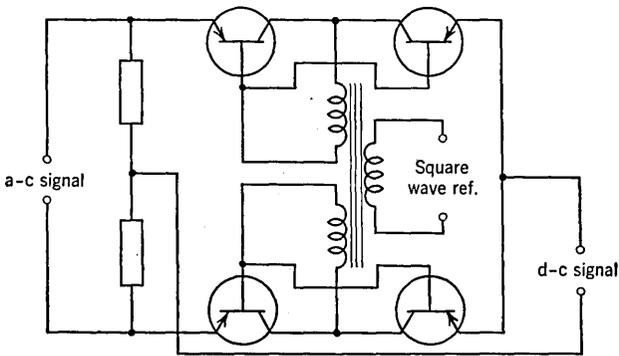
FEEDBACK CONTROL



Vacuum Tube Diode



Diodes



Transistor

FIG. 31. Modulators and demodulators.

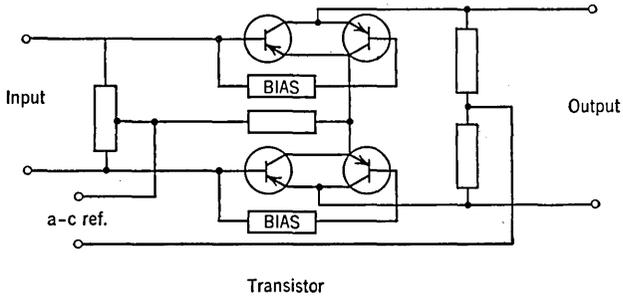
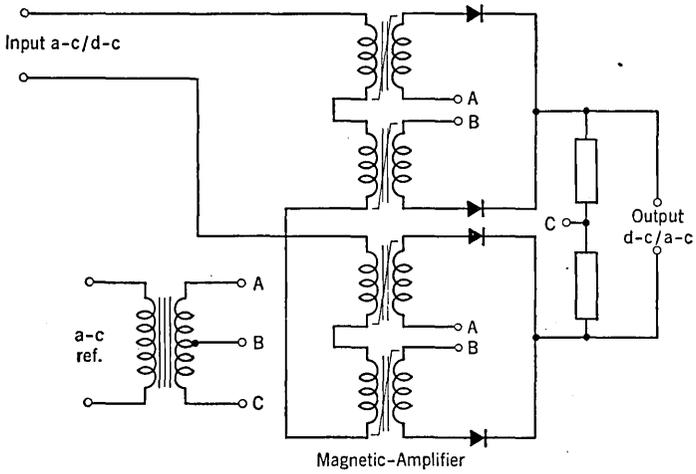
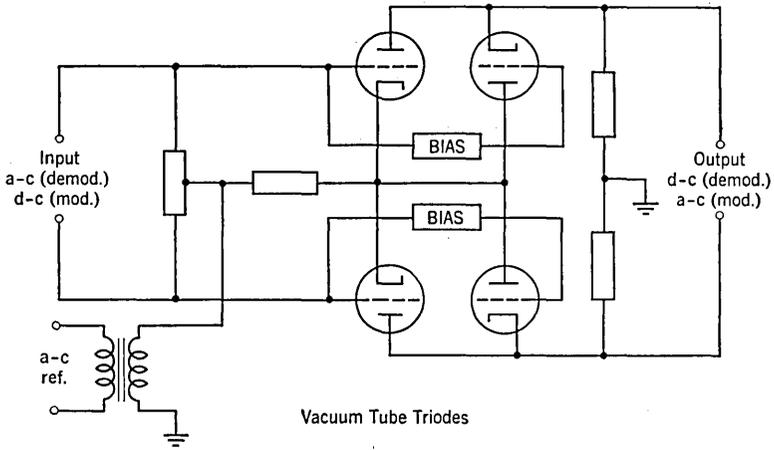


FIG. 31. Modulators and demodulators (continued).

1. Modulators generate perfect suppressed carrier signals; i.e., harmonics besides the sideband frequencies, self-generated noise, and serious phase shifts in the modulator are neglected.
2. The motor acts as a perfect demodulator with a prescribed static and dynamic relationship between the driving voltage envelope and the output shaft position.
3. The carrier frequency and magnitude remain constant.

**Suppressed Carrier System.** A simple suppressed carrier open loop system is shown in Fig. 32. If the input to the preamplifier is the product of the carrier and the reference input, then

$$(M \cos a\omega_c t)(V \cos \omega_c t) = MV(\cos a\omega_c t)(\cos \omega_c t).$$

Expanding this yields

$$(81) \quad MV[\cos (1+a)\omega_c t + \cos (1-a)\omega_c t].$$

As indicated by eq. (81) this type of modulator has an output containing only the two sideband frequencies, the sum and difference of the modulating

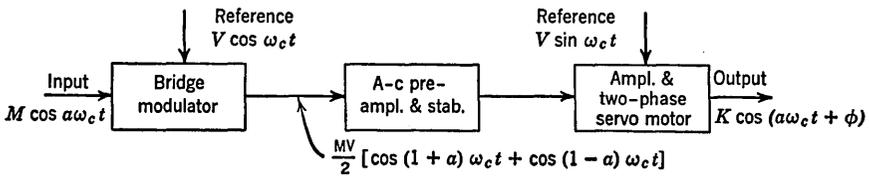


FIG. 32. Open loop carrier system (Ref. 3b).

and carrier frequencies,  $(1 - a)\omega_c$  and  $(1 + a)\omega_c$ , and no carrier frequency,  $\omega_c$ . This gives rise to the name suppressed carrier.

A typical suppressed-carrier feedback system using tachometer stabilization is shown in Fig. 33. The signal equations are indicated at the significant points in the system.

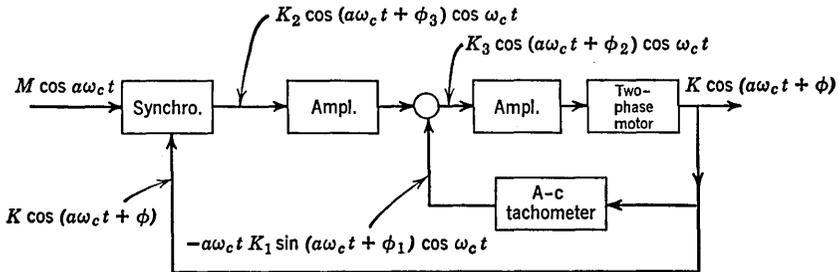


FIG. 33. Typical a-c servo with tachometer feedback (Ref. 3b).

**System Analysis and Design.** For the purposes of system analysis the a-c components can usually be treated in the same manner as analogous d-c components. For instance as shown in Fig. 34, the speed-torque curves of a d-c and an a-c machine are analogous, and the analysis of the a-c machine can proceed in the same manner as the analysis of the d-c machine. See Table 5, Sect. 1. Note, however, that the a-c machine

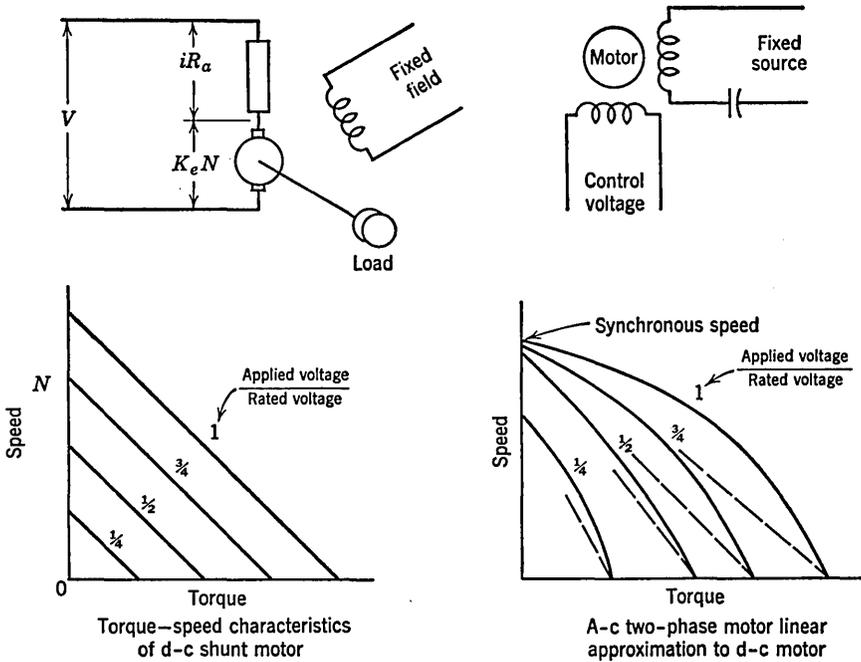


FIG. 34. D-c analogy to a-c two-phase motor.

characteristics are nonlinear and that to derive a linear transfer function, the analysis must be carried out on a linearized, incremental change basis. The linearization methods of Sect. 2, Chap. 25, can be used.

There are torques at other frequencies besides the signal frequency produced in the motor. These torques are at frequencies  $(2 + a)\omega_c$  and  $(2 - a)\omega_c$  and normally because of the high frequency and low amplitude produce little mechanical motion. However, the associated currents can produce important heating effects. Similarly because the motor tends to discriminate against quadrature control signals, the currents produce little torque but the heating and/or saturating effects of quadrature currents can be important.

The a-c and d-c analogy can be extended to a-c tachometers. An a-c tachometer with a control signal of a frequency  $\omega_c$  and an amplitude pro-

portional to  $\omega_c$  affects system performance in a manner similar to a d-c tachometer.

Alternating-current stabilizing networks and a-c system stability are treated in Chap. 23, Sect. 3.

Noise, quadrature voltage or carrier phase shift, variations in carrier frequency, and pickup present major problems in a-c system design, and their consideration dictates as much as the stability analysis the form and characteristics selected. As a result it is desirable to define the environment and operating requirements before investigating the system stability.

#### ACKNOWLEDGMENTS

The cooperation of the following is gratefully acknowledged in granting permission to reproduce material in this chapter:

American Institute of Physics. From: *Journal of Applied Physics* (part of Sect. 5).

General Electric Company. From: *Servomechanisms and Regulating System Design* by H. Chestnut and R. W. Mayer (John Wiley & Sons, New York) (Tables 3, 4, 5, 16, 19; Figures 11, 23, 26, 28, 29, 30, 32, 33).

McGraw-Hill Book Company. From: *Automatic Feedback Control* by W. R. Ahrendt and J. F. Taplin (Table 21); *Servomechanism Practice* by W. R. Ahrendt (Tables 1, 20); *Servomechanism Analysis* by G. J. Thaler and R. G. Brown (Tables 2, 18).

*Westinghouse Engineer* (Tables 6, 7, 8).

John Wiley & Sons. From: *Transients in Linear Systems* by H. F. Gardner and J. L. Barnes (Tables 9, 15).

American Institute of Electrical Engineers. From: *Electrical Engineering* (Table 17).

Bureau of Aeronautics, United States Navy. From report prepared by Northrop Aircraft Co. (Tables 13 to 15, Figures 16 to 19).

#### REFERENCES

1. E. A. Guillemin, *Synthesis of RC Networks, J. Math. Phys.*, **28**, 22-42 (1949).
2. G. J. Thaler and R. G. Brown, *Servomechanism Analysis*, Chap. 1, McGraw-Hill, New York, 1953.
3. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design: (a) Vol. I, 1951, (b) Vol. II, 1955, Wiley, New York.*
4. J. R. Ketchum and R. T. Craig, *Simulation of Linearized Dynamics of Gas-Turbine Engines, Nall. Advisory Comm. Aeronaut., Tech. Notes 2826*, November 1952.
5. L. M. Toss, How to reckon basic process dynamics, *Control Eng.*, **3**, 50-55 (1956).
6. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. II, Chap. 1, Wiley, New York, 1955.
7. J. B. Reswick, Determine System Dynamics—without Upset, *Control. Eng.*, **2**, 50-57 (1955).
8. J. G. Truxal, *Automatic Feedback Control System Synthesis*, McGraw-Hill, New York, 1955.

9. W. M. Gaines, Frequency response methods in design of turbojet engine controls, *Second Feedback Controls System Conference*, Am. Inst. Elec. Engrs., April 1954.
10. W. R. Ahrendt, *Servomechanism Practice*, McGraw-Hill, New York, 1954.
11. S. W. Herwald, Forms and principles of servomechanisms, *Westinghouse Eng.*, **6**, 149-155 (1946).
12. W. R. Evans, *Control System Dynamics*, McGraw-Hill, New York, 1954.
13. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. 1, Chap. 3, Wiley, New York, 1951.
14. S. B. Crary, *Power System Stability*, Vol. 1, p. 1, Wiley, New York, 1945.
15. M. F. Gardner and J. L. Barnes, *Transients in Linear Systems*, Vol. 1, Appendix A, Wiley, New York, 1942.
16. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. 1, Chap. 4, Wiley, New York, 1951.
17. M. F. Gardner and J. L. Barnes, *Transients in Linear Systems*, Vol. 1, Chaps. 3-6, Wiley, New York, 1942.
18. G. J. Thaler, *Elements of Servomechanism Theory*, Chap. 3, McGraw-Hill, New York, 1955.
19. G. J. Thaler and R. G. Brown, *Servomechanism Analysis*, Chap. 4, McGraw-Hill, New York, 1953.
20. Methods of Analysis and Synthesis of Piloted Aircraft Flight Control Systems, BuAer Rept. AE 61-4I, March 1952, Appendix, Sect. A.
21. M. F. Gardner and J. L. Barnes, *Transients in Linear Systems*, Vol. 1, Chap. 8, Wiley, New York, 1942.
22. E. Weber, *Linear Transient Analysis*, Vol. 1, Chap. 2, Wiley, New York, 1954.
23. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. 1, Chaps. 9 and 10, Wiley, New York, 1951.
24. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. 1, Chaps. 12 and 13, Wiley, New York, 1951.
25. G. S. Brown and D. P. Campbell, *Principles of Servomechanisms*, Chap. 8, Wiley, New York, 1948.
26. H. M. James, N. B. Nichols, and R. S. Phillips, *Theory of Servomechanisms*, Chap. 4, McGraw-Hill, New York, 1947.
27. M. F. Gardner and J. L. Barnes, *Transients in Linear Systems*, Vol. I, Chaps. 2 and 7, Wiley, New York, 1942.
28. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. I, Chap. 7, Wiley, New York, 1951.
29. A.I.E.E. Standards Subcommittee on Terminology and Nomenclature of the Feedback Control Committee, Am. Inst. Elec. Engrs., January 1950.  
See also *Letter Symbols for Feedback Control Systems*, ASA Y10.13-1955, American Standards Association, New York, July 1955.
30. IRE 26.S2 Standards on Terminology for Feedback Control Systems, 1955, *Proc. I.R.E.*, **44**, 107-109 (1956).
31. IRE 26.S1 Standards on Graphical and Letter Symbols for Feedback Control Systems, 1955, *Proc. I.R.E.*, **43**, 1608-1609 (1955).
32. T. D. Graybeal, Transformation of Block Diagram Network, *Elec. Eng.*, **70**, 985-990 (1951).
33. T. M. Stout, A Block Diagram Approach to Network Analysis, *Trans. Am. Inst. Elec. Engrs.*, Application and Industry, **71**, 255-260 (1952).
34. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. 1, Chap. 8, Wiley, New York, 1951.

35. G. J. Thaler and R. G. Brown, *Servomechanism Analysis*, Chap. 7, McGraw-Hill, New York, 1953.
36. J. G. Truxal, *Automatic Feedback Control System Synthesis*, Chap. 1, McGraw-Hill, New York, 1955.
37. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. II, Chap. 2, Wiley, New York, 1955.
38. P. E. Smith, Jr., *Design Regulating Systems by Error Coefficients*, *Control Eng.*, **2**, 69-74 (1955).
39. W. R. Ahrendt, *Servomechanism Practice*, Chap. 14, McGraw-Hill, New York, 1954.
40. W. R. Ahrendt and J. F. Taplin, *Automatic Feedback Control*, Chap. 7, McGraw-Hill, New York, 1951.
41. E. Arthurs and L. H. Martin, Closed expansion of the convolution integral (A generalization of servomechanism error coefficients), *J. Appl. Phys.*, **26**, 58 (1955).
42. J. G. Truxal, *Automatic Feedback Control System Synthesis*, Chap. 5, McGraw-Hill, New York, 1955.
43. R. A. Bruns and R. M. Saunders, *Analysis of Feedback Control Systems*, McGraw-Hill, New York, 1955.
44. M. Panzer, Envelope transfer function analysis in a-c servosystems, *Trans. Am. Inst. Elec. Engrs.*, **75**, 274-279 (1956).
45. S. S. L. Chang, Transient analysis of a-c servomechanisms, *Trans. Am. Inst. Elec. Engrs.*, **74**, 30-37 (1955).
46. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. II, Chap. 6, Wiley, New York, 1955.
47. R. J. W. Koopman, Operating characteristics of two-phase servo motors, *Trans. Am. Inst. Elec. Engrs.*, **68**, Pt. I, 319-329 (1949).
48. A. Hopkin, Transient response of small two-phase induction motors, *Trans. Am. Inst. Elec. Engrs.*, **70**, Pt. I, 881-886 (1951).
49. L. O. Brown, Transfer function for a two-phase induction servo motor, *Trans. Am. Inst. Elec. Engrs.*, **70**, Pt. 2, 1890-1893 (1951).
50. R. H. Frazier, Analysis of the drag-cup a-c tachometer, *Trans. Am. Inst. Elec. Engrs.*, **70**, Pt. 2, 1894-1906 (1951).
51. S. A. Davis, Using a two-phase motor as a tachometer, *Control Eng.*, **2**, 75-76 (1955).
52. J. G. Truxal, *Automatic Feedback Control System Synthesis*, Chap. 6, McGraw-Hill, New York, 1955.
53. G. M. Attura, Effects of carrier shifts on derivative networks for AC servomechanisms, *Trans. Am. Inst. Elec. Engrs.*, **70**, Pt. 1, 612-618 (1951).
54. C. S. Draper, W. McKay, and S. Lees, *Instrument Engineering*, Vol. II, McGraw-Hill, New York, 1953.

## Stability

W. E. Sollecito and S. G. Reque

1. Introduction	21-01
2. Classical Solution Approach	21-02
3. Routh's Criterion	21-05
4. Nyquist Stability Criterion	21-09
5. Bode Attenuation Diagram Approach	21-29
6. Root Locus Method	21-46
7. Miscellaneous Stability Criteria	21-71
8. Closed Loop Response from Open Loop Response	21-72
References	21-81

### 1. INTRODUCTION

**Definition of Stability.** A *stable* system is one wherein all transients decay to zero in the steady state. An *unstable* system is here loosely defined as one in which the response variable increases without bound with a bounded signal input.

**Reason for Stability Analysis.** The primary objective of a control system design is to devise a system such that a controlled variable is related to a command signal in a desired manner within permissible tolerances. If power elements with reliable, unchanging characteristics were available, the problems of control system design would be much simplified. Since, in the main, the characteristics of power elements change with time, temperature, load, pressure, etc., a feedback element is employed to

remove the deleterious effects of change in element characteristics. To improve performance of the system, a natural solution is to increase the gain or amplification in the system. The combination of a closed loop and high gain leads to problems of instability.

**Purpose of Stability Analysis.** To be a satisfactory control system, the system deviation resulting from any normally encountered deviation stimulus must reduce with increasing time to a small value within acceptable tolerance. It is the purpose of stability studies to indicate a *system's dynamic behavior*, and if this behavior is improper or inadequate, the studies should point the way toward proper system revision to improve performance.

**Methods of Stability Analysis.** The methods of studying stability presented in this chapter are restricted to linear systems. A *linear* system is one in which the output due to simultaneous inputs is the same as the

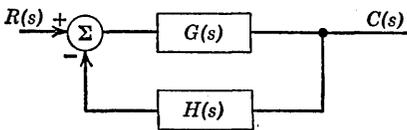


FIG. 1. General negative feedback system.

sum of the several outputs due to the inputs acting alone. In other words, a linear system is one which may be described by ordinary linear differential equations wherein the *theorem of superposition* holds true. For nonlinear systems, see Chap. 25.

Consider the general feedback system shown in Fig. 1;  $s$  is the Laplace transform complex variable. The transfer characteristics are given by

$$(1) \quad \frac{C(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)}.$$

The stability of a system is uniquely defined by those values of  $s$  which make

$$(2) \quad 1 + G(s)H(s) = 0.$$

All the methods of stability analysis, therefore, confine themselves to investigation of eq. (2), in one fashion or another. The techniques can be classified in two general categories: those which obtain the *explicit values of the roots* of eq. (2) and those which obtain *information about the bounded regions* wherein all the roots lie. In the first category belong the *classical approach* and the *root locus method*. In the second category belong *Routh's criteria*, *Nyquist's criteria*, *Bode's method*, and many others. The relative merits of each will be discussed as each method is examined in detail.

## 2. CLASSICAL SOLUTION APPROACH

As shown in Chap. 20, it is possible to relate the controlled variable to the command (reference) variable by a differential equation. For the sta-

bility studies involved here, assume this is a linear differential equation of the form

$$(3) \quad a_0 \frac{d^n x(t)}{dt^n} + a_1 \frac{d^{n-1} x(t)}{dt^{n-1}} + \cdots + a_n x(t) \\ = b_0 \frac{d^m y(t)}{dt^m} + b_1 \frac{d^{m-1} y(t)}{dt^{m-1}} + \cdots + b_m y(t).$$

Since this equation is linear, the solution may be broken into the sum of two solutions, the *particular* solution and the *complementary* or *homogeneous* solution.

The *particular* solution,  $x_{ss}$ , also called the *forced response* or the *steady-state* solution, is of the form

$$(4) \quad x_{ss} = f(x).$$

In other words, the forced response is of the same character as the reference. For *example*, if  $y$  is sinusoidal,  $x_{ss}$  is also sinusoidal.

**Characteristic Equation.** When the operator  $p$  is substituted for  $d/dt$  and  $y$  is set equal to zero, eq. (3) becomes

$$(5) \quad a_0 p^n x(t) + a_1 p^{n-1} x(t) + a_2 p^{n-2} x(t) + \cdots + a_{n-1} p x(t) + a_n x(t) = 0$$

or

$$(6) \quad [a_0 p^n + a_1 p^{n-1} + a_2 p^{n-2} + \cdots + a_{n-1} p + a_n] x(t) = 0.$$

This is the *characteristic equation* (see Chap. 20) of the system. The *complementary solution*,  $x_t$ , also called the *homogeneous* solution or *transient* solution, is of the form

$$(7) \quad x_t = A_1 e^{p_1 t} + A_2 e^{p_2 t} + A_3 e^{p_3 t} + \cdots + A_n e^{p_n t}.$$

The exponents  $p_1, p_2, p_3, \cdots, p_n$  are the roots of eq. (6). The coefficients  $A_1, A_2, A_3, \cdots, A_n$  depend upon the initial conditions of the system and the forcing function,  $y$ .

Note that when multiple roots occur, say  $p_1 = p_2$ , the transient solution is of the form

$$(8) \quad x_t = A_1 e^{p_1 t} + A_2 t e^{p_2 t} + A_3 e^{p_3 t} + \cdots + A_n e^{p_n t}.$$

The total solution is the sum of the two parts

$$(9) \quad x = x_{ss} + x_t.$$

**Relation of Stability to Characteristic Equation.** Instability has been defined as the output becoming large without bound for bounded input. Since the steady-state solution is of the same character as the forcing function, only the transient solution can provide terms which in-

crease without bound for bounded input. This occurs when any of the roots  $p_1, p_2, p_3, \dots, p_n$  have positive real parts because the corresponding exponential terms in eq. (7) or (8) tend to infinity as  $t$  becomes infinite. Because  $A_1, A_2, A_3, \dots, A_n$  are finite values depending on initial conditions and  $y_{ss}$  is bounded for a bounded input, *system stability is dependent only upon the nature of the characteristic equation of the system!* In other words, *system stability is uniquely determined by the behavior of the exponential terms in the transient response given by eq. (7) or (8).*

a. If all the roots have negative real parts, all the exponential terms decay to zero as time increases. This is a stable system.

b. If any of the roots have positive real parts, the corresponding exponential terms increase without limit. This is an unstable system.

c. If any of the roots are purely imaginary, the corresponding terms oscillate at constant amplitude. This condition is the dividing point between a stable and an unstable system. It is here also considered unstable.

d. If it so happens that multiple roots occur, i.e.,  $p_1 = p_2$ , which are purely imaginary, the output increases without bound. Again this is an unstable condition.

*The fundamental problem in ascertaining system stability is therefore one of determining the nature of the roots of the characteristic equation of a given system.* The straightforward method of determining stability of a system consists of the following steps:

a. Write the differential equation of the system relating input and output variables.

b. Substitute  $p$  for  $d/dt$  and equate the input signal to zero. This is the characteristic equation of the system in operational form.

c. Obtain the roots of the characteristic equation with assigned values for all constants.

d. Examine the roots. If all roots have negative real parts, the system is stable. If any of the roots have zero or positive real parts, the system is unstable.

Figure 2 shows the regions of root location for stable and unstable systems.

*Note.* If the Laplace transform method of analysis had been used, the conclusions would have been identical except that the complex variable  $s$  would replace the operator  $p$ . Equation (2) would yield:

$$(10) \quad [a_0s^n + a_1s^{n-1} + \dots + a_n]X(s) = [b_0s^m + b_1s^{m-1} + \dots + b_m]Y(s).$$

As an input-output ratio similar to eq. (1) this is

$$(11) \quad \frac{C(s)}{R(s)} = \frac{X(s)}{Y(s)} = \frac{b_0s^m + b_1s^{m-1} + \dots + b_m}{a_0s^n + a_1s^{n-1} + a_2s^{n-2} + \dots + a_n}.$$

Solution of the equation resulting from setting the denominator of eq. (11) to zero is *exactly* the same as solution of eq. (2).  $[1 + G(s)H(s)]$  is a fraction of polynomials in  $s$  where the numerator polynomial is the characteristic function of the system. As shown in Fig. 2, stability is uniquely defined by those values of  $s$  which satisfy eq. (2). Because of more universal acceptance, the complex variable  $s$  will be used in place of the operator  $p$  in all the following methods of stability analysis.

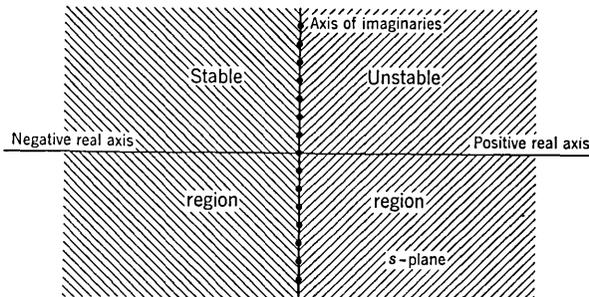


FIG. 2. Pictorial representation of stability definition.

**Relative Merits of the Classical Solution Approach.** This method of determining stability has the advantage of being theoretically exact, but suffers from two major disadvantages:

a. A great amount of labor is required to factor equations of degree higher than 3.

b. To factor any higher order equations, the coefficients must be numerical values. The loss of system parameters in literal form obscures the ways to improve system performance should redesign become necessary.

### 3. ROUTH'S CRITERION (Refs. 1, 2, 3)

In 1877 E. J. Routh developed an algebraic method for determining whether a polynomial has roots with positive real parts. This method does not reveal the exact values of the roots but shows the bounded regions wherein they are located. Reference to Fig. 2 shows that this is all that is necessary to determine whether a system is stable or not. If all roots lie in the left half  $s$ -plane, the system is stable.

#### Application of the Routh Criterion.

*Step 1.* Write the characteristic equation in the form

$$(12) \quad [a_0s^n + a_1s^{n-1} + a_2s^{n-2} + \cdots + a_{n-1}s + a_n]X(s) = 0.$$

Remove all the zero roots, i.e., the roots that occur at  $s = 0$ . If the zero roots do occur, they can easily be recognized because  $s$  or some mul-

title of  $s$  will be common to all terms in eq. (12). For example, if  $a_n = 0$  in eq. (12),  $s$  would be common to all terms and could be placed outside the brackets.

*Step 2.* Examine eq. (12) to see that all the coefficients of  $s$  are non-zero and of the same sign. If this is not true, an unstable system is immediately indicated.

*Step 3.* Arrange the coefficients in an array of the form:

Index							
$n$ :	$a_0$	$a_2$	$a_4$	$a_6$	$\cdot$	$\cdot$	$\cdot$
$n - 1$ :	$a_1$	$a_3$	$a_5$	$a_7$	$\cdot$	$\cdot$	$\cdot$
$n - 2$ :	$b_1$	$b_2$	$b_3$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$n - 3$ :	$c_1$	$c_2$	$c_3$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$n - 4$ :	$d_1$	$d_2$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$n - 5$ :	$e_1$	$e_2$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$n - 6$ :	$f_1$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$n - 7$ :	$g_1$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$

The index number indicates the highest order of  $s$  in a row.

The first two rows consist of all the terms in the given equation and the rest are calculated in the following fashion.

$$b_1 = \frac{a_1 a_2 - a_0 a_3}{a_1}, \quad b_2 = \frac{a_1 a_4 - a_0 a_5}{a_1}, \text{ etc.}$$

$$c_1 = \frac{b_1 a_3 - a_1 b_2}{b_1}, \quad c_2 = \frac{b_1 a_5 - a_1 b_3}{b_1}, \text{ etc.}$$

$$d_1 = \frac{c_1 b_2 - b_1 c_2}{c_1}, \quad d_2 = \frac{c_1 b_3 - b_1 c_3}{c_1}, \text{ etc.}$$

etc.

Notice that two terms in the first column are used in each calculation. As the term to be calculated shifts to the right, the additional two terms in the formula shift to the right also. The formulas for calculation of terms in a row use only those terms in the two rows immediately above. The process is continued for  $(n + 1)$  rows where  $n$  is the order of the characteristic equation.

*Step 4.* After the array has been completed, stability can be investigated by inspection of terms in the first column. *The number of changes in sign of the terms in the first column is the number of roots with positive real parts.* This constitutes Routh's criterion.

EXAMPLE. Given the fourth order equation

$$8s^4 + 2s^3 + 3s^2 + s + 5 = 0.$$

The array becomes:

Index				
4:	+ 8	3	5	$\left[ b_1 = \frac{2 \cdot 3 - 8 \cdot 1}{2} = \frac{-2}{2} = -1 \right]$
3:	+ 2	1		$\left[ b_2 = \frac{2 \cdot 5 - 8 \cdot 0}{2} = \frac{2 \cdot 5}{2} = 5 \right]$
2:	- 1	5		$\left[ c_1 = \frac{-1 \cdot 1 - 2 \cdot 5}{-1} = \frac{-11}{-1} = 11 \right]$
1:	+11			$\left[ d_1 = \frac{11 \cdot 5 - (-1 \cdot 0)}{11} = \frac{11 \cdot 5}{11} = 5 \right]$
0:	+ 5			

There are two changes of sign in column one (between indexes 3 and 2, and 2 and 1), therefore the equation must have two roots with positive real parts. Since a fourth order equation has four roots, the remaining two roots must lie in the left half  $s$ -plane.

*Note.* A generalization can be made from this example. The last term (+5) came down through the array without change. Since all the coefficients in the equation are positive, the first two terms in column one are positive. Only terms of index 2 and 1 in column one can be negative. Thus, a maximum of two sign changes can occur. Therefore, one can conclude that if all terms in a fourth order equation are nonzero and of the same sign, at least two roots must lie in the left half  $s$ -plane. This conclusion is of no great import in itself but it merely points the way to intelligent use of this method of analysis.

**Special Cases in Applying the Routh Criterion.** Because the Routh criterion can be used to advantage in other commonly used stability studies, it is worth while to pursue the criterion in greater detail here.

*a. Row multiplication.* Any row may be multiplied by a positive constant without affecting the criterion. This may be used to decrease the arithmetic labor involved.

*b.* When the first term in a row is zero and other terms in the same row are not zero. To continue the process, replace the first column zero by an arbitrarily small positive constant,  $\Delta$ , and continue the calculations. Examine the complete array in the usual fashion. If necessary,  $\Delta$  may be assigned any arbitrarily small value. This number may be positive or negative but is customarily assumed positive.

c. When *all terms in a row are zero*. This special case arises when roots lying radially opposite and equidistant from the origin occur as shown in Fig. 3. A pair of conjugate pure imaginary roots is of this category. When a row of zeros occurs, take the preceding row of coefficients and form a subsidiary function. This subsidiary function is the polynomial in  $s$  having as coefficients the terms of a row; the exponent of the highest power of  $s$  is the index of the row and successive powers of  $s$  decrease by two.

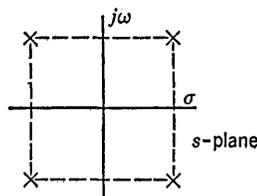


FIG. 3. Roots radially opposite and equidistant from origin.

EXAMPLE. The subsidiary function of the row with index 3 of the preceding example is

$$f(s) = 2s^3 + s,$$

whereas the subsidiary function of the row with index 2 is

$$f(s) = -s^2 + 5.$$

Upon formation of the subsidiary function of the row preceding the row of zeros, differentiate it with respect to  $s$  and replace the row of zeros by the corresponding coefficients of the differentiated function. Proceed in the usual manner. The index numbers remain unaltered. Upon completion of the array, the number of changes in sign indicates the number of roots in the right half  $s$ -plane. The remaining roots are either in the left half  $s$ -plane or on the axis of imaginaries. One of several procedures can be utilized to determine the number of each (Ref. 4). A straightforward approach is as follows.

In the original equation replace  $s$  by  $-s$ . This substitution rotates all the roots of the equation through 180 degrees. Those roots of the original equation in the left half  $s$ -plane are now in the right half  $s$ -plane. Application of Routh's criterion to this new equation determines the number of these roots. Thus, the number of roots of the original equation in the left half  $s$ -plane has been ascertained. The total number of roots is equal to the order of the original equation. Therefore the number of roots on the axis is equal to the total minus the sum of those in the right and left half  $s$ -planes.

**Relative Merits of Routh's Criterion.** This criterion serves as a quick check on absolute system stability. It can also be used to advantage in the more powerful Nyquist criterion. It nicely avoids the necessity for factoring an equation to determine the nature of its roots. This method does not provide a clear indication of system performance and does not clearly show the ways to improve a design should improvement be required.

#### 4. NYQUIST STABILITY CRITERION

This powerful criterion is based on the fact that the frequency response of the open loop transfer function indicates the stability characteristics of the closed loop system. In Fig. 1 the open loop transfer function is represented by  $G(s)H(s)$ .

##### Restrictions on the General Nyquist Criterion.

a.  $G(s)H(s)$  must be the ratio of the transforms of linear differential equations.

b.  $G(s)H(s)$  must be single valued and an analytic function (Ref. 5) for all values of  $s$  having zero or positive real parts except at possible discrete points (Ref. 4).

**Basic Definitions.** In general  $G(s)H(s)$  is a fraction of rational polynomials in  $s$ .

$$(13) \quad G(s) = \frac{N_1(s)}{D_1(s)} = \frac{K_1(s + s_1)(s + s_3) \cdots}{(s + s_2)(s + s_4)(s + s_6) \cdots}$$

$$(14) \quad H(s) = \frac{N_2(s)}{D_2(s)} = \frac{K_2(s + s'_1)(s + s'_3) \cdots}{(s + s'_2)(s + s'_4)(s + s'_6) \cdots}$$

The all important eq. (2) can be written as

$$(15) \quad 1 + G(s)H(s) = 1 + \frac{N_1(s)N_2(s)}{D_1(s)D_2(s)}$$

$$(16) \quad 1 + G(s)H(s) = \frac{D_1(s)D_2(s) + N_1(s)N_2(s)}{D_1(s)D_2(s)}$$

*Characteristic Function.*  $[D_1(s)D_2(s) + N_1(s)N_2(s)]$  represents the characteristic function of the closed loop system of Fig. 1. The characteristic equation is merely the characteristic function set equal to zero.

*Zeros.* The factors  $(s + s_1)$ ,  $(s + s_3)$ ,  $\cdots$ , represented by  $N_1(s)$  are called zeros of  $G(s)$ . This terminology arises because when  $s$  takes on the value of a root of  $N_1(s)$ , i.e.,  $-s_1$ ,  $-s_3$ ,  $\cdots$ ,  $N_1(s)$  equals zero and  $G(s)$  does likewise per eq. (13).

*Poles.* The factors  $(s + s_2)$ ,  $(s + s_4)$ ,  $\cdots$ , represented by  $D_1(s)$  are called poles of  $G(s)$ . When  $s$  takes on the value of a root of  $D_1(s)$ , i.e.,  $-s_2$ ,  $-s_4$ ,  $\cdots$ ,  $D_1(s)$  equals zero and  $G(s)$  goes to infinity per eq. (13). This rise to infinity is called a pole.

*Note.* Per eq. (16), poles of  $G(s)H(s)$  are also poles of  $[1 + G(s)H(s)]$  whereas zeros of  $[1 + G(s)H(s)]$  are unknown and their nature to be determined by the stability criterion. Zeros of  $[1 + G(s)H(s)]$  are poles of  $C(s)/R(s)$ .

**Nyquist Criterion. General Procedure.**

a. Plot  $G(s)H(s)$  for  $s$  traversing the boundary of the entire right half  $s$ -plane in a clockwise direction. (See following note.)

b. Draw a vector,  $\bar{V}$ , from  $(-1 + j0)$  [the minus one point in the  $G(s)H(s)$ -plane] to  $G(s)H(s)$  and observe the angular rotation of this vector for the above values of  $s$ .

c. Let  $R$  be the net number of revolutions of this vector.  $R$  is positive for counterclockwise revolutions and negative for clockwise revolutions.

d. Determine the number of poles of  $G(s)H(s)$  in the right half  $s$ -plane, i.e., poles with positive real parts. Call this integer number  $P$ . If necessary, Routh's criterion may be used to determine this.

e. The number of zeros of  $[1 + G(s)H(s)]$ ,  $Z$ , is determined from the equation

$$(17) \quad Z = P - R.$$

f. The system is stable if and only if  $Z = 0$ , i.e., if the number of counterclockwise revolutions of  $G(s)H(s)$  about the  $-1$  point is equal to the number of poles of  $G(s)H(s)$  in the right half  $s$ -plane.

*Note.* If  $G(s)H(s)$  has any poles on the  $j\omega$ -axis (i.e., pure imaginary roots), when  $s$  is taking on values up the  $j\omega$ -axis, it must bypass these points. It is customary to make  $s$  traverse a small semicircle to the right of these points as shown in Fig. 4. If  $G(s)H(s)$  ever does have poles on

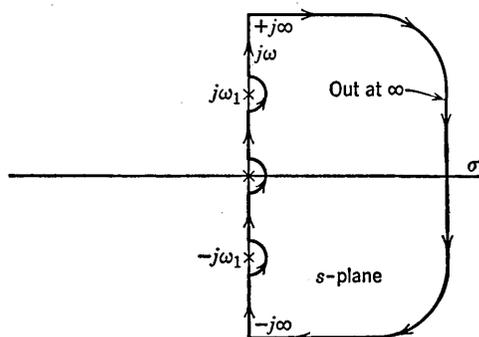


FIG. 4. Traversal of  $s$  for the Nyquist plot where  $G(s)H(s)$  has poles at  $\pm j\omega_1$  and 0.

the  $j\omega$ -axis whose values are unknown, it is almost as much effort to determine these as it is to find the zeros of  $[1 + G(s)H(s)]$  directly. In this case, *Dzung's criteria* (Refs. 19, 20) may be a better approach for stability analysis. Fortunately this condition arises infrequently.

**The Physical Meaning of Making  $s$  Traverse the  $j\omega$ -Axis.** In short, it is obtaining the steady-state frequency response of the open loop transfer

function  $G(s)H(s)$ . Consider the case shown in Fig. 5.  $A \sin \omega t$  is the input and in the steady-state condition,  $B \sin(\omega t + \theta)$  is the output. To be theoretically exact, the *steady-state condition* is the condition that exists after an infinite time has elapsed. This allows all the transients to die out to absolute zero for a stable system. For practical consideration, steady state occurs after the transients have settled down to arbitrarily

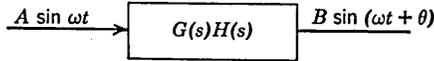


FIG. 5. Frequency response measurements.

small values. Comparison of the ratio of the output sinusoid to the input sinusoid reveals that a *gain change*,  $B/A$ , and a *phase shift*,  $\theta$ , have occurred. This gain change and phase shift are due to  $G(s)H(s)$  and can be considered as the *magnitude* and *direction* of a vector. This is the vector notation of the steady-state behavior of  $G(s)H(s)$ .

**Sinusoidal Input Variable.** The Laplace transform of the input variable is

$$(18) \quad \mathcal{L}^{-1}[\sin \omega t] = \frac{\omega}{s^2 + \omega^2} = \frac{\omega}{(s + j\omega)(s - j\omega)}$$

The graphical representation of the Laplace transform of this sinusoid is a pair of points on the imaginary axis a distance of  $\pm\omega$  from the origin. As the frequency of the sinusoid varies,  $\omega$  varies with it. See Fig. 6.

Consider the case where

$$(19) \quad G(s)H(s) = \frac{K}{(s + s_2)(s + s_4)}$$

The poles of  $G(s)H(s)$  are plotted at  $-s_2$ , and  $-s_4$  in the  $s$ -plane in Fig. 6. In this same figure are plotted the poles of the input sinusoid whose frequencies are successively  $\omega_0, \omega_1, \omega_2, \omega_3, \omega_4, \dots, \omega_n$ . The corresponding vectors representing the gain change and phase shift  $G(j\omega)H(j\omega)$  for each frequency are plotted in Fig. 7. One method to obtain  $G(s)H(s)$  for any particular  $s$  is to substitute the particular value of  $s$  in eq. (19). An equivalent, but more illuminating, procedure is to consider each factor in eq. (19) as a separate vector. In Fig. 6 are shown the two factor vectors for  $s = j\omega_2$ . As  $s$  assumes values up the  $j\omega$ -axis, the vectors from the roots increase in magnitude and phase. Since these vectors appear in the denominator of eq. (19), as  $s$  traverses up the  $j\omega$ -axis, the magnitude of  $G(s)H(s)$  decreases whereas its phase becomes increasingly negative. For this particular transform given by eq. (19), for positive  $\omega$ ,  $G(j\omega)H(j\omega)$  lies in the third and fourth quadrants. The entire curve expands or contracts with respective increase or decrease in the gain constant  $K$ .

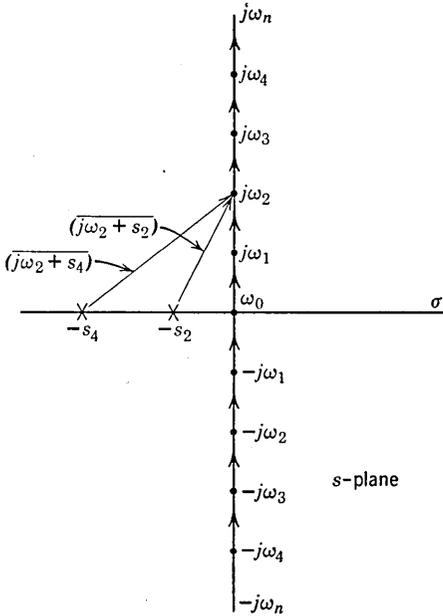


FIG. 6. s-Plane plot of input sinusoid and vectors of  $G(s)H(s)$  factors.

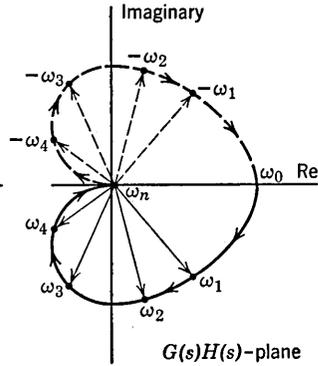


FIG. 7.  $G(s)H(s)$ -Plane plot of frequency response of  $G(s)H(s)$ .

**Conformal Mapping.** Mathematically,  $G(s)H(s)$  is a function which transforms a point in the  $s$ -plane to a point in the  $G(s)H(s)$ -plane. This mapping of points or curves in one plane to points or curves in another plane is called conformal mapping. The line along the  $j\omega$ -axis in the  $s$ -plane maps into the curve in the  $G(s)H(s)$ -plane shown in Fig. 7 by use of the transform  $G(s)H(s)$ . An important point to remember is that any curve in the  $s$ -plane produces a corresponding curve in the  $G(s)H(s)$ -plane. The curve in the  $s$ -plane which lies on the  $j\omega$ -axis corresponds to the input function being a variable frequency sinusoid. The shape of the corresponding curve in the  $G(s)H(s)$ -plane depends on the particular fraction of polynomials represented by  $G(s)H(s)$ .

Figure 8 shows other lines along which  $s$  might vary. Figure 9 shows the corresponding curves of  $G(s)H(s)$  for  $G(s)H(s)$  given by eq. (19). Points on line (1) correspond to input functions of the form.

$$(20) \quad e^{-\sigma_1 t} \sin \omega t$$

whose Laplace transform is

$$(21) \quad \mathcal{L}^{-1}[e^{-\sigma_1 t} \sin \omega t] = \frac{\omega}{(s + \sigma_1 + j\omega)(s + \sigma_1 - j\omega)}$$

Points on line (2) correspond to input functions whose Laplace transform is

$$(22) \quad \mathcal{L}^{-1}[e^{\sigma_1 t} \sin \omega t] = \frac{\omega}{(s - \sigma_1 + j\omega)(s - \sigma_1 - j\omega)}$$

The conformal mapping procedure obtains definite corresponding curves for  $G(s)H(s)$  as shown in Fig. 9.

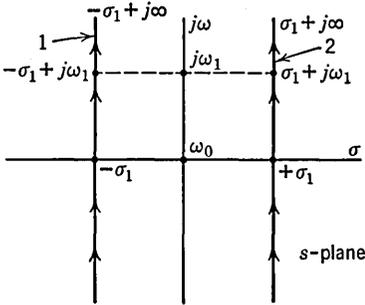


FIG. 8. Particular paths of  $s$  in  $s$ -plane.

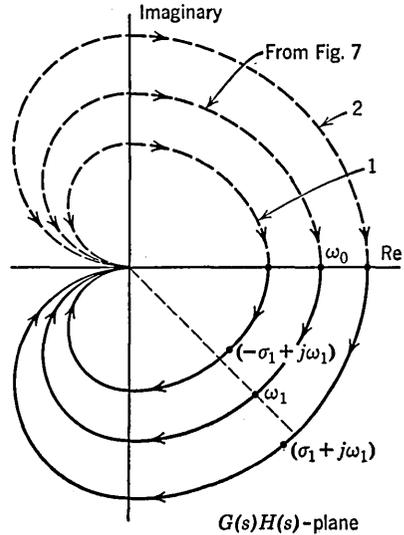


FIG. 9. Plot of  $G(s)H(s)$  for paths of  $s$  in Fig. 8.

**Principles of Nyquist Criterion.** By use of conformal mapping principles it can be shown (Ref. 6) that if  $s$  is made to traverse the boundaries of a given area, observation of the behavior of the vector from the  $-1$  point to  $G(s)H(s)$  in the  $G(s)H(s)$ -plane indicates how many zeros of  $[1 + G(s)H(s)]$  lie in the area whose boundaries were traversed by  $s$ .

Refer to Fig. 10, where  $s$  is made to traverse the boundary of area  $A$ , and the corresponding path of  $G(s)H(s)$  is as shown in Fig. 11. Observation of the net rotation of the vector  $\bar{V}$  about the  $-1$  point gives a clear indication of the roots of  $[1 + G(s)H(s)]$  in area  $A$ . For every pole of  $G(s)H(s)$  located in area  $A$ ,  $\bar{V}$  will experience *one net counterclockwise* rotation about the  $-1$  point. For every zero of  $[1 + G(s)H(s)]$  in area  $A$ ,  $\bar{V}$  will experience *one net clockwise* rotation about the  $-1$  point. Therefore if the number of poles,  $P$ , of  $G(s)H(s)$  in area  $A$  is known, the number of zeros of  $[1 + G(s)H(s)]$  in area  $A$  can be found by subtracting from  $P$  the number of net revolutions of  $\bar{V}$  about the  $-1$  point. If area  $A$  is

made to encompass the entire right half of the  $s$ -plane, existence of zeros of  $[1 + G(s)H(s)]$  in this area can be determined from the above procedure and stability of the closed loop system can be ascertained!

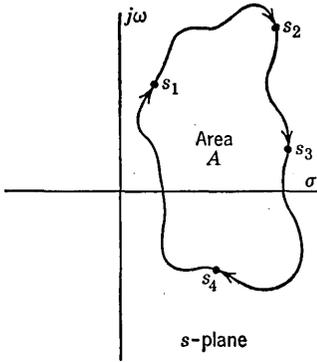


FIG. 10. Arbitrary path of  $s$  in the  $s$ -plane.

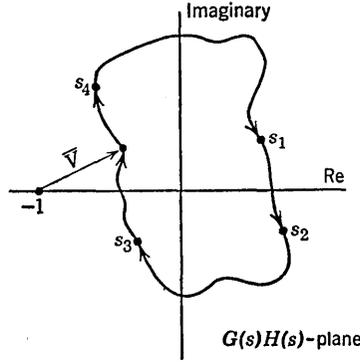


FIG. 11. Corresponding path of  $G(s)H(s)$ .

The left portion of the boundary in Fig. 12 corresponds to making the input to  $G(s)H(s)$  a sinusoid. The traversal out at infinity is only of mathematical importance because infinite values are difficult to handle in physical equipment. For practical purposes, that finite region relatively close to the origin is of major importance as will be more clearly demonstrated in the Bode approach to stability analysis.

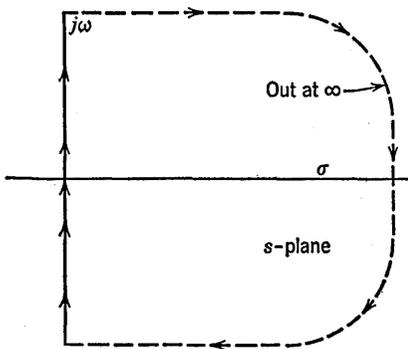


FIG. 12. Path of  $s$  enclosing entire right half  $s$ -plane.

**Summary.** Stability is uniquely defined by those values of  $s$  which make

$$1 + G(s)H(s) = 0.$$

To ascertain existence of zeros of  $[1 + G(s)H(s)]$  in the right half  $s$ -plane, Nyquist's criterion requires  $s$  to traverse the boundary of the entire right half  $s$ -plane. The portion of the boundary of major importance, the  $j\omega$ -axis, corresponds to a sinusoidal input function. Therefore, the frequency response of the open loop transfer function  $G(s)H(s)$  gives clear indication of stability of the closed loop system. This is most fortunate because constant amplitude variable frequency generators are much easier to build than exponentially varying variable frequency

generators. Experimental procedures are thereby more easily implemented.

**Application of Nyquist Stability Criterion.**

EXAMPLE 1. Given

$$G(s)H(s) = \frac{K(s + s_1)}{s(s + s_2)(s + s_4)(s + s_6)}$$

In Fig. 13 consider  $s$  in the region from  $b$  to  $c$ . As  $\omega$  becomes increasingly large,

$$\lim_{s \rightarrow j\infty} [G(s)H(s)] = \frac{K}{s^3} = 0 / -270^\circ.$$

In this region  $G(s)H(s)$  approaches zero asymptotically to the  $-270$ -degree direction, i.e., the  $+j\beta$ -line. As  $s$  traverses the boundary  $c-d-e$

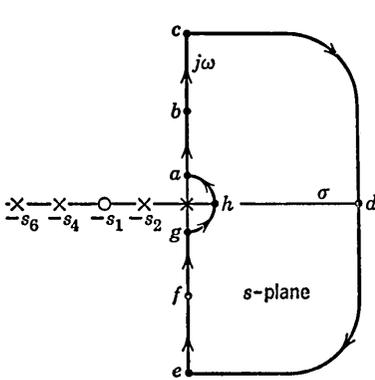


FIG. 13. s-Plane plot

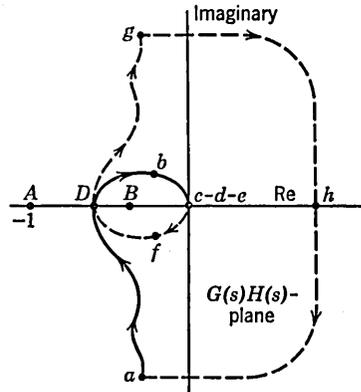


FIG. 14. Nyquist plot of  $G(s)H(s)$ .

$$G(s)H(s) = \frac{K(s + s_1)}{s(s + s_2)(s + s_4)(s + s_6)}$$

out at infinity, the  $G(s)H(s)$  vector rotates 540 degrees in the counter-clockwise direction, but since the magnitude is zero, this rotation is unobservable. The region  $e-f$  is the conjugate of  $c$  to  $b$ . In the region  $f$  to  $g$  there is a continuous curve which wiggles a bit because of the pole zero locations as shown in Fig. 13. At point  $g$  the  $s$  traverse takes a 90-degree turn to the right. In conformal mapping, angles are preserved in the small, therefore the  $G(s)H(s)$  plot also takes a 90-degree turn to the right. In the region  $g-h-a$ ,  $G(s)H(s)$  behaves like  $K/s$ .

$$\lim_{s \rightarrow 0} [G(s)H(s)] = \frac{K}{s}$$

In other words, the movement of  $s$  is very close to the pole at the origin so the vectors of the poles, and zeros relatively far away do not experience great change. The vector of the pole at the origin experiences a 180-degree change in the counterclockwise sense. Since this vector is in the denominator of  $G(s)H(s)$ ,  $G(s)H(s)$  experiences a 180-degree change in the clockwise sense.

The region  $a$  to  $b$  is the conjugate of  $g$  to  $f$ . The  $G(s)H(s)$  plot is usually plotted solid for  $s = +j\omega$  and dotted for the rest of the boundary. From Fig. 13 it is apparent that  $G(s)H(s)$  has no poles in the right half  $s$ -plane.  $P = 0$ . Notice that the zero of  $G(s)H(s)$  is not considered at all. If the gain constant is such that the  $-1$  is at point  $A$  in Fig. 14,  $R = 0$ . Therefore

$$Z = P - R = 0 - 0 = 0,$$

and the closed loop system is stable.

If the gain constant is raised such that the  $-1$  point is at  $B$ , there are two clockwise encirclements of the  $-1$  point,

$$Z = P - R = 0 - (-2) = +2,$$

and the closed loop system is unstable and has two poles in the right half  $s$ -plane.

If the gain were adjusted such that the  $-1$  point were at  $D$ , i.e., the  $G(s)H(s)$  curve passes right through the  $-1$  point,  $R$  is indeterminate. This condition produces a constant amplitude sinusoidal oscillation in the closed loop system. A change in the gain constant is like changing the calibration on the coordinate axes.

EXAMPLE 2. Given

$$G(s)H(s) = \frac{K(s - 10)}{s^2 + 100}.$$

In the region  $a$  to  $b$  in Fig. 15

$$G(s)H(s) = \frac{K(-10 + j\omega)}{100 - \omega^2} = \frac{K\sqrt{\omega^2 + 100}}{100 - \omega^2} / \tan^{-1}(\omega/-10).$$

$$\text{For } \omega = 0, G(s)H(s) = \frac{K}{10} / -180^\circ.$$

As  $\omega$  increases, the magnitude of  $G(s)H(s)$  increases and the phase angle becomes more negative.

As  $\omega$  approaches 10,  $G(s)H(s)$  approaches infinity along the  $+135$ -degree line. In the region  $b$ - $c$ - $d$

$$\lim_{s \rightarrow j10} [G(s)H(s)] = \frac{K}{(s - j10)}.$$

Therefore, since  $s$  takes a 90-degree right turn and proceeds 180 degrees counterclockwise,  $G(s)H(s)$  takes a right turn and proceeds 180 degrees clockwise. In the region  $d-e$ ,  $G(s)H(s)$  is well behaved and proceeds to zero as  $s$  approaches  $j\infty$ . As  $s \rightarrow \infty$ ,

$$\lim_{s \rightarrow j\infty} [G(s)H(s)] = \frac{K}{s} = 0/\underline{-90^\circ}.$$

Along  $e-f-g$ ,  $G(s)H(s)$  remains at zero. The rest of the curve is the conjugate image of  $e$  to  $a$ . For this system  $P = 0$ . From Fig. 16,  $R = 0$  for the  $-1$  point at  $A$ . This system is stable.

For the  $-1$  point at  $B$ ,  $R = -1$  and  $Z = 1$ . This system is unstable.

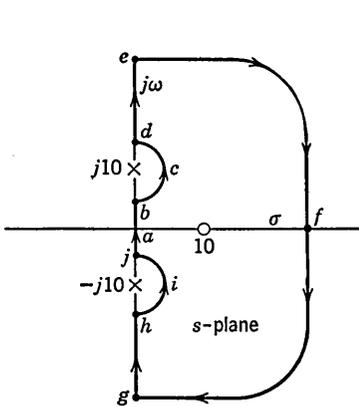


FIG. 15.  $s$ -Plane plot

$$G(s)H(s) = \frac{K(s - 10)}{s^2 + 100}.$$

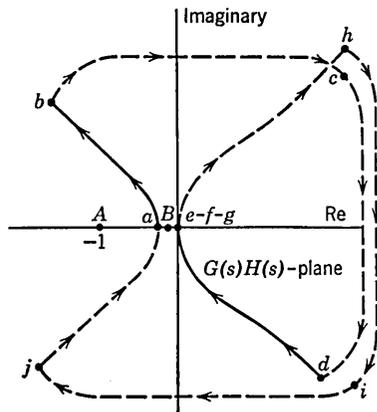


FIG. 16. Nyquist plot of  $G(s)H(s)$ .

**Practical Considerations in Plotting Diagrams.** If one should ever find that the number of counterclockwise encirclements of the  $-1$  point is greater than  $P$ , he may correctly infer that he has made a mistake in calculating either  $P$  or  $R$ !

The procedure in drawing the Nyquist diagrams is first to draw in the approximate shape of the  $G(s)H(s)$  curve for the prescribed traversal of  $s$ . The labor involved is by no means negligible. To avoid unnecessary labor, the reader is advised to learn first how to use the following Bode diagrams and apply them to obtain the exact Nyquist plot when necessary. The Bode approach is presented after the Nyquist criterion for ease in presenting the requisite theory. Subsequent usage should by no means be affected by order of theoretical presentation.

Strictly speaking, the small semicircles about poles of  $G(s)H(s)$  on the  $j\omega$ -axis and also the traversal of  $s$  out at infinity do not correspond to constant amplitude sinusoidal input. The polar plot used in the Nyquist criterion is therefore not strictly a frequency response plot. For purposes of simple definition, these exceptions are overlooked.

**Abbreviated Nyquist Stability Criterion.** When the open loop transfer function is stable by itself

$$P = 0,$$

and the criterion for stability reduces to

$$R = 0.$$

**STATEMENT 1.** *For a stable open loop transfer function, the closed loop system will be stable if there are no encirclements of the  $-1$  point in the  $G(s)H(s)$ -plane for  $s = j\omega$ .*

The criterion may be further reduced to observing the behavior of  $G(s)H(s)$  for positive  $\omega$  in the region where the magnitude of  $G(s)H(s)$  is near unity. The additional restriction is that  $G(s)H(s)$  becomes a constant less than 1 (or zero) as  $s$  becomes increasingly large. This restriction means that in eq. (15) the order of  $N_1(s)N_2(s)$  is less than or equal to the order of  $D_1(s)D_2(s)$ . Where the respective orders are equal, the product of gain constants,  $K_1K_2$ , from eqs. (13) and (14) must be less than 1.

For the cases that fall within the above-mentioned restrictions (and there are many), the criterion can be restated.

**STATEMENT 2.** *In the region of frequencies where  $G(j\omega)H(j\omega)$  is near the unit circle, the system is stable if the  $-1$  point is not encircled.*

**STATEMENT 3.** *If the further restriction is imposed that  $G(j\omega)H(j\omega)$  is well behaved in the region of the unit circle, then stability is indicated by the phase angle of  $G(j\omega)H(j\omega)$  for positive values of  $\omega$  when it crosses the unit circle. For phase angles less than  $-180$  degrees at unit circle crossover the system is stable. For phase angles more negative than  $-180$  degrees at unit circle crossover, the system is unstable. In Fig. 17,  $G_1(s)H_1(s)$  represents a stable closed loop system whereas  $G_2(s)H_2(s)$  represents an unstable system. A well-behaved  $G(s)H(s)$  is loosely defined as one that does not wander too much in the region of the unit circle. A not too well-behaved open loop transfer function is shown in Fig. 18. For systems of this type, the general Nyquist criterion should be used. Adequate information about system stability is contained in Fig. 18, but more than the first unit circle crossover must be inspected.*

**Phase Margin.** For those systems that do fall within the abbreviated criterion, additional definitions have evolved.

The *phase* of  $G(j\omega)H(j\omega)$ , measured with respect to the positive real axis

and defined as positive in the counterclockwise sense, is given as  $\theta$ . The *phase margin* is the phase of  $G(j\omega)H(j\omega)$  at unit circle crossover and is measured with respect to the direction of the  $-1$  point:

$$(23) \quad \gamma = 180^\circ + \theta.$$

In Fig. 17,  $G_1(s)H_1(s)$  has a positive phase margin whereas  $G_2(s)H_2(s)$  has a negative phase margin. *Phase margin at unit circle crossover evidences system stability with plus and minus values indicating stable and unstable systems respectively.* Zero phase margin at unit circle crossover means that  $G(j\omega)H(j\omega)$  passes through the  $-1$  point and therefore that the closed loop system will sustain a constant amplitude oscillation.

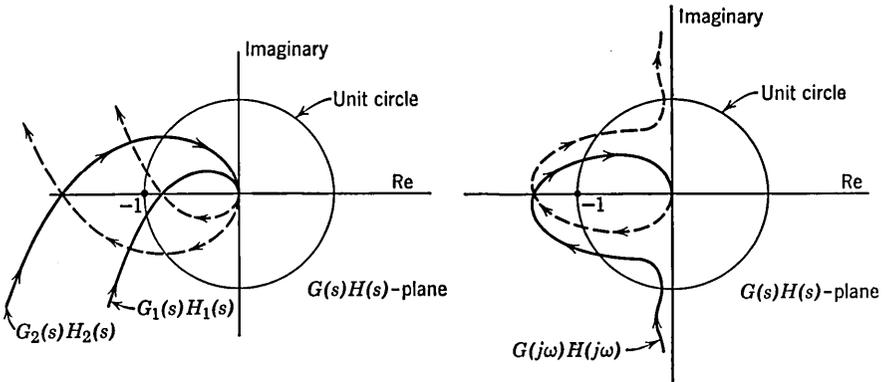


FIG. 17. Unit circle in the  $G(s)H(s)$ -plane. FIG. 18. Not too well-behaved  $G(s)H(s)$ .

**EXAMPLE.** In Fig. 18 the phase margin at the first unit circle crossover is positive. The  $-1$  point is encircled so the system is unstable. This example illustrates the case where inspection of only the first unit circle crossover could lead to erroneous conclusions.

**Gain Margin.** A second point of particular significance is the gain or magnitude of  $G(j\omega)H(j\omega)$  where it crosses the negative real axis. This is  $\sigma_1$  in Fig. 19. The reciprocal of this value is the *gain margin* of the system. The gain constant of  $G(s)H(s)$  could be raised by a value  $1/\sigma_1$  before instability arose.

The  $-1$  point can be considered a vector of unit magnitude and direction of  $-180$  degrees. Note that the phase margin is defined with relation to the magnitude of the  $-1$  point whereas the gain margin is defined with relation to the direction of the  $-1$  point.

**Conditional Stability, Unconditional Stability.** A *conditionally stable* system is one where instability can come about by either an increase or decrease in system gain. An *unconditionally stable* system is one where

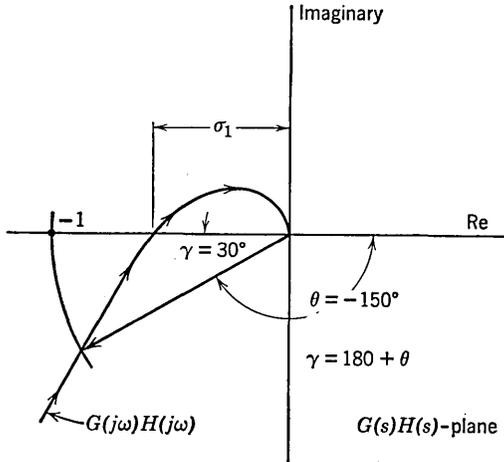


Fig. 19. Determination of phase margin, gain margin.

instability can come about only for an increase in system gain. Figures 20 and 21 illustrate these cases.

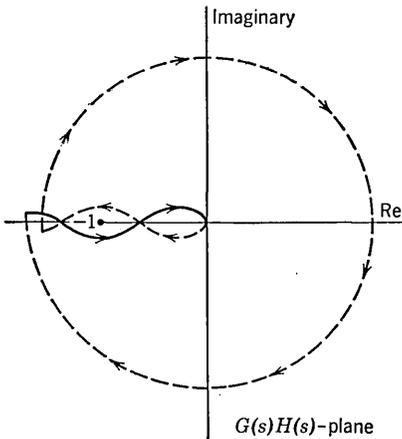


Fig. 20. Conditionally stable system:

$$G(s)H(s) = \frac{K(s + s_1)(s + s_3)}{s^2(s + s_2)(s + s_4)(s + s_6)}$$

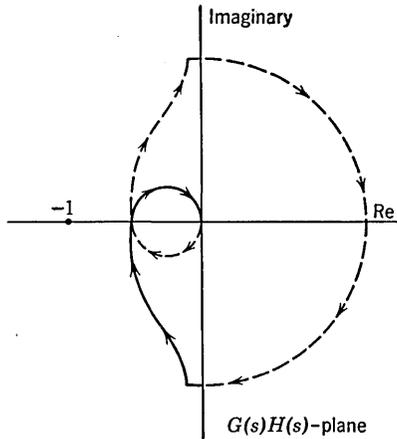


Fig. 21. Unconditionally stable system:

$$G(s)H(s) = \frac{K}{s(s + s_2)(s + s_4)}$$

**Inverse Polar Plots.** The preceding Nyquist diagrams are polar plots of  $G(s)H(s)$ . These diagrams led to ascertainment of the nature of the zeros of eq. (2),

$$1 + G(s)H(s) = 0.$$

If in this equation both sides are divided by  $G(s)H(s)$ ,

$$(24) \quad \frac{1}{G(s)H(s)} + 1 = 0.$$

Let  $G'(s)H'(s)$  represent the inverse of  $G(s)H(s)$

$$(25) \quad G'(s)H'(s) + 1 = 0.$$

The above mathematical manipulations cannot alter the factors of eq. (2). The zeros of eq. (24) or (25) are exactly the same zeros of eq. (2). Investigation of system stability via the inverse polar plot leads to conclusions identical to those arrived at by use of the direct polar plot of  $G(s)H(s)$ . In certain design applications use of the inverse loop transfer function may more clearly demonstrate effects of design changes.

**Polar Plots of Some Common Open Loop Transfer Functions.**

The following plots represent some commonly encountered system functions. Once the reader recognizes how these were generated, he should be ready to handle any newly encountered situation. In the  $G(s)H(s)$ -plane plots are the letters *A*, *B*, *C*. These represent possible locations of the  $-1$  point dependent upon the value of the gain constant *K*. Stability is indicated for various locations of the  $-1$  point. See Figs. 22–36. See also Figs. 13–16.

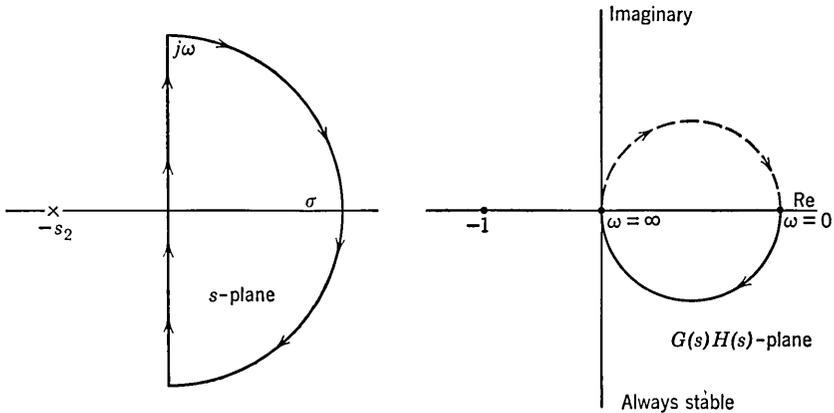


FIG. 22. Polar plot of

$$G(s)H(s) = \frac{K}{(s + s_2)}, \quad P = 0.$$

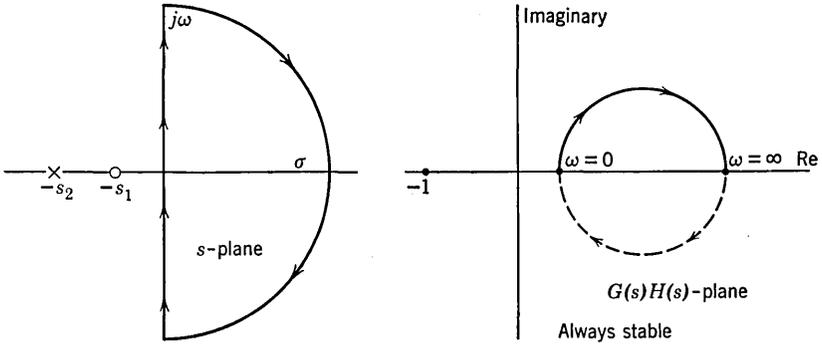


FIG. 23. Polar plot of

$$G(s)H(s) = \frac{K(s + s_1)}{(s + s_2)}, \quad P = 0.$$

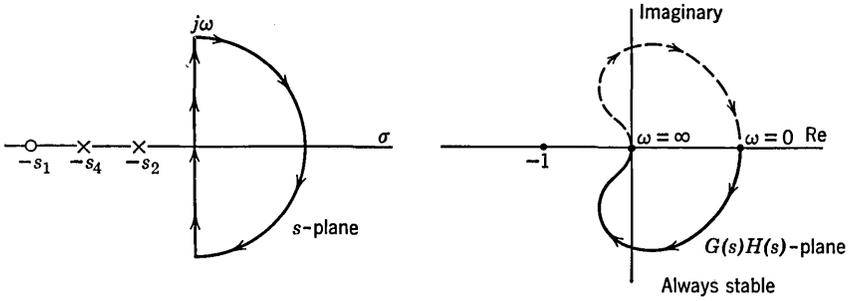


FIG. 24. Polar plot of

$$G(s)H(s) = \frac{K(s + s_1)}{(s + s_2)(s + s_4)}, \quad P = 0.$$

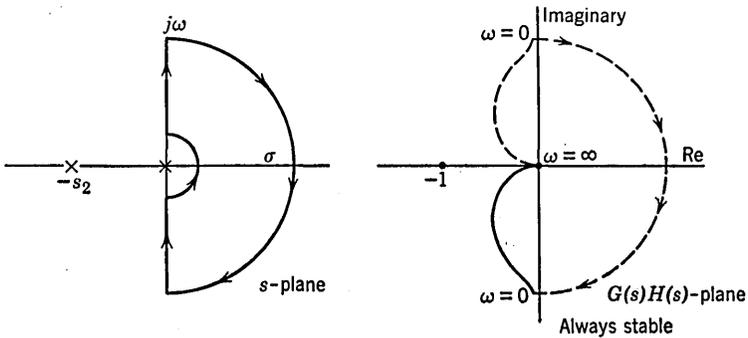


FIG. 25. Polar plot of

$$G(s)H(s) = \frac{K}{s(s + s_2)}, \quad P = 0.$$

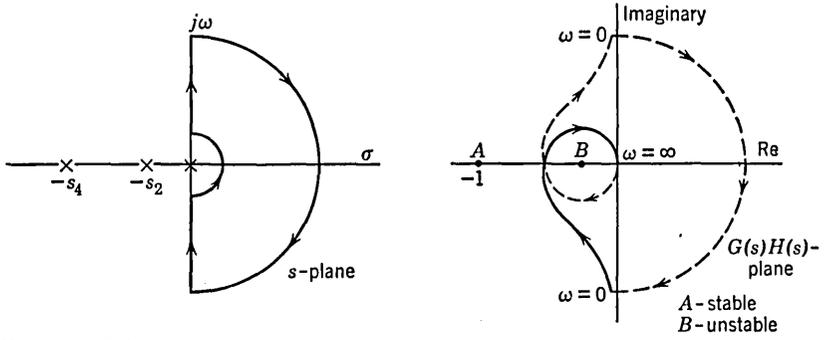


FIG. 26. Polar plot of

$$G(s)H(s) = \frac{K}{s(s + s_2)(s + s_4)}, \quad P = 0.$$

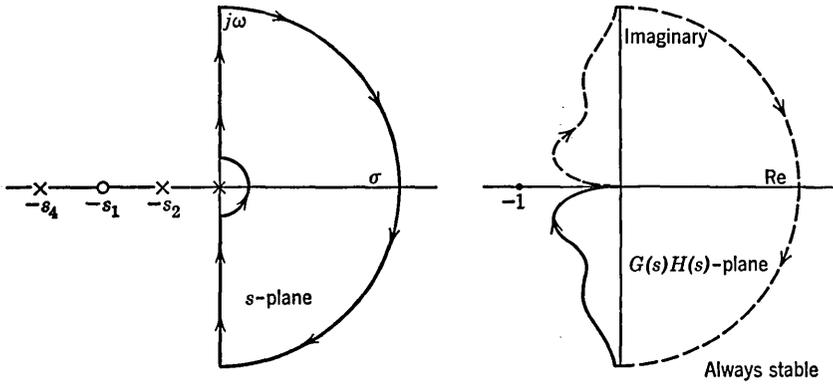


FIG. 27. Polar plot of

$$G(s)H(s) = \frac{K(s + s_1)}{s(s + s_2)(s + s_4)}, \quad P = 0.$$

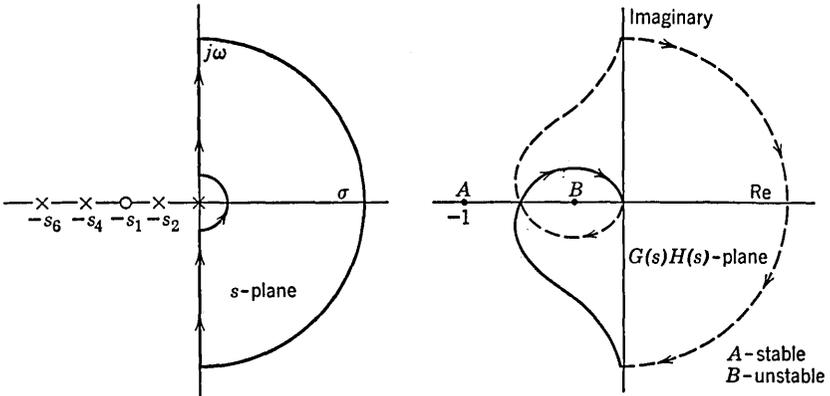


FIG. 28. Polar plot of

$$G(s)H(s) = \frac{K(s + s_1)}{s(s + s_2)(s + s_4)(s + s_6)}, \quad P = 0.$$

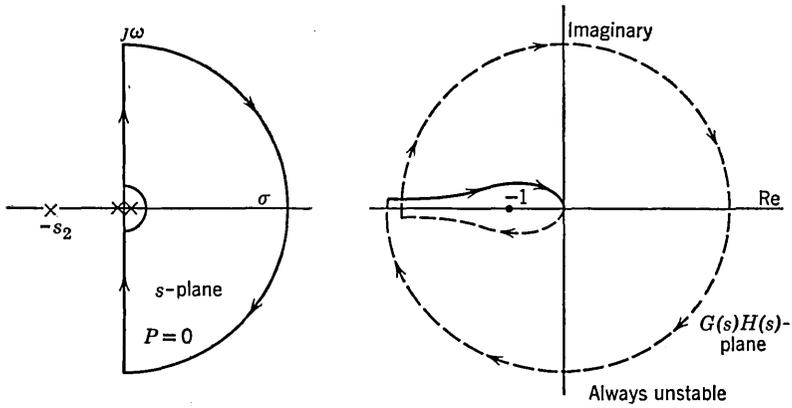


FIG. 29. Polar plot of

$$G(s)H(s) = \frac{K}{s^2(s + s_2)}$$

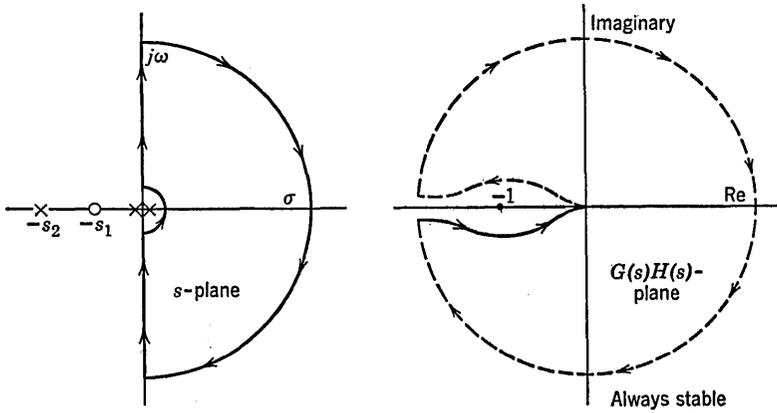


FIG. 30. Polar plot of

$$G(s)H(s) = \frac{K(s + s_1)}{s^2(s + s_2)}, \quad P = 0.$$

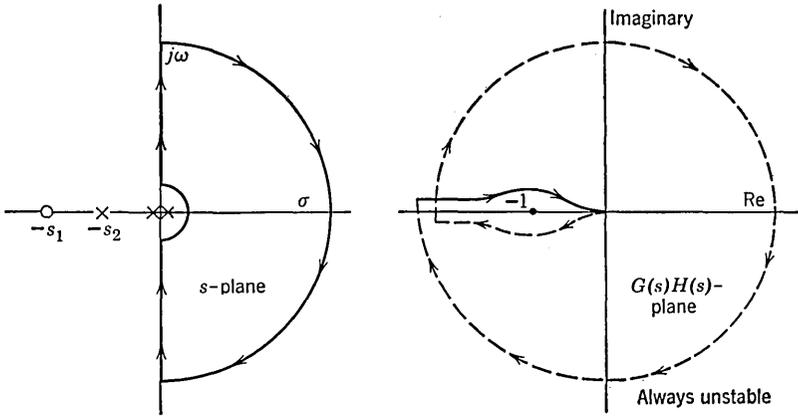


FIG. 31. Polar plot of

$$G(s)H(s) = \frac{K(s + s_1)}{s^2(s + s_2)}, \quad P = 0.$$

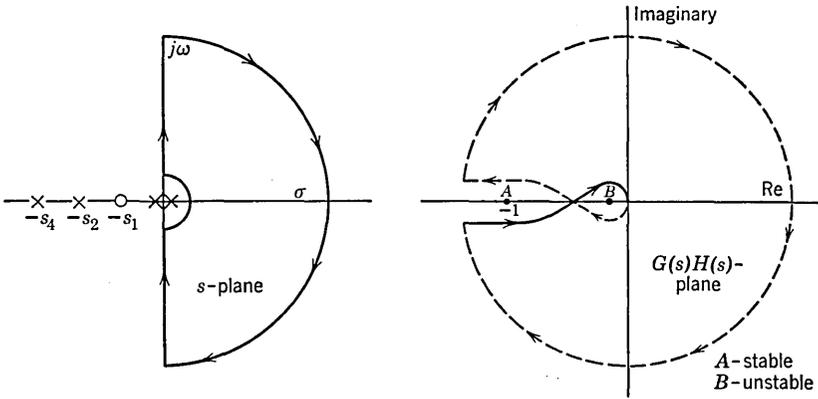


FIG. 32. Polar plot of

$$G(s)H(s) = \frac{K(s + s_1)}{s^2(s + s_2)(s + s_4)}, \quad P = 0.$$

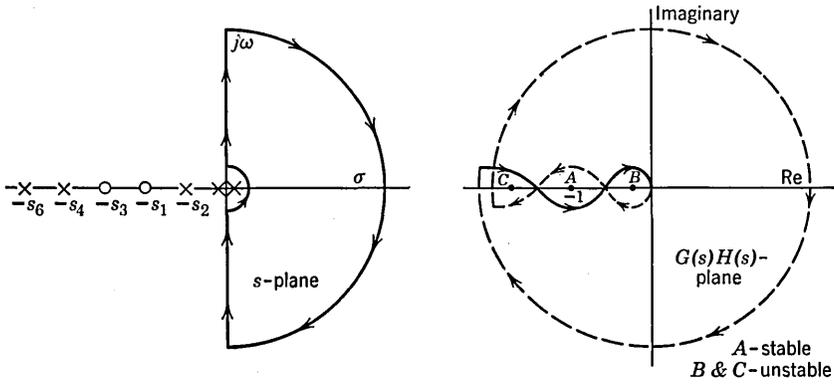


FIG. 33. Polar plot of

$$G(s)H(s) = \frac{K(s + s_1)(s + s_3)}{s^2(s + s_2)(s + s_4)(s + s_6)}, \quad P = 0.$$

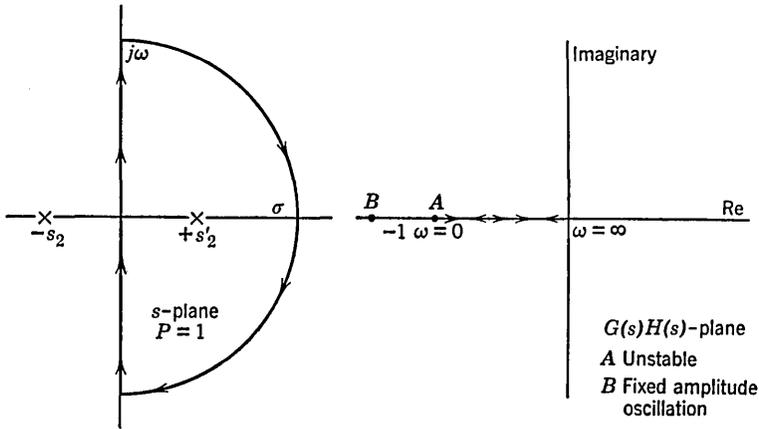


FIG. 34. Polar plot of

$$G(s)H(s) = \frac{K}{(s + s_2)(s + s'_2)}$$

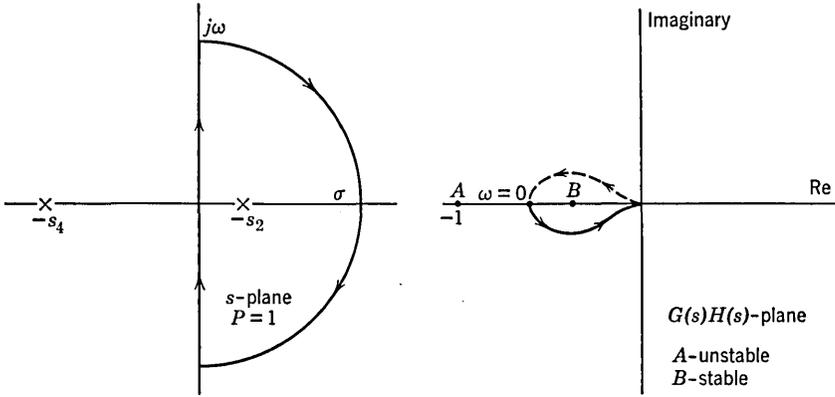


FIG. 35. Polar plot of

$$G(s)H(s) = \frac{K}{(s + s_2)(s + s_4)}$$

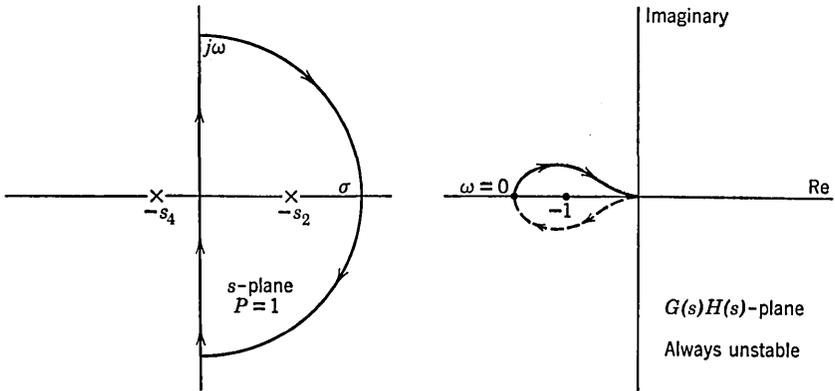


FIG. 36. Polar plot of

$$G(s)H(s) = \frac{K}{(s + s_2)(s + s_4)}$$

**Multiple Input, Multiloop Systems.** A system of this category is shown in Fig. 37.

Stability characteristics of any linear system are independent of where the input or output functions are located. All closed loop transfer functions of a given system have the same characteristic equation. Therefore, when merely investigating stability, one can select one input and one output and proceed to reduce the multiloop system to the basic form shown in Fig. 1. Stability analysis proceeds in the aforementioned fashion.

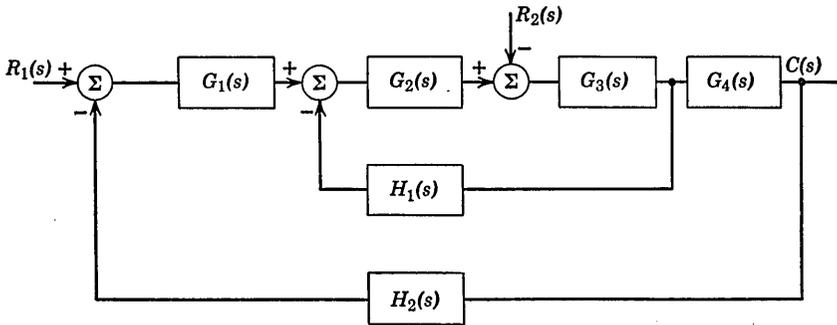


FIG. 37. Multiple input, multiloop system.

It is important not to mislead the reader into believing that transfer characteristics are independent of input and output locations. If  $R_1(s)$  is a reference command and  $R_2(s)$  is a corruption disturbance, it is very possible to design the system so that  $R_1(s)$  is accepted, whereas  $R_2(s)$  is rejected. In essence, this is the purpose of system design as will be shown in the following chapters.

**Relative Merits of the Nyquist Criterion.** This powerful criterion reveals the stability characteristics of practically all linear systems. It also presents a clear indication of ways to improve the system design should this be desired. System design and compensation will be discussed fully in Chap. 23.

Systems defined by linear differential equations whose coefficients are time variant cannot be handled directly by this method. Fortunately, the greatest number of systems encountered in practice are not of this category.

The one major disadvantage of this criterion is the large amount of labor involved in drawing the polar plots or Nyquist diagrams. *When a complete detailed stability analysis is required, the following Bode diagrams should be used to aid in construction of the requisite polar plots.*

## 5. BODE ATTENUATION DIAGRAM APPROACH

This stability analysis procedure is exactly the same as that used in the abbreviated Nyquist criterion except that the information is examined in a different fashion. The information presentation is changed in a way which allows design modification without the large amount of labor attendant in drawing the Nyquist diagrams.

The Nyquist complex plane diagrams (see Sect. 4) contain three pieces of information. The first pertains to the path or contour that the complex variable,  $s$ , traversed. The second and third pieces of information consist of the magnitude and phase of the transform,  $G(s)H(s)$ , for each value of  $s$ . When  $s$  traversed the axis of imaginaries, it took on values of real frequency,  $\omega$ . The corresponding plot of  $G(j\omega)H(j\omega)$  was calibrated with respect to  $\omega$ . See Figs. 6 and 7.

The same information contained in Fig. 7 can be conveyed by means of two diagrams with  $\omega$  a common parameter as in Fig. 38. Because of

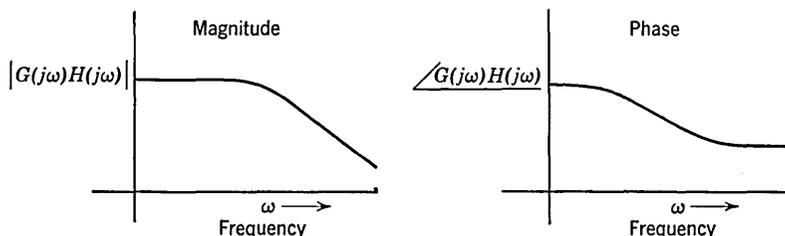


FIG. 38. Bode attenuation diagrams, frequency response curves.

the fundamental work done by Bode (Refs. 7 and 8) in this field, these diagrams are commonly called *Bode diagrams*. Since these diagrams provide information only for  $s$  on the  $j\omega$ -axis, they are true frequency response curves.

**Fundamentals of Bode Diagrams.** Consider first the logarithm of a complex number  $G(j\omega)$ .

$$(26) \quad G(j\omega) = \rho + j\beta$$

$$(27) \quad \log G(j\omega) = \log M(\omega) + j[\theta(\omega) + 2\pi n], \quad n = 0, 1, 2, 3, \dots$$

$M(\omega)$  is the magnitude of the complex number (or vector) and  $\theta(\omega)$  is its phase angle.  $2\pi n$  occurs because an integer number of revolutions causes the vector  $G(j\omega)$  to lie in the same place. The principal value of the logarithm is

$$(28) \quad \log G(j\omega) = \log M(\omega) + j\theta(\omega).$$

If  $G(j\omega)$  consists of the product of two complex numbers,  $G_1(j\omega)$  and  $G_2(j\omega)$ , the logarithm becomes

$$(29) \quad \log G_1(j\omega)G_2(j\omega) = \log M_1(\omega) + \log M_2(\omega) + j[\theta_1(\omega) + \theta_2(\omega)].$$

In other words, the logarithm of the product of two complex numbers is the sum of the individual magnitude logarithms plus  $j$  times the sum of the separate phase angles. Note that the logarithm of a complex number is another complex number. Therefore, in adding complex numbers, real parts add to real parts and imaginary parts add to imaginary parts.

Equation (29) shows that by introducing the logarithm concept, the difficult process of multiplication is replaced by the simpler process of addition.

Equation (28) shows that the log-magnitude  $[\log M(\omega)]$  and the phase  $[\theta(\omega)]$  are separate functions of the common parameter  $\omega$ . They can be plotted separately as in Fig. 38. For minimum phase networks (Ref. 9), the magnitude and phase are so uniquely related (Refs. 7 and 8) that specification of one also specifies the other. The definition of a minimum phase network can be complex (Ref. 9), but a simple "not-all-inclusive" definition is given here. *A minimum phase network is one where for a specified magnitude characteristic, the phase is the minimum possible at all frequencies. In short, the transform of a minimum phase network is one with no poles or zeros in the right half  $s$ -plane.*

**Bode Diagram Stability Analysis.** Consider those well-behaved transfer functions discussed for the abbreviated Nyquist criterion and shown in Fig. 17. Stability was uniquely related to the phase angle at unit circle crossover or to the magnitude of  $G(j\omega)H(j\omega)$  when its phase angle was  $-180$  degrees. To repeat, *for well-behaved, minimum phase  $G(s)H(s)$ , the negative feedback, closed loop system is stable if the phase angle at unit circle crossover is less negative than  $-180$  degrees or whose magnitude is less than unity when the phase angle is  $-180$  degrees.*

To use this important information, it is necessary only to transform the unit circle and the negative real axis into the frequency response diagrams. Whenever  $G(j\omega)H(j\omega)$  lies on the unit circle, its magnitude is equal to 1 regardless of its phase or the frequency involved. The logarithm of this magnitude is zero; therefore the unit circle in the  $G(s)H(s)$ -plane corresponds to the line where  $\log M = 0$  in the log magnitude plot.

In terms of the decibel concept, the unit circle corresponds to the zero-decibel line.

Whenever  $G(j\omega)H(j\omega)$  lies on the negative real axis, its phase is  $-180^\circ$ . Therefore the negative real axis in the  $G(s)H(s)$ -plane corresponds to the  $-180$ -degree line in the phase plot. These important stability landmarks are shown by heavy lines in Fig. 39.

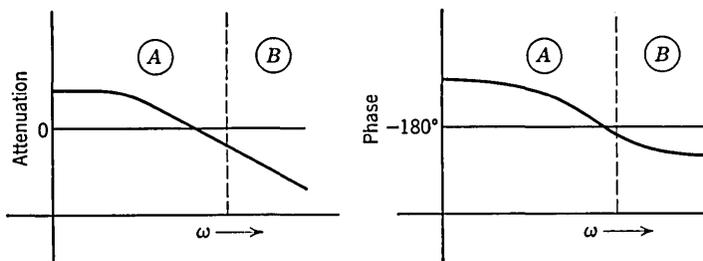


FIG. 39. Minimum phase response curve regions.

**The Nyquist Stability Criterion Rephrased in Terms of the Bode Diagrams.** For well-behaved, minimum phase  $G(s)H(s)$ , the closed loop system is stable if at the frequency where the log magnitude of  $G(j\omega)H(j\omega)$  is equal to zero, its phase angle is less than  $-180$  degrees. Or, the system is stable if at the frequency where the phase angle of  $G(j\omega)H(j\omega)$  is  $-180^\circ$ , the log magnitude is less than zero. If the condition arises where the phase angle is equal to  $-180$  degrees and the log magnitude is zero at a frequency  $\omega_0$ , the closed loop will sustain a constant amplitude oscillation at a frequency  $\omega_0$ . This condition corresponds to  $G(j\omega)H(j\omega)$  passing through the  $-1$  point in the Nyquist plot.

If the phase angle of  $G(j\omega)H(j\omega)$  is defined in terms of phase margin as given by eq. (23) and Fig. 19, the stability criterion is commonly expressed as follows. At gain crossover (the point where the magnitude curve crosses the log  $M = 0$  line) a positive phase margin indicates a stable system whereas a negative phase margin indicates an unstable system.

By use of Bode diagrams it is possible to deduce whether a system is or is not stable in situations more complicated than that wherein  $G(s)H(s)$  is a well-behaved, minimum phase network. When complicated situations arise, final conclusions should be checked by use of the general Nyquist criterion or the Routh criterion.

**Mechanics of Drawing Bode Diagrams.** When given a transform  $G(s)H(s)$ , the most straightforward procedure in drawing Bode diagrams is to pick values of  $j\omega$ , substitute into  $G(j\omega)H(j\omega)$ , and grind out the complex algebra. Fortunately this laborious procedure is not required frequently because  $G(s)H(s)$  is usually known in factored form. There are four basic building blocks used in drawing Bode diagrams.

1.  $K^{\pm 1}$ , a pure gain constant.
2.  $s^{\pm 1}$ , a pure differentiation or pure integration.
3.  $(s + \omega_0)^{\pm 1}$ , a simple lead or simple lag.
4.  $(s^2 + 2\zeta\omega_0s + \omega_0^2)^{\pm 1}$ , a quadratic lead or quadratic lag.

In reverting these basic factors to logarithmic plots it would be entirely possible to use logarithms to the base  $e$  and to use the common multiplier

of 1. Since the decibel concept was in vogue and orders of 10 are easier to handle than orders of  $e$ , logarithms to the base 10 were used and the multiplying factor was taken as 20. A decibel is equal to

$$(30) \quad \text{Decibels} = \text{db} = 10 \log_{10} \frac{P_0}{P_i} = 20 \log_{10} \frac{V_0}{V_i}.$$

Transfer functions in general are more similar to voltage ratios,  $V_0/V_i$ , than to power ratios,  $P_0/P_i$ , therefore the multiplying factor of 20 is commonly used. Some writers (Ref. 10) would rather use the multiplying factor of 10 and units of decilogs, but this seems to be of small consequence in stability analysis.

**The First Building Block: The Pure Gain Constant.**

$$(31) \quad 20 \log K^{\pm 1} = \pm 20 \log K.$$

The logarithm of a pure gain constant is independent of frequency and therefore plots as a horizontal line in the magnitude and phase curves.  $K$  has zero phase angle if it is positive and  $-180$  degrees if it is negative.

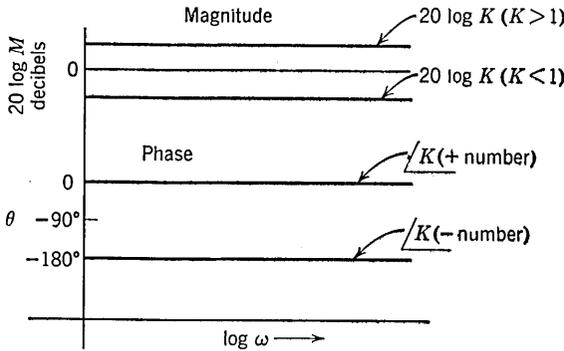


Fig. 40. Bode diagram of a pure gain constant.

It might appear here that logarithms of negative real numbers exist. Note that a purely real logarithm of a negative number is an impossibility, but a complex logarithm of a complex number is possible and given by eqs. (27) and (28). A pure gain constant is a complex number with zero imaginary part. Bode plots of pure gain constants are shown in Fig. 40.

*Note.* Log  $\omega$  is plotted at the bottom of the figure and is common to both curves.

**The Second Building Block: Pure Differentiation or Pure Integration.**

$$(32) \quad 20 \log s^{\pm 1} \Big|_{s=j\omega} = \pm 20 \log \omega \pm j\pi/2.$$

Equation (32) shows that a pure differentiation,  $s$ , introduces a constant phase angle of  $+90$  degrees whereas the pure integration,  $1/s$ , introduces a constant phase angle of  $-90$  degrees.

$\log \omega$  increases by a factor of 1 for every decade change in  $\omega$ ; therefore the log-magnitude changes by 20 db per decade change in  $\omega$ . If the log-magnitude is plotted to a linear scale and  $\omega$  is plotted to a logarithmic scale, the magnitude curves for  $s^{\pm 1}$  consist of straight lines with  $\pm 20$  db

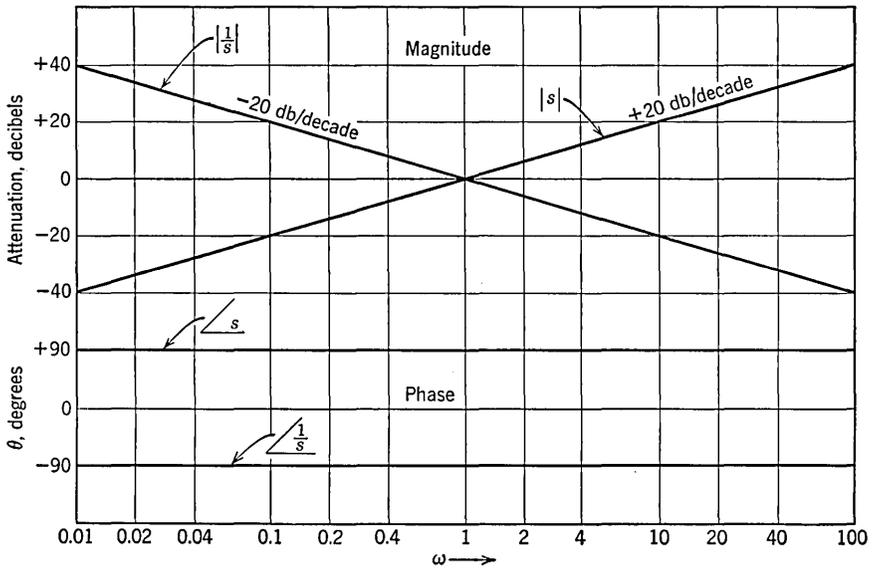


FIG. 41. Bode diagram of  $s^{\pm 1}$  factor.

per decade slopes. Since  $\log 2$  is approximately equal to 0.3 and  $20 \times 0.3 = 6$ , the slope is also commonly referred to as 6 db per octave. These curves are shown in Fig. 41. Semilog paper lends itself very nicely to this application. Notice that the curves pass through 0 db at  $\omega = 1$ . This occurs because

$$(33) \quad 20 \log 1 = 0.$$

**The Third Building Block: Simple Lead or Simple Lag.** This factor is put in nondimensional form by multiplying by  $\omega_0/\omega_0$ .

$$(34) \quad (s + \omega_0)^{\pm 1} = \frac{\omega_0}{\omega_0} (s + \omega_0)^{\pm 1} = \omega_0^{\pm 1} \left( \frac{s}{\omega_0} + 1 \right)^{\pm 1}.$$

Since  $\omega_0$  is a constant, it can be handled as a pure gain factor. The logarithm of the frequency variant part becomes

$$(35) \quad 20 \log \left( \frac{s}{\omega_0} + 1 \right)^{\pm 1} \Big|_{s=j\omega} = \pm 20 \log \sqrt{\left( \frac{\omega}{\omega_0} \right)^2 + 1} + j \tan^{-1} \frac{\omega}{\omega_0}.$$

Consider first the magnitude expression

$$(36) \quad M(\omega) = \pm 20 \log \sqrt{\left( \frac{\omega}{\omega_0} \right)^2 + 1} = \pm 10 \log \left[ \left( \frac{\omega}{\omega_0} \right)^2 + 1 \right].$$

Substitution of values of  $\omega$  in this expression leads to the log magnitude curves drawn in Fig. 42. The curves are drawn for the simple lag. To obtain values for the simple lead, reverse the sign of the ordinate values.

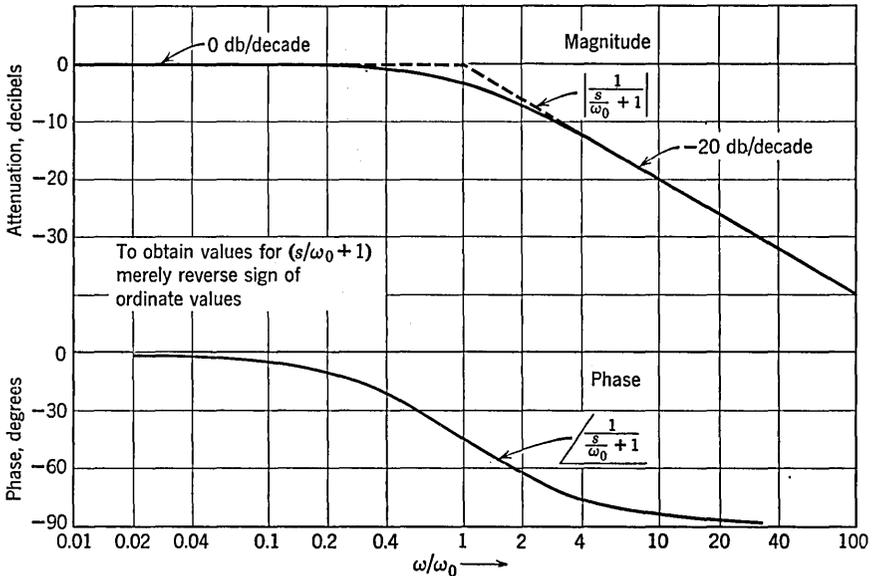


FIG. 42. Bode diagram of  $\frac{1}{[(s/\omega_0) + 1]}$ .

The straight lines are the asymptotes to the log magnitude curves. For  $\omega$  much less than  $\omega_0$ ,  $\omega/\omega_0$  is much less than 1 and can be neglected in eq. (36). Therefore

$$(37) \quad M(\omega) \Big|_{\omega \ll \omega_0} \approx \pm 10 \log 1 = 0 \text{ db/decade.}$$

For  $\omega$  much greater than  $\omega_0$ ,  $\omega/\omega_0$  is much greater than 1. Therefore from eq. (36)

$$(38) \quad M(\omega) \Big|_{\omega \gg \omega_0} \approx \pm 20 \log \frac{\omega}{\omega_0}.$$

This is similar to the expression for a pure differentiation or pure integration and leads to a slope of  $\pm 20$  db per decade change in the non-dimensional parameter  $\omega/\omega_0$ . The point where the asymptote curves meet is at  $\omega/\omega_0 = 1$  and since the slope breaks from 0 db per decade to  $\pm 20$  db per decade at this frequency,  $\omega_0$  is called the *break frequency*. Since the junction resembles a corner, it is also called a *corner frequency*. Note that a pure differentiation or integration corresponds to a factor with a break frequency at  $\omega = 0$ ; therefore all that can be seen on a Bode diagram for this factor is the  $\pm 20$  db per decade slope.

At  $\omega/\omega_0 = 1$  the exact curve lies  $\pm 3$  db away from the asymptote curves. At  $\omega/\omega_0 = 2$  or  $1/2$ , the correction is a  $\pm 1$  db. Normally, it is sufficient to plot in the asymptote lines, the three corrections at  $\omega/\omega_0 = 1/2, 1, 2$  and sketch in the curve freehand.

From eq. (35) the phase angle of a simple lead or lag is

$$(39) \quad \theta(\omega) = \pm \tan^{-1} \frac{\omega}{\omega_0}.$$

For	$\omega = 0$	$\theta(\omega) = 0^\circ,$
	$\omega = \infty$	$\theta(\omega) = \pm 90^\circ,$
	$\omega = \omega_0$	$\theta(\omega) = \pm 45^\circ.$

The phase angle curve can be calculated from eq. (39) and is shown in Fig. 42. To simplify the labor involved, templates can be cut to the  $\theta(\omega)$  shape and used to draw in the phase curve for simple lead or lag factors (commonly referred to as simple breaks). Caution must be exercised in that each template is good only for a given calibration on the log  $\omega$  and  $\theta(\omega)$  axes.

The servomechanisms scale shown in Fig. 43 can also be used to obtain the phase angle of simple breaks. The 45-degree arrowhead is placed at the frequency at which the phase is desired, and the phase is read from the scale at the break frequency. A scale, like a template, is good only in conjunction with the calibration on the log  $\omega$  axis for which it was designed.

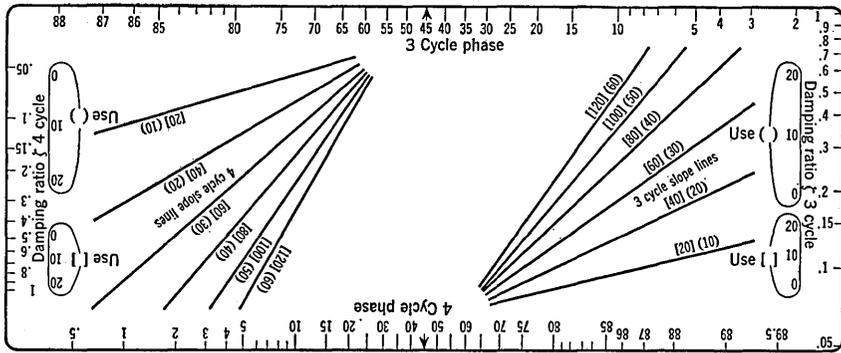


FIG. 43. Plastic servomechanisms scale. (Developed at the General Electric Company by J. J. Hogle and W. E. Sollecito.)

The plastic scale shown in Fig. 43 has scales on the top and bottom sides for 3-cycle and 4-cycle semilog paper for a particular calibration on the logarithmic axis.

**The Fourth Building Block: A Quadratic Lead or Quadratic Lag.**

This factor  $(s^2 + 2\zeta\omega_0s + \omega_0^2)^{\pm 1}$ , in nondimensional form is

$$\begin{aligned}
 (40) \quad & 20 \log \left( \frac{s^2}{\omega_0^2} + 2\zeta \frac{s}{\omega_0} + 1 \right)^{\pm 1} \Big|_{s=j\omega} \\
 & = \pm 20 \log \sqrt{\left( \frac{2\zeta\omega}{\omega_0} \right)^2 + \left( 1 - \frac{\omega^2}{\omega_0^2} \right)^2} \pm j \tan^{-1} \frac{2\zeta\omega_0\omega}{\omega_0^2 - \omega^2}.
 \end{aligned}$$

As evidenced in eq. (40), the magnitude and phase functions are dependent not only on  $\omega$  but also on the damping ratio,  $\zeta$ . For a given damping ratio, there is only one magnitude and one phase curve. The quadratic factor

$$\left( \frac{s^2}{\omega_0^2} + 2\zeta \frac{s}{\omega_0} + 1 \right)^{-1}$$

is plotted in Figs. 44 and 45 as a family of curves with  $\zeta$  an independent parameter. For

$$\left( \frac{s^2}{\omega_0^2} + 2\zeta \frac{s}{\omega_0} + 1 \right)^{+1}$$

the curves would merely be inverted because all values would take on opposite sign as shown by eq. (40).

The magnitude asymptote lines, shown as heavy dashed lines in Fig. 44 meet at  $\omega/\omega_0 = 1$ . This  $\omega_0$  is referred to as a *complex* or *quadratic*

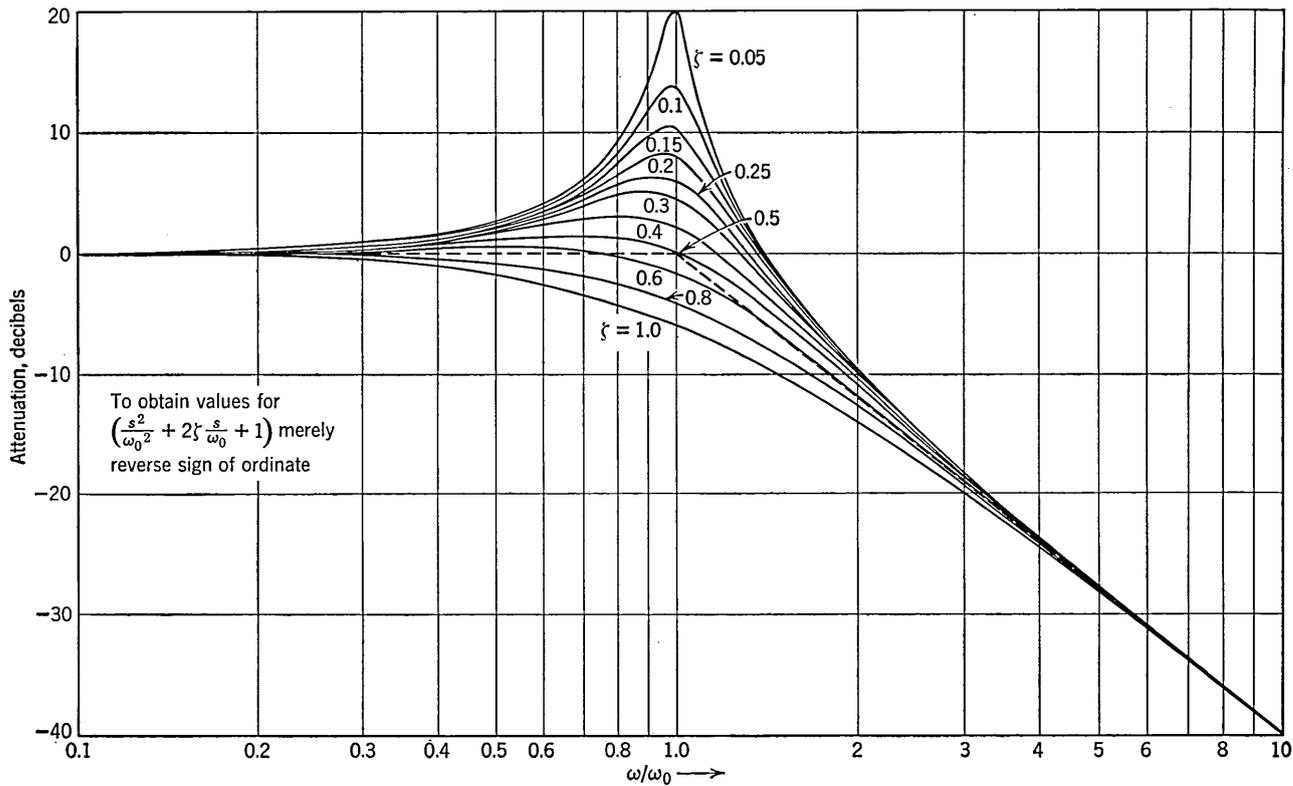


FIG. 44. Magnitude curve of  $[(s^2/\omega_0^2) + 2\zeta(s/\omega_0) + 1]^{-1}$ .

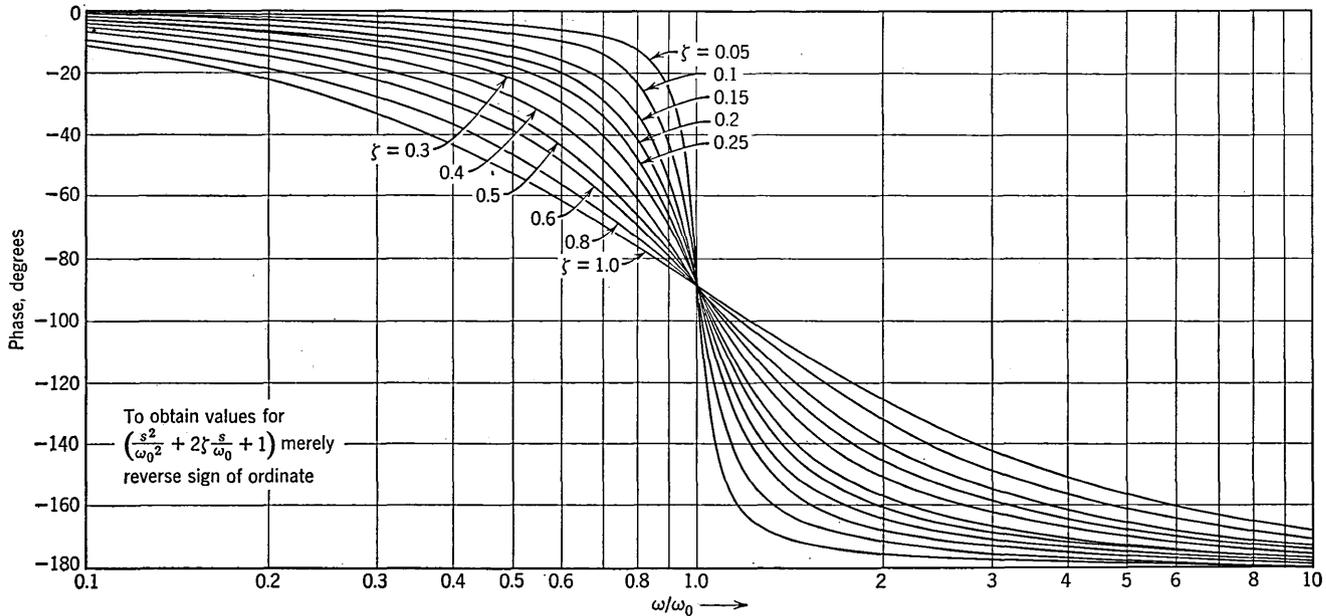


FIG. 45. Phase curves of  $[(s^2/\omega_0^2) + 2\zeta(s/\omega_0) + 1]^{-1}$ .

break frequency. From eq. (40)

$$(41) \quad M(\omega) = \pm 20 \log \sqrt{\left(\frac{2\zeta\omega}{\omega_0}\right)^2 + \left(1 - \frac{\omega^2}{\omega_0^2}\right)^2}.$$

For  $\omega/\omega_0$  much less than 1, all  $\omega/\omega_0$  terms are negligible and

$$(42) \quad M(\omega) \Big|_{\frac{\omega}{\omega_0} \ll 1} \approx \pm 20 \log \sqrt{1} = 0 \text{ db/decade.}$$

For  $\omega/\omega_0$  much greater than 1, the dominant term is

$$(43) \quad M(\omega) \Big|_{\frac{\omega}{\omega_0} \gg 1} \approx \pm 20 \log \sqrt{\left(-\frac{\omega^2}{\omega_0^2}\right)^2} = \pm 40 \log \frac{\omega}{\omega_0}.$$

This is merely twice the slope of a simple break for the similar assumption. Therefore, the asymptote of a complex break is  $\pm 40$  db per decade for large  $\omega/\omega_0$ .

When a quadratic factor is encountered, a first approximation is made by considering  $\zeta = 1$ , which means that a simple break of multiplicity 2 occurs at  $\omega_0$ . For more accurate work, the data in Figs. 44 and 45 must be used. In this case,  $\zeta$  is calculated from the given quadratic factor and the requisite magnitude and phase information obtained from the corresponding  $\zeta$  curves.

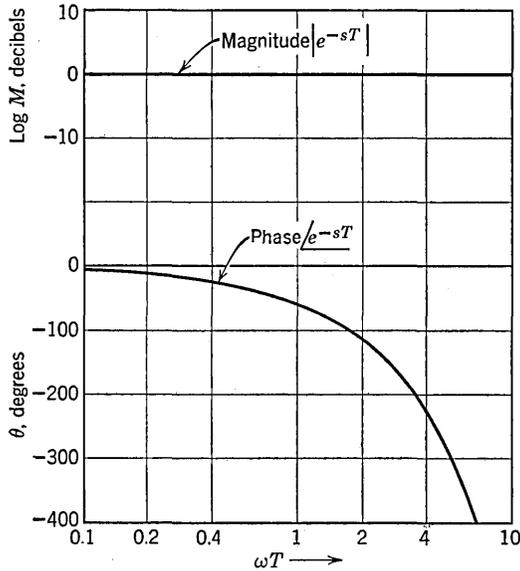
The scale shown in Fig. 43 can be used to obtain phase of quadratic factors by proper use of the additional scales on the right and left sides. Since a graph must be kept for magnitude information, it seems logical to use a graph for phase information also.

It is possible to plot graphs of magnitude correction terms for simple and complex factors to an expanded scale to improve accuracy. Since the corresponding phase correction curves offer small expanded scale possibilities, it is usually of little value.

**The Transform of a Pure Time Delay  $e^{-sT}$ .** This is of particular interest in many cases.

$$(44) \quad 20 \log e^{-sT} \Big|_{s=j\omega} = 20 \log 1 - j\omega T = 0 - j\omega T.$$

Equation (44) shows that the magnitude is independent of frequency and the phase is linearly related to frequency. The magnitude and phase curves are shown in Fig. 46. This function falls within the limitations of the Nyquist criterion, and stability can be investigated in the usual fashion.

FIG. 46. Bode diagram of  $e^{-sT}$ .

### Application of Bode Diagrams.

EXAMPLE 1. Draw the Bode diagram of

$$G(s)H(s) = \frac{316}{s(s+10)}.$$

First put the simple lag factor in nondimensional form

$$G(s)H(s) = \frac{31.6}{s[(s/10) + 1]}.$$

Separate individual factors

$$G(s)H(s) = 31.6 \cdot \frac{1}{s} \cdot \frac{1}{(s/10) + 1}.$$

The asymptotic approximate and the exact Bode diagrams of these individual factors are shown in Fig. 47. The composite  $G(s)H(s)$ , in heavy solid lines, is merely the summation of all the separate magnitude and phase curves as indicated by eq. (29). At gain crossover  $\omega = 16$ ,  $\theta = -147^\circ$  for the exact curve whereas  $\omega = 18$  and  $\theta = -150^\circ$  for the approximate curve. In most cases the approximate answers are sufficiently accurate because in practice the transfer functions represent average values and will not correspond exactly with delivered equipment. Also, as equipment wears in normal use, the transfer characteristics change. For these

reasons the designer must usually provide a margin of safety and some adjustments which will permit small changes when required for improved system performance.

The system shown in Fig. 47 is stable for all values of loop gain because the phase angle approaches  $-180^\circ$  asymptotically. The Nyquist plot of a similar transfer function is shown in Fig. 25. It is well to keep both representations in mind.

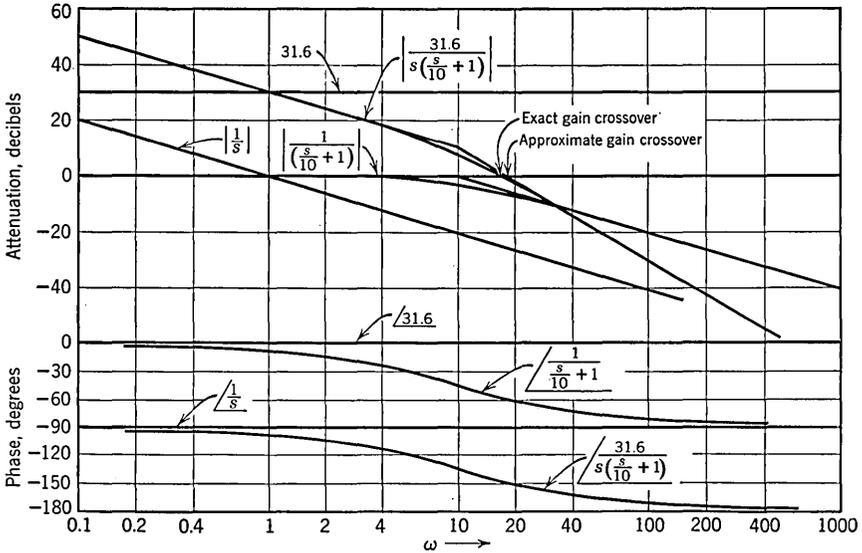


FIG. 47. Bode diagram of  $G(s)H(s) = 31.6 \cdot (1/s) \cdot [1/(s/10) + 1]$ .

EXAMPLE 2. Given the following loop transfer function, determine  $K$  such that gain crossover occurs at  $\theta = -135^\circ$  at  $\omega$  greater than 100 rad/sec.

$$G(s)H(s) = \frac{K(s + 80)}{(s^2 + 6s + 100)(s + 400)}$$

Again, by nondimensionalizing and separating factors

$$G(s)H(s) = \frac{K \cdot 80}{100 \cdot 400} \cdot \frac{1}{[(s/10)^2 + (0.6s/10) + 1]} \cdot \left(\frac{s}{80} + 1\right) \cdot \frac{1}{(s/400) + 1}$$

The frequency response curves are shown in Fig. 48. The composite curves can be drawn without resort to drawing all the individual curves as done in Fig. 47. Neglect the constant term and consider first the

asymptote approximations to the separate factors. There is a quadratic lag break at 10, a simple lead break at 80 and a simple lag break at 400. The approximate curve is flat out to 10, breaks down to  $-40$  db per decade at 10, breaks to  $-20$  db per decade at 80 because of the simple lead, and then breaks back to  $-40$  db per decade at 400 and continues on at this slope. The servomechanism scale shown in Fig. 43 is very useful in drawing these asymptote lines. The exact curve is drawn in by obtaining correction

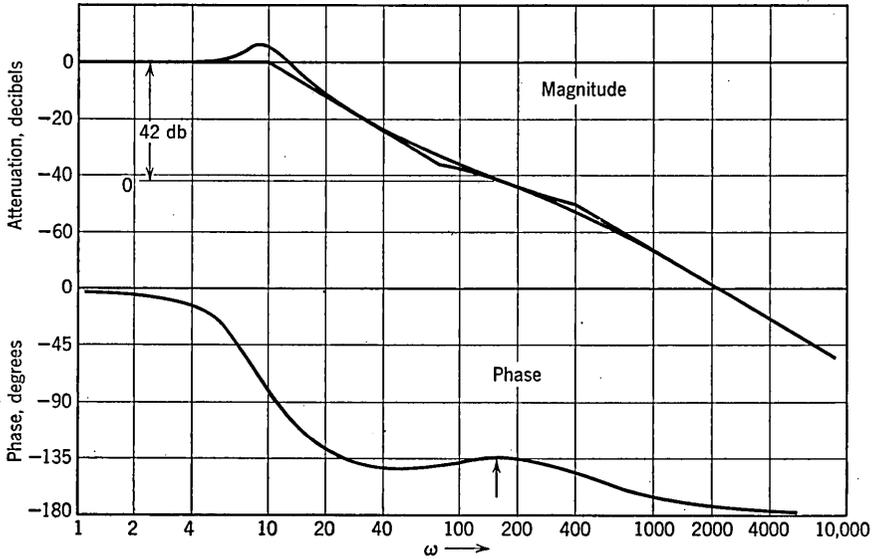


FIG. 48. Bode diagram of  $G(s)H(s) = \frac{K}{500} \frac{(s/80) + 1}{[(s/10)^2 + (0.6s/10) + 1][(s/400) + 1]}$ .

terms for the quadratic lag for  $\zeta = 0.3$  from Fig. 44 and for the simple lead and simple lag from Fig. 42. The exact phase curve is drawn in by use of the servomechanism scale shown in Fig. 43 for the simple breaks and the phase curve for the quadratic lag by use of the phase curve for  $\zeta = 0.3$  in Fig. 45. The arrowhead on the servomechanism scale is placed at the frequency where the phase is desired, and the phase contribution by the simple breaks is read at the break frequencies. The lead and lag terms contribute positive and negative phase angles respectively.

To set the gain constant such that gain crossover occurs at  $\theta = -135^\circ$  at  $\omega$  greater than 100, the entire magnitude curve is shifted up until this occurs. Instead of shifting the magnitude curve up, it is simpler to shift the zero db line down. This corresponds to recalibration of the db axis. The required amount of 0-db line shift corresponds to  $K/500$ . To

meet the requirements of the example  $K/500 = 42$  db.  $K$  therefore must equal 63,000.

**Use of Bode Diagrams in Drawing Nyquist Plots.** When system design is attempted by use of the Nyquist diagrams, it soon becomes apparent that the labor involved in drawing the diagrams is excessive. This comes about because design changes come in terms of multiplying factors which are laborious to incorporate because multiplication is a relatively complex process. The logarithm concept of the Bode diagrams reduces

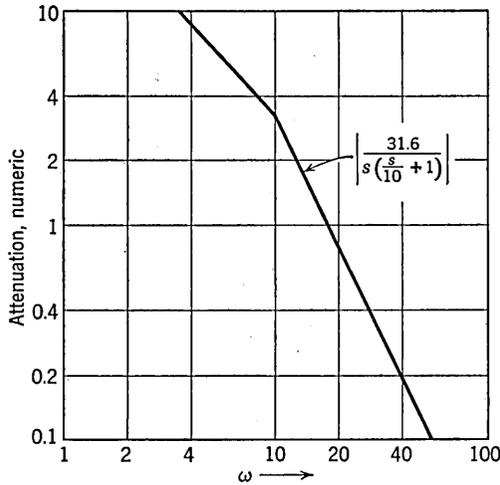


FIG. 49. Reproduction of magnitude of Fig. 47.

multiplication to the simple process of addition. It is advantageous first to plot a Bode diagram and transfer the values from this diagram to the polar plot when information is desired in such form. The major stumbling block to this procedure is the conversion of decibels to gain numbers. It is possible to plot the Bode magnitude diagrams on log-log paper, as shown in Fig. 49 and thereby circumvent the use of decibels. Gain factors are clearly brought to view.

This approach suffers from two major disadvantages. First, the plot of phase is still best accomplished on semilog paper, therefore separate scales would be required for the magnitude and phase curves. Use of decibels allows a single semilog paper to be used for both plots. Second, in adding the magnitudes of two factors, a pair of dividers or some such device would become necessary. Shift of the zero line is not as simple as it is with the decibel scale.

A more useful approach is to use a scale as shown in Fig. 50. The scale is transparent. Three possible scale factors on the decibel scale are avail-

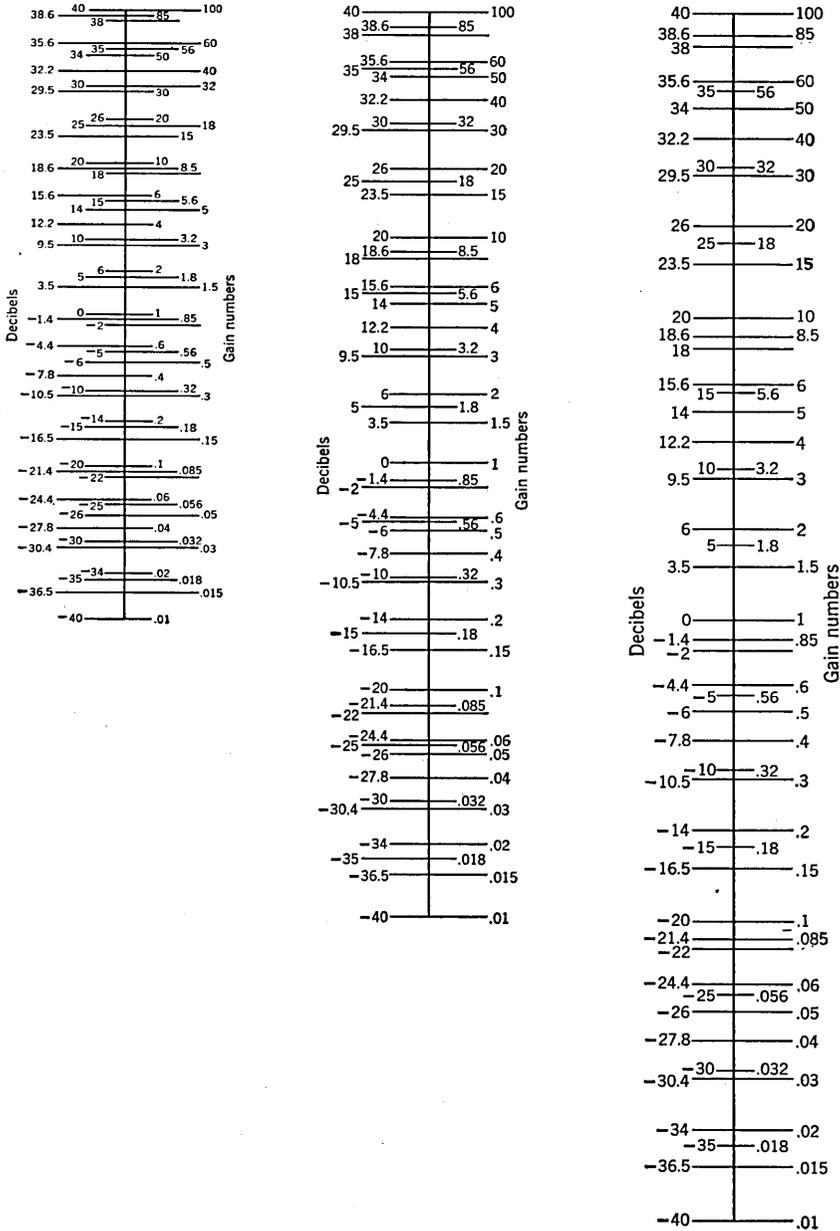


Fig. 50. Gain-decibel conversion scale.

able. The scale is placed vertically on the Bode magnitude diagram and values of gain read directly.

Another approach is that of a graph as shown in Fig. 51. Values of decibels are read off the magnitude curve and the graph is used to convert decibels to gain numbers.

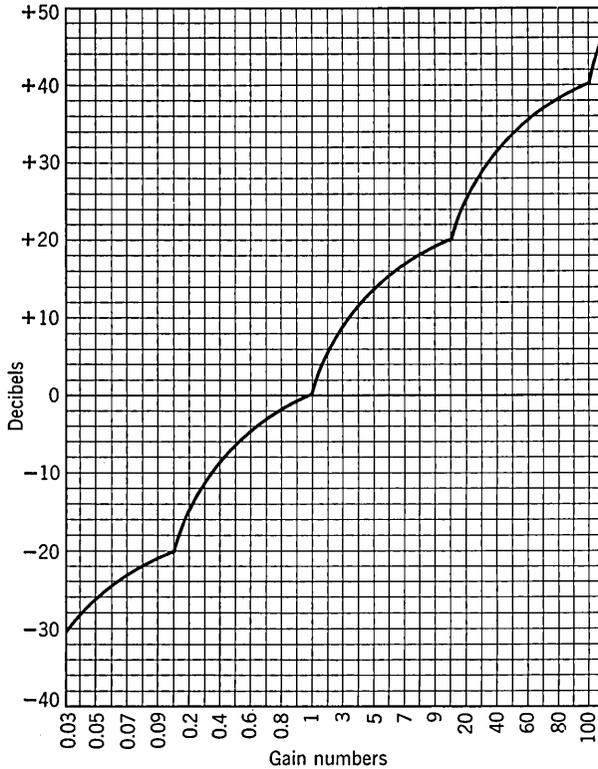


Fig. 51. Gain-decibel conversion chart.

**Relative Merits of the Bode Diagram Approach.** This approach is by far the most generally used method in system design. Design modifications can be analyzed with a minimum of labor involved in drawing the diagrams. The approximate curves allow the designer to investigate a host of designs in a short time. The most promising approximate designs are then investigated more exactly. The host of curves provides an indication of how system performance will change in the presence of some nonlinearities in the system.

This approach is based on the limitations of the abbreviated Nyquist criterion, and when complex situations arise, it is best to revert to the

Nyquist criterion or Routh criterion for exact stability evaluation. The Bode diagrams can be used to draw the Nyquist diagrams.

## 6. ROOT LOCUS METHOD

This method (Refs. 11–13), developed by Evans, provides a means for obtaining the roots of the characteristic equation of a closed loop system, the values of which clearly indicate system stability. Essentially, the method assumes that a chosen complex number is a root of the characteristic equation and tests to see if it can be. If this test is favorable, one constant is changed to a value such that the complex number is a root of the equation. This constant is the loop gain of the closed loop system.

The complex numbers which represent possible roots of the characteristic equation, when plotted in the  $s$ -plane and identified with the necessary corresponding loop gain, form curves which are the loci of the characteristic equation roots. These roots represent poles of the closed loop response which clearly indicate system stability and transient performance.

This plot in the  $s$ -plane provides a rapid evaluation of the effects of varying the gain in a system. It provides a graphical representation of the predominant features of a closed loop system, i.e., its poles, and when system behavior is inadequate, provides a clear indication of proper compensation.

For a system whose loop transfer function is given by

$$G(s)H(s) = \frac{K}{s(T_2s + 1)(T_4s + 1)}$$

the root locus plot is shown in Fig. 52. This plot shows that as the loop

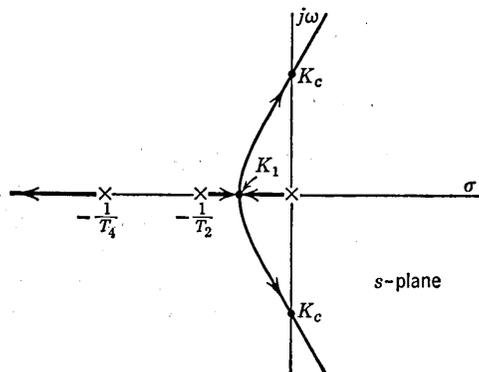


Fig. 52. Root locus for

$$G(s)H(s) = \frac{K}{s(T_2s + 1)(T_4s + 1)}$$

gain,  $K$ , is increased, the closed loop poles move in the direction of the arrowheads. For all values of gain less than  $K_c$ , the closed loop poles lie in the left half  $s$ -plane, and the system is stable. For all values of gain greater than  $K_c$ , the system is unstable.  $K_c$  is a critical gain factor. Also shown by the plot is the fact that, as the gain varies from  $K_1$  to  $K_c$ , the closed loop poles have a decreasing damping factor. Therefore, one can expect the transient response to be more oscillatory and to have a longer settling time as the gain is increased in this region.

**Theory of Construction.** Consider the general negative feedback system shown in Fig. 1.

Assuming that the feed forward and feedback transfer functions are composed of fractions of rational polynomials in  $s$ , i.e.,

$$(45) \quad G(s) = \frac{a_{m_1}s^{m_1} + a_{m_1-1}s^{m_1-1} + \dots + a_1s + a_0}{b_{n_1}s^{n_1} + b_{n_1-1}s^{n_1-1} + \dots + b_1s + b_0} = \frac{N_1(s)}{D_1(s)},$$

$$(46) \quad H(s) = \frac{d_{m_2}s^{m_2} + d_{m_2-1}s^{m_2-1} + \dots + d_1s + d_0}{e_{n_2}s^{n_2} + e_{n_2-1}s^{n_2-1} + \dots + e_1s + e_0} = \frac{N_2(s)}{D_2(s)}.$$

The closed loop response is

$$(47) \quad \frac{C(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)} = \frac{N_1(s)D_2(s)}{D_1(s)D_2(s) + N_1(s)N_2(s)}.$$

The root locus method obtains the roots of the fractional equation  $[1 + G(s)H(s) = 0]$ , which are the roots of the characteristic function  $[D_1(s)D_2(s) + N_1(s)N_2(s)]$ . It is of interest to note that the closed loop response has numerator factors (zeros) which are identical to the zeros of the feed forward transfer function and poles of the feedback transfer function,  $N_1(s)$  and  $D_2(s)$  respectively.

To find the closed loop poles

$$1 + G(s)H(s) = 0.$$

Therefore

$$(48) \quad G(s)H(s) = -1 = 1/\underline{\pm N\pi}, \quad N = 1, 3, 5, 7, \dots$$

For this identity to exist, the angle of  $G(s)H(s)$  must lie along the negative real axis of the  $s$ -plane. *This constitutes the angle condition:*

$$(49) \quad \underline{G(s)H(s)} = \pm N\pi, \quad N = 1, 3, 5, 7, \dots$$

Also, the magnitude of  $G(s)H(s)$  must be unity. *This constitutes the magnitude condition:*

$$(50) \quad |G(s)H(s)| = 1.$$

In general,

$$(51) \quad G(s)H(s) = \frac{K(s + s_1)(s + s_3)(s + s_5) \cdots (s + s_{2m-1})}{(s + s_2)(s + s_4)(s + s_6) \cdots (s + s_{2n})},$$

where

$$(52) \quad \frac{K(s_1)(s_3)(s_5) \cdots (s_{2m-1})}{(s_2)(s_4)(s_6) \cdots (s_{2n})}$$

represents the loop gain. Each factor in eq. (51) represents a vector in the  $s$ -plane as shown in Fig. 53.

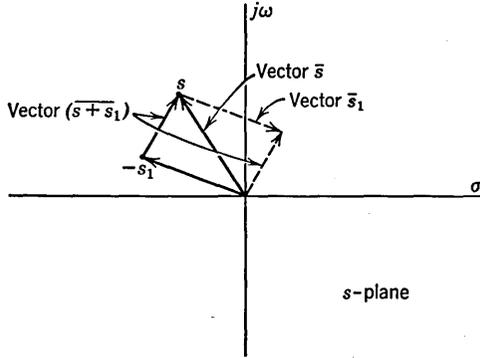


Fig. 53. Vector representation of typical polynomial factors.

The angle condition, eq. (49), requires that

(53)

$$\left[ \angle \frac{1}{s + s_1} + \angle \frac{1}{s + s_3} + \cdots + \angle \frac{1}{s + s_{2m-1}} \right] - \left[ \angle \frac{1}{s + s_2} + \cdots + \angle \frac{1}{s + s_{2n}} \right] = \pm N\pi$$

or

$$(54) \quad \sum_{k=1}^{k=m} \angle \frac{1}{(s + s_{2k-1})} - \sum_{i=1}^{i=n} \angle \frac{1}{(s + s_{2i})} = \pm N\pi.$$

This states that for an exploratory point,  $s$ , to lie on the root locus, the summation of the angles of the zeros minus the summation of the angles of the poles of the open loop response *must be an odd multiple of  $\pi$* .

**Procedure.** The procedure is to plot the poles and zeros of  $G(s)H(s)$  in the  $s$ -plane, choose an exploratory point,  $s$ , lay off the factor vectors [note that the factor vector arrowheads always lie on the exploratory point  $s$  in Fig. 53], sum the angles with the proper sense; if they add up to  $\pm N\pi$ , the point is on the root locus. If not, move the exploratory point over and repeat. This constitutes assuming that a chosen complex number is a root of the characteristic equation and testing to see if it can be, i.e., if it satisfies the angle condition.

When a point is located that does satisfy the angle condition, then the vector magnitudes are measured and values are substituted in the magnitude condition, eq. (50), to calibrate the constant  $K$ . [Note.  $K$  determines loop gain, but it is not defined as such per eq. (52).]

$$(55) \quad |K| = \frac{|(s + s_2)| |(s + s_4)| \cdots |(s + s_{2n})|}{|(s + s_1)| |(s + s_3)| \cdots |(s + s_{2m-1})|}$$

Repetition of the above steps should ascertain the complete locus.

When the locus is completed, the actual  $K$  of the given system is determined from  $G(s)H(s)$ . By reference to the root locus plot, the closed loop poles are then obtained by inspection.

**General Theorems for Construction.** At first glance, it appears as though a root locus may lie anywhere in the complex plane and to discover it may be a hit-or-miss proposition. Fortunately, the locus must take on certain definite patterns governed by the number and location of the open loop poles and zeros. The following general theorems aid in ascertaining the approximate root locus.

**THEOREM 1.** *Number of branches of the locus is equal to the number of closed loop poles.* A branch is a separate portion of the root locus which has all values of loop gain on it. For a given loop gain only one pole may exist on one branch of the complete root locus plot. The number of branches is therefore equal to the degree of the characteristic equation in  $s$  because this determines the total number of poles. Reference to Fig. 54 shows that for a given loop gain,  $K_1$ , there are four closed loop poles, one on each of the four branches labeled (1), (2), (3), (4).  $K$  is the loop gain because all factors are in nondimensional form.

**THEOREM 2.** *The locus starts at open loop poles or infinity ( $K = 0$ ) and ends at open loop zeros or infinity ( $K = \infty$ ).* Inspection of the magnitude eq. (55) shows that at open loop poles  $K$  is zero because one of the numerator magnitudes becomes zero.  $K$  is infinite at open loop zeros because a zero magnitude term appears in the denominator. For the locus to start at infinity it is imperative that  $G(s)H(s)$  have more zeros than poles, i.e., its numerator would be of higher degree than its denominator. Equation (55) shows that for this case and for  $s$  approaching infinity,  $K$  approaches zero. For the locus to end at infinity, it is imperative that  $G(s)H(s)$  have more poles than zeros. In this case eq. (55) shows that for  $s$  approaching infinity  $K$  approaches infinity also. Figure 54 shows the case where the loop transfer function has four poles and one zero. The branches start at the poles for a loop gain of zero. As the loop gain increases to infinity, branch (2) goes along the real axis from the pole to the zero while the other three branches tend toward infinity.

**THEOREM 3.** For the locus to exist on the real axis, the sum of poles and zeros to the right of the exploratory point must be odd. This is so because conjugate complex roots together contribute zero angle when the exploratory point is on the real axis. Only poles and zeros on the real axis to the right of the exploratory point contribute angle ( $180^\circ$  each), there-

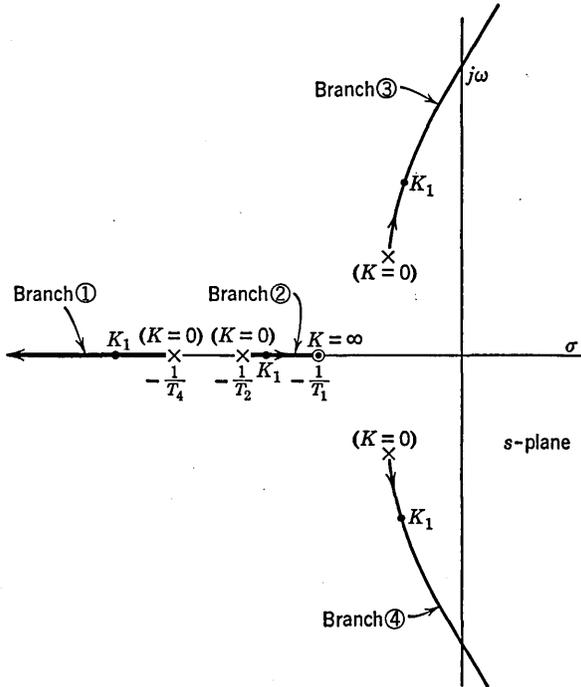


FIG. 54. Root locus plot for

$$G(s)H(s) = \frac{K(T_1s + 1)}{(T_2s + 1)(T_4s + 1)[(s^2/\omega_0^2) + 2\zeta(s/\omega_0) + 1]}$$

fore, the above conclusion. By again referring to Fig. 54 it is seen that where the locus exists on the real axis there are either five or three poles and zeros to the right.

**THEOREM 4.** The locus is symmetrical with respect to the real axis. The characteristic equation is a rational polynomial in  $s$  with real coefficients. Therefore, the roots, when complex, must occur as conjugate pairs. In Fig. 54, branch (3) is the image of branch (4) about the real axis.

**THEOREM 5.** The locus leaves an open loop pole or approaches an open loop zero in the direction given by  $\pm N\pi$  minus the sum of angles of vectors from remaining poles and zeros to the pole or zero in question. Consider the exploratory point  $s$  to be very close to an open loop pole. As  $s$  circles the



When the exploratory point is extremely far from the cluster of open loop poles and zeros, they all contribute essentially the same amount of angle. Since these must add up to  $\pm 180$  degrees or some odd multiple, the foregoing conclusion exists. In Fig. 55 it is seen that  $\pm 60^\circ$  and  $180^\circ$  are the directions of the asymptote lines. Part of the 180-degree line also happens to be a branch of the locus.

**THEOREM 7.** *Asymptote lines cross real axis at*

$$\frac{\sum_{i=1}^{i=n} \sigma_i - \sum_{k=1}^{k=m} \sigma_k}{n - m} \quad \begin{array}{l} \sigma_i = \text{real part of } i\text{th poles,} \\ \sigma_k = \text{real part of } k\text{th zero.} \end{array}$$

This corresponds to the centroid of the pole-zero cluster. In Fig. 55 the asymptote lines cross the axis at  $-7$ .

**Practical Considerations.** If the loop transfer function has many poles and zeros, some of which are located relatively far from the main cluster and from the  $j\omega$ -axis, a first order approximation can be made to the exact locus by omitting the distant poles and zeros. This procedure requires good engineering judgment. The advantage lies in quicker ascertainment of the important part of the locus which can be drawn to an expanded scale.

The procedure to be used in drawing a root locus is to plot the poles and zeros of the open loop response. From the preceding generalizations, sketch in the loci. Graphically determine the exact loci. With the open loop gain constant pick off the closed loop poles. The closed loop response is then made up of the poles obtained from the plot and the zeros and multiplying constant from inspection of  $G(s)$  and  $H(s)$  per eq. (47). Multiple loops are handled by first reducing the minor loops to transfer functions in factored form. It is of interest to remember here that the root locus method is a graphical procedure of factoring the characteristic equation of a system. The minor loop transfer functions are then included as blocks in the major loop and the major loop root locus is then plotted.

**Donahue's Analytical Procedure to Calculate the Root Loci.** A relatively simple analytical means of plotting a root locus has been developed (Ref. 34). This method determines a point on the locus by shifting the  $j\omega$ -axis a given distance,  $\sigma_1$ , and then calculating the frequency,  $\omega_1$ , at which the locus crosses the  $j\omega$ -axis in the  $s$ -plane. The requisite loop gain is then calculated from  $\omega_1$ . Successive points on the locus are obtained by successive shifting of the  $j\omega$ -axis. Tables 1 and 2 have been derived by Donahue to aid in the calculations (Ref. 34).

Referring to the general single loop negative feedback system of Fig. 1, let

$$(56) \quad G(s)H(s) = \frac{K(R_N + I_N)}{R_D + I_D},$$

where  $R_N$  is the real part of the numerator (even powers of  $s$ ) and  $I_N$  is the imaginary part of the numerator (odd powers of  $s$ ). The denominator follows similar notation.

For a point to lie on the root locus, from eq. (48),

$$(57) \quad \frac{K(R_N + I_N)}{R_D + I_D} = -1.$$

Therefore

$$(58) \quad KR_N + KI_N = -R_D - I_D.$$

Equating real part to real part and imaginary to imaginary

$$(59) \quad K = -\frac{R_D}{R_N} = -\frac{I_D}{I_N}$$

and

$$(60) \quad R_D I_N = R_N I_D.$$

Equations (59) and (60) provide means of solution for frequency and gain at the root locus crossing of the  $j\omega$ -axis.

EXAMPLE. Let

$$(61) \quad G(s) = \frac{K(s + s_1)}{(s + s_2)}, \quad H(s) = \frac{1}{(s + s_4)},$$

$$(62) \quad G(s)H(s) = \frac{K(s + s_1)}{(s + s_2)(s + s_4)} = \frac{K(s + h_0)}{s^2 + r_1s + r_0},$$

$$\text{where} \quad h_0 = s_1, \quad r_1 = s_2 + s_4, \quad r_0 = s_2s_4$$

from the preceding

$$R_N = h_0, \quad I_N = s, \quad R_D = s^2 + r_0, \quad I_D = r_1s.$$

Substitution in eqs. (59) and (60) gives

$$(63) \quad K = -\frac{(s^2 + r_0)}{h_0},$$

$$(64) \quad (s^2 + r_0)s = h_0r_1s.$$

TABLE 1. AID FOR ROOT LOCUS CAL

$(m,n)$	General Form	$a_0$	$a_1$
(1,2)	$\frac{K(s+h_0)}{(s^2+r_1s+r_0)}$	$r_0+r_1\sigma+\sigma^2$	$r_1+2\sigma$
(2,2)	$\frac{K(s^2+h_1s+h_0)}{(s^2+r_1s+r_0)}$	$r_0+r_1\sigma+\sigma^2$	$r_1+2\sigma$
(0,3)	$\frac{K}{(s^3+r_2s^2+r_1s+r_0)}$	$r_0+r_1\sigma+r_2\sigma^2+\sigma^3$	$r_1+2r_2\sigma+3\sigma^2$
(1,3)	$\frac{K(s+h_0)}{(s^3+r_2s^2+r_1s+r_0)}$	$r_0+r_1\sigma+r_2\sigma^2+\sigma^3$	$r_1+2r_2\sigma+3\sigma^2$
(2,3)	$\frac{K(s^2+h_1s+h_0)}{(s^3+r_2s^2+r_1s+r_0)}$	$r_0+r_1\sigma+r_2\sigma^2+\sigma^3$	$r_1+2r_2\sigma+3\sigma^2$
(3,3)	$\frac{K(s^3+h_2s^2+h_1s+h_0)}{(s^3+r_2s^2+r_1s+r_0)}$	$r_0+r_1\sigma+r_2\sigma^2+\sigma^3$	$r_1+2r_2\sigma+3\sigma^2$
(0,4)	$\frac{K}{(s^4+r_3s^3+r_2s^2+r_1s+r_0)}$	$r_0+r_1\sigma+r_2\sigma^2+r_3\sigma^3+\sigma^4$	$r_1+2r_2\sigma+3r_3\sigma^2+4\sigma^3$
(1,4)	$\frac{K(s+h_0)}{(s^4+r_3s^3+r_2s^2+r_1s+r_0)}$	$r_0+r_1\sigma+r_2\sigma^2+r_3\sigma^3+\sigma^4$	$r_1+2r_2\sigma+3r_3\sigma^2+4\sigma^3$
(2,4)	$\frac{K(s^2+h_1s+h_0)}{(s^4+r_3s^3+r_2s^2+r_1s+r_0)}$	$r_0+r_1\sigma+r_2\sigma^2+r_3\sigma^3+\sigma^4$	$r_1+2r_2\sigma+3r_3\sigma^2+4\sigma^3$
(0,5)	$\frac{K}{(s^5+r_4s^4+r_3s^3+r_2s^2+r_1s+r_0)}$	$r_0+r_1\sigma+r_2\sigma^2+r_3\sigma^3+r_4\sigma^4+\sigma^5$	$r_1+2r_2\sigma+3r_3\sigma^2+4r_4\sigma^3+5\sigma^4$
(1,5)	$\frac{K(s+h_0)}{(s^5+r_4s^4+r_3s^3+r_2s^2+r_1s+r_0)}$	$r_0+r_1\sigma+r_2\sigma^2+r_3\sigma^3+r_4\sigma^4+\sigma^5$	$r_1+2r_2\sigma+3r_3\sigma^2+4r_4\sigma^3+5\sigma^4$
(0,6)	$\frac{K}{(s^6+r_5s^5+r_4s^4+r_3s^3+r_2s^2+r_1s+r_0)}$	$r_0+r_1\sigma+r_2\sigma^2+r_3\sigma^3+r_4\sigma^4+r_5\sigma^5+\sigma^6$	$r_1+2r_2\sigma+3r_3\sigma^2+4r_4\sigma^3+5r_5\sigma^4+6\sigma^5$
$(m,n)$	$\frac{K \left[ \sum_{i=0}^m h_i s^i \right]}{\left[ \sum_{i=0}^n r_i s^i \right]}$	$\left[ \sum_{i=0}^n r_i \sigma^i \right]$	$\left[ \sum_{i=1}^n i r_i \sigma^{(i-1)} \right]$

CULATION BY DONAHUE PROCEDURE

$a_2$	$a_3$	$a_4$	$a_5$	$b_0$	$b_1$	$b_2$
0	0	0	0	$h_0 + \sigma$	0	0
0	0	0	0	$h_0 + h_1\sigma$ $+ \sigma^2$	$h_1 + 2\sigma$	0
$r_2 + 3\sigma$	0	0	0	0	0	0
$r_2 + 3\sigma$	0	0	0	$h_0 + \sigma$	0	0
$r_2 + 3\sigma$	0	0	0	$h_0 + h_1\sigma$ $+ \sigma^2$	$h_1 + 2\sigma$	0
$r_2 + 3\sigma$	0	0	0	$h_0 + h_1\sigma$ $+ h_2\sigma^2 + \sigma^3$	$h_1 + 2h_2\sigma$ $+ 3\sigma^2$	$h_2 + 3\sigma$
$r_2 + 3r_3\sigma + 6\sigma^2$	$r_3 + 4\sigma$	0	0	0	0	0
$r_2 + 3r_3\sigma + 6\sigma^2$	$r_3 + 4\sigma$	0	0	$h_0 + \sigma$	0	0
$r_2 + 3r_3\sigma + 6\sigma^2$	$r_3 + 4\sigma$	0	0	$h_0 + h_1\sigma$ $+ \sigma^2$	$h_1 + 2\sigma$	0
$r_2 + 3r_3\sigma$ $+ 6r_4\sigma^2 + 10\sigma^3$	$r_3 + 4r_4\sigma$ $+ 10\sigma^2$	$r_4 + 5\sigma$	0	0	0	0
$r_2 + 3r_3\sigma$ $+ 6r_4\sigma^2 + 10\sigma^3$	$r_3 + 4r_4\sigma$ $+ 10\sigma^2$	$r_4 + 5\sigma$	0	$h_0 + \sigma$	0	0
$r_2 + 3r_3\sigma + 6r_4\sigma^2$ $+ 10r_5\sigma^3 + 15\sigma^4$	$r_3 + 4r_4\sigma$ $+ 10\sigma^2 + 20\sigma^3$	$r_4 + 5r_5\sigma$ $+ 15\sigma^2$	$r_5$ $+ 6\sigma$	0	0	0

$$\left[ \sum_{i=2}^n \left( \sum_{k=1}^{i-1} k \right) r_i \sigma^{(i-2)} \right]$$

$$\left[ \sum_{i=0}^m h_i \sigma^i \right] \left[ \sum_{i=1}^n i h_i \sigma^{(i-1)} \right] \left[ \sum_{i=2}^m \left( \sum_{k=1}^{i-1} k \right) h_i \sigma^{(i-2)} \right]$$

In eqs. (63) and (64) let  $s = j\omega$

$$(65) \quad K = \frac{\omega^2 - r_0}{h_0},$$

$$(66) \quad \omega^2 = r_0 - r_1 h_0.$$

Equations (65) and (66) define the gain and frequency at which the root locus crosses the imaginary axis.

To calculate other points on the locus, shift the  $j\omega$ -axis by replacing  $s$  with  $(s + \sigma)$

$$(67) \quad G(s + \sigma)H(s + \sigma) = \frac{K(s + \sigma + h_0)}{(s + \sigma)^2 + r_1(s + \sigma) + r_0},$$

which reduces to

$$(68) \quad G(s + \sigma)H(s + \sigma) = \frac{K(s + b_0)}{s^2 + a_1 s + a_0},$$

where

$$a_0 = r_0 + \sigma r_1 + \sigma^2, \quad a_1 = r_1 + 2\sigma, \quad b_0 = h_0 + \sigma.$$

By analogy to eqs. (62), (65), and (66)

$$(69) \quad -K = \frac{1}{b_0} (a_0 - \omega^2),$$

$$(70) \quad \omega^2 = a_0 - a_1 b_0.$$

**Tables for Donahue Procedure.** This example gives rise to the first row of Tables 1 and 2. For  $\sigma = 0$  eqs. (69) and (70) reduce to eqs. (65) and (66). Table 1 has the parameters  $a_0$ ,  $a_1$ ,  $a_2$ , etc., and  $b_0$ ,  $b_1$ ,  $b_2$ , etc., determined in terms of the original numerator and denominator power series coefficients for each of several types of loop transfer functions. Table 2 gives  $\omega^2$  and  $-K$  in terms of these  $a$  and  $b$  parameters.

The procedure therefore consists of reducing  $G(s)H(s)$  to a fraction of two power series, identifying this with the proper row in Table 1, substituting in values of  $\sigma$ , which lead to calculation of the parameters  $a$  and  $b$  and subsequent solution of  $\omega^2$  and  $-K$  per Table 2.  $\sigma_1$ ,  $\omega_1$ , and  $K_1$  provide a point on the root locus. The occasion may arise that for a given  $\sigma$ , there may exist no real  $\omega$  or positive  $K$ . This merely signifies that no locus exists in this portion of the  $s$ -plane.

It will be noted that the last line in Table 1 has the general equation which can be used to evaluate the  $a$ 's and  $b$ 's for additional transfer functions. But since the corresponding general equations for the  $K$  and  $\omega^2$  are missing, the above serves more as a check on new derivations than as a means of avoiding work.

TABLE 2. AID FOR ROOT LOCUS CALCULATION BY DONAHUE PROCEDURE

$(m, n)$	$\omega^2$	$-K$
(1, 2)	$[a_0 - b_0a_1]$	$\frac{1}{b_0} [a_0 - \omega^2]$
(2, 2)	$\frac{[b_0a_1 - b_1a_0]}{[a_1 - b_1]}$	$\frac{[a_0 - \omega^2]}{[b_0 - \omega^2]}$
(0, 3)	$a_1$	$[a_0 - a_2\omega^2]$
(1, 3)	$\frac{[b_0a_1 - a_0]}{[b_0 - a_2]}$	$\frac{1}{b_0} [a_0 - a_2\omega^2]$
(2, 3)	$\frac{1}{2}[(b_0 + a_1 - b_1a_2) \pm \sqrt{(b_0 + a_1 - b_1a_2)^2 - 4(b_0a_1 - b_1a_0)}]$	$\frac{[a_0 - a_2\omega^2]}{[b_0 - \omega^2]}$
(3, 3)	$\frac{1}{2(b_2 - a_2)} [(b_2a_1 + b_0 - a_0 - b_1a_2) \pm \sqrt{(b_2a_1 + b_0 - a_0 - b_1a_2)^2 - 4(b_2 - a_2)(b_0a_1 - b_1a_0)}]$	$\frac{[a_0 - a_2\omega^2]}{[b_0 - b_2\omega^2]}$
(0, 4)	$\left[ \frac{a_1}{a_3} \right]$	$[\omega^4 - a_2\omega^2 + a_0]$
(1, 4)	$\frac{1}{2}[(a_2 - b_0a_3) \pm \sqrt{(b_0a_3 - a_2)^2 + 4(b_0a_1 - a_0)}]$	$\frac{1}{b_0} [\omega^4 - a_2\omega^2 + a_0]$
(2, 4)	$\frac{1}{2(a_3 - b_1)} [(b_0a_3 + a_1 - b_1a_2) \pm \sqrt{(b_0a_3 + a_1 - b_1a_2)^2 - 4(a_3 - b_1)(b_0a_1 - b_1a_0)}]$	$\frac{[\omega^4 - a_2\omega^2 + a_0]}{[b_0 - \omega^2]}$
(0, 5)	$\frac{1}{2}[a_3 \pm \sqrt{a_3^2 - 4a_1}]$	$[a_4\omega^4 - a_2\omega^2 + a_0]$
(1, 5)	$\frac{1}{2(b_0 - a_4)} [(b_0a_3 - a_2) \pm \sqrt{(b_0a_3 - a_2)^2 - 4(b_0 - a_4)(b_0a_1 - a_0)}]$	$\frac{1}{b_0} [a_4\omega^4 - a_2\omega^2 + a_0]$
(0, 6)	$\frac{1}{2a_5} [a_3 \pm \sqrt{a_3^2 - 4a_1a_5}]$	$[-\omega^6 + a_4\omega^4 - a_2\omega^2 + a_0]$

**Construction Aids**

From the discussion in the preceding subsections, it may be inferred that locating points that satisfy the angle condition is a time-consuming procedure. To aid in this respect, a simple device can be constructed which mechanically sums the vector angles.

**Mechanical Angle Summer.** (See Fig. 56.) This device is made of clear plastic. The arm rotates on the disk with a slight drag. To use, place the pin point at an exploratory point  $s$  with arm pointing horizontally to the left and the zero degree arrowhead aligned under the arm centerline.

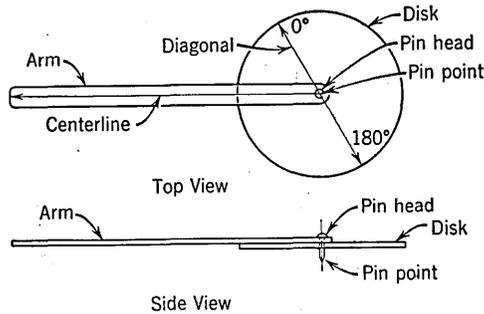


Fig. 56. Mechanical angle summer.

To sum angles of pole vectors, hold the disk in place and rotate arm centerline to a pole root. Release disk and return arm to neutral. Friction causes disk to rotate with arm. To sum angles of zero vectors, reverse the order of disk rotation. Rotate arm centerline to a zero root (disk free to rotate). Hold disk and return arm to neutral. When all roots have been successively accounted for and the arm has been returned to the neutral position, the 180-degree arrowhead should lie under the arm centerline for a point to be on the root locus.

**The Spirule** shown in Fig. 57 (Ref. 13) performs the above operation plus the additional feature of calibrating the locus. A logarithmic spiral curve on the arm permits the logarithm of a length to be obtained as an

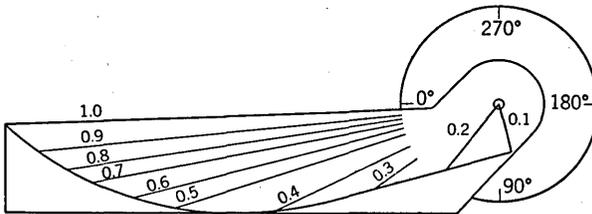


Fig. 57. Spirule. (Developed by W. Evans. Available from the Spirule Company, Whittier, Calif.)

angle, so that the addition of such angles corresponds to adding logarithms. The root locus is calibrated rather simply with this addition.

**Conductive Paper Disk.** (See Figs. 58 and 59.) To minimize further the labor involved and therefore enhance its use, machines have been devised which perform the necessary operations automatically. One such machine uses the fundamental idea that the electric potential de-

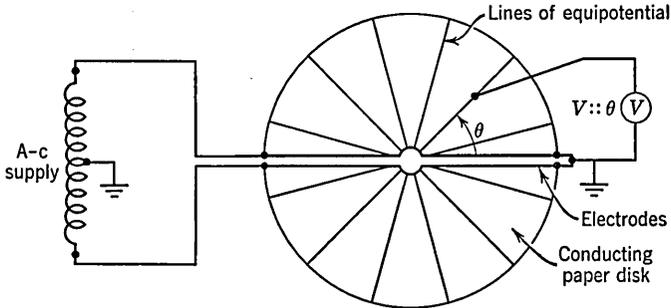


FIG. 58. Angle measurement on a conductive paper disk.

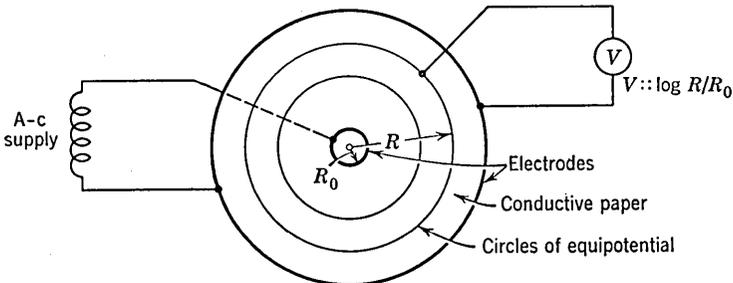


FIG. 59. Magnitude measurement on a conductive paper disk.

veloped on a conducting paper disk could represent angles or logarithms of lengths. This principle has been tried with success (Ref. 35).

**A Mechanical Plotting Machine.** This machine, described in Ref. 14, is a simple mechanical instrument which sums angles of vectors by using the principle that torque developed by a rotational spring is proportional to the angle of rotation of the spring. The machine is simple in construction, portable, and it requires no auxiliary power.

**A Compact Analog Machine.** Described in Ref. 15 and called the "Complex Plane Analyzer" this machine can, among other functions, be used to obtain a root locus plot. The principle involved is that of reducing vector multiplication to two independent summations of phase and log magnitude. To this end, a logarithmic potentiometer is used to measure

magnitude and a linear one measures phase. Capacitors are individually charged with voltages representing these quantities. Summation of capacitor voltages produces the required overall products and quotients. The machine is simple, rugged, can also be used to plot phase loci, and is available commercially.

**Some Common Root Loci**

The following plots (Figs. 60-91) are presented to aid in checking some of the preceding theorems, to present some general loci and to show in general how redistribution or variation in number of poles and zeros affects the plot.

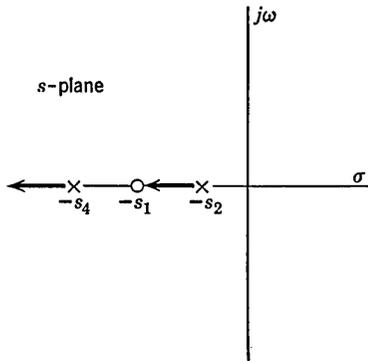


FIG. 60. Root locus plot for  

$$G(s)H(s) = \frac{K(s + s_1)}{(s + s_2)(s + s_4)}$$

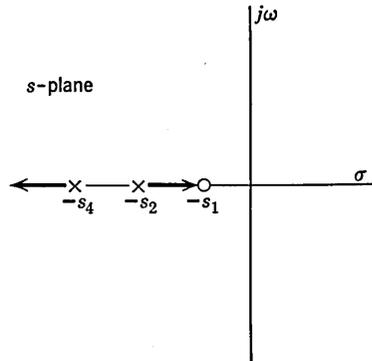


FIG. 61. Root locus plot for  

$$G(s)H(s) = \frac{K(s + s_2)}{(s + s_1)(s + s_4)}$$

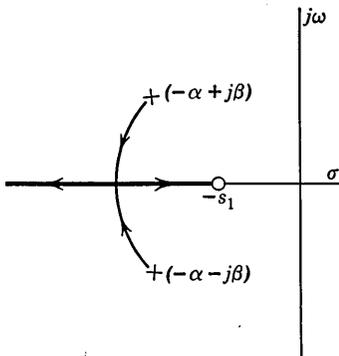


FIG. 62. Root locus plot for  

$$G(s)H(s) = \frac{K(s + s_1)}{(s + \alpha + j\beta)(s + \alpha - j\beta)}$$

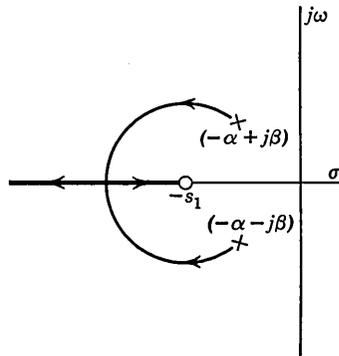


FIG. 63. Root locus plot for  

$$G(s)H(s) = \frac{K(s + s_1)}{(s + \alpha + j\beta)(s + \alpha - j\beta)}$$

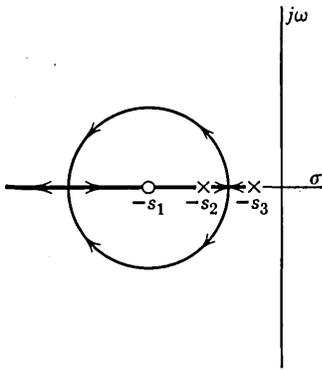


FIG. 64. Root locus plot for

$$G(s)H(s) = \frac{K(s + s_1)}{(s + s_2)(s + s_3)}$$

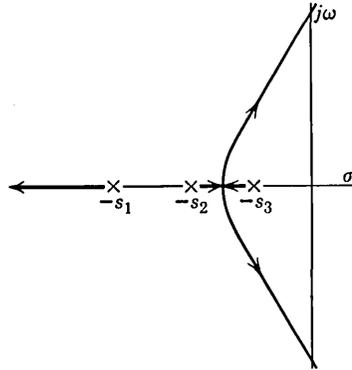


FIG. 65. Root locus plot for

$$G(s)H(s) = \frac{K}{(s + s_1)(s + s_2)(s + s_3)}$$

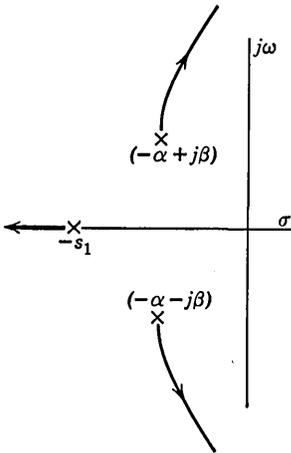


FIG. 66. Root locus plot for

$$G(s)H(s) = \frac{K}{\{(s + s_1)(s + \alpha + j\beta)\} \times \{(s + \alpha - j\beta)\}}$$

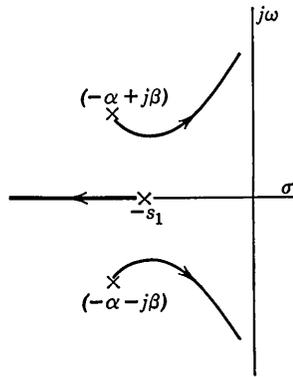


FIG. 67. Root locus plot for

$$G(s)H(s) = \frac{K}{\{(s + s_1)(s + \alpha + j\beta)\} \times \{(s + \alpha - j\beta)\}}$$

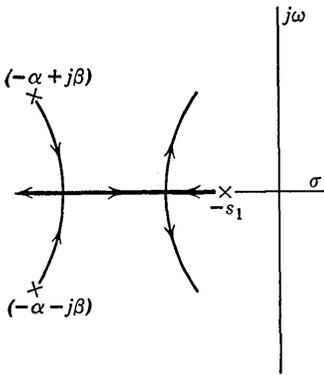


FIG. 68. Root locus plot for

$$G(s)H(s) = \frac{K}{\{(s + \alpha + j\beta)(s + \alpha - j\beta)\} \times (s + s_1)}$$

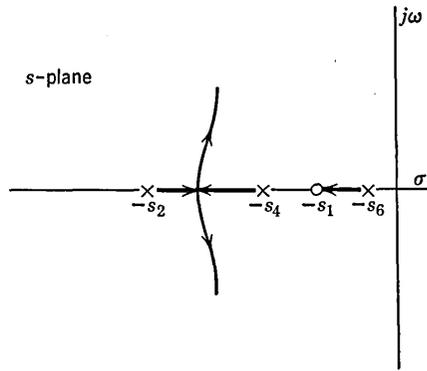


FIG. 69. Root locus plot for

$$G(s)H(s) = \frac{K(s + s_1)}{(s + s_2)(s + s_4)(s + s_6)}$$

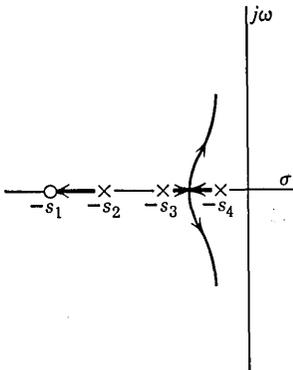


FIG. 70. Root locus plot for

$$G(s)H(s) = \frac{K(s + s_1)}{(s + s_2)(s + s_3)(s + s_4)}$$

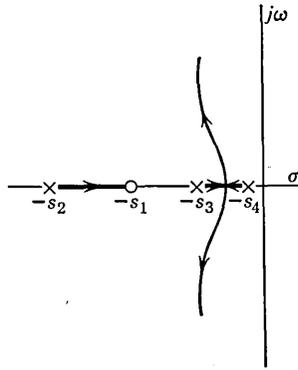


FIG. 71. Root locus plot for

$$G(s)H(s) = \frac{K(s + s_1)}{(s + s_2)(s + s_3)(s + s_4)}$$

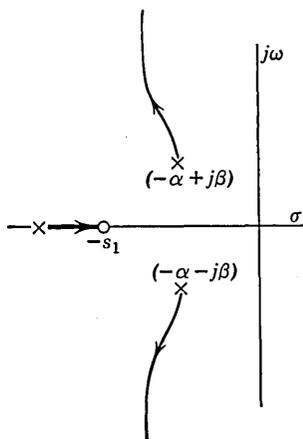


FIG. 72. Root locus plot for  $G(s)H(s) = \frac{K(s + s_1)}{\{(s + \alpha + j\beta) \times (s + \alpha - j\beta)(s + s_2)\}}$ .

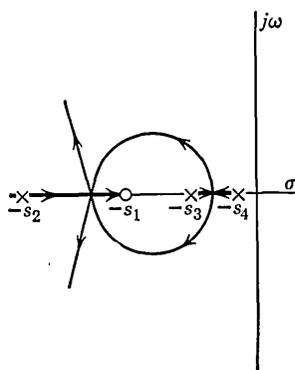


FIG. 73. Root locus plot for  $G(s)H(s) = \frac{K(s + s_1)}{(s + s_2)(s + s_3)(s + s_4)}$ .

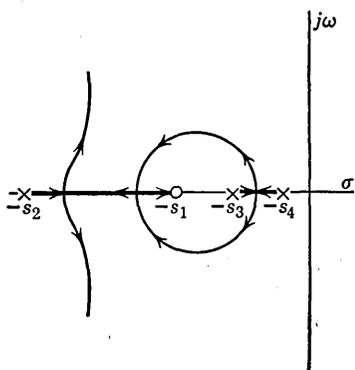


FIG. 74. Root locus plot for  $G(s)H(s) = \frac{K(s + s_1)}{(s + s_2)(s + s_3)(s + s_4)}$ .

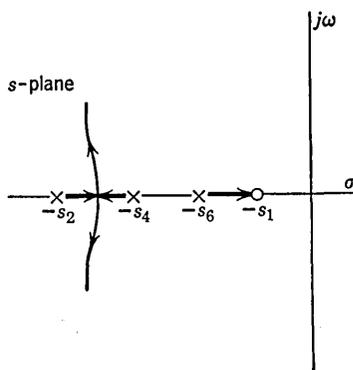


FIG. 75. Root locus plot for  $G(s)H(s) = \frac{K(s + s_1)}{(s + s_2)(s + s_4)(s + s_6)}$ .

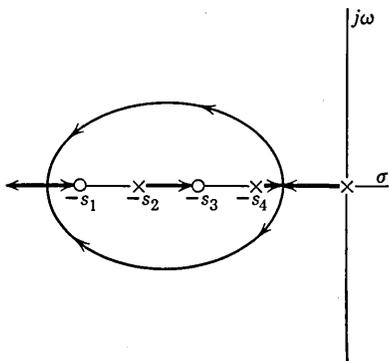


FIG. 76. Root locus plot for

$$G(s)H(s) = \frac{K(s + s_1)(s + s_3)}{s(s + s_2)(s + s_4)}$$

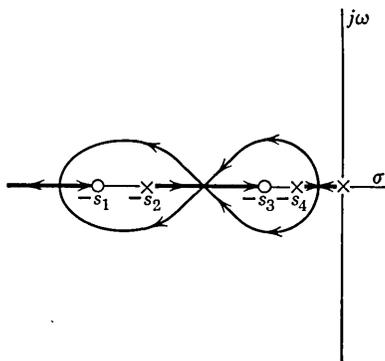


FIG. 77. Root locus plot for

$$G(s)H(s) = \frac{K(s + s_1)(s + s_3)}{s(s + s_2)(s + s_4)}$$

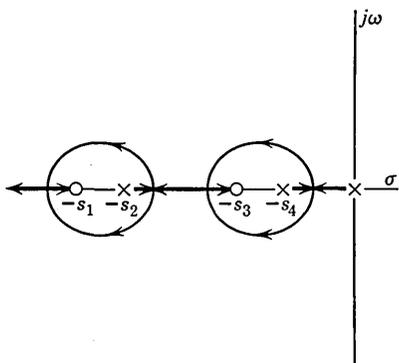


FIG. 78. Root locus plot for

$$G(s)H(s) = \frac{K(s + s_1)(s + s_3)}{s(s + s_2)(s + s_4)}$$

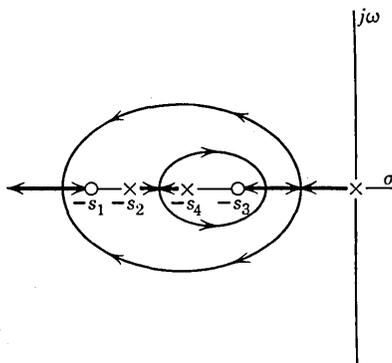


FIG. 79. Root locus plot for

$$G(s)H(s) = \frac{K(s + s_1)(s + s_3)}{s(s + s_2)(s + s_4)}$$

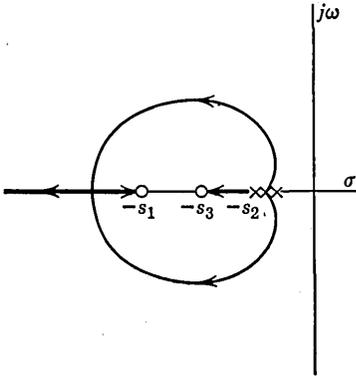


FIG. 80. Root locus plot for

$$G(s)H(s) = \frac{K(s + s_1)(s + s_3)}{(s + s_2)^3}$$

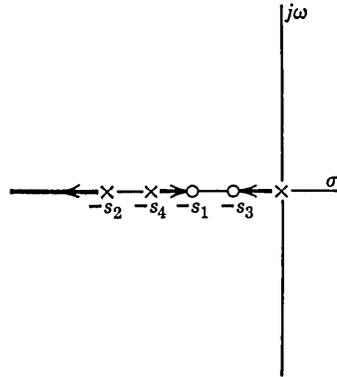


FIG. 81. Root locus plot for

$$G(s)H(s) = \frac{K(s + s_1)(s + s_3)}{s(s + s_2)(s + s_4)}$$

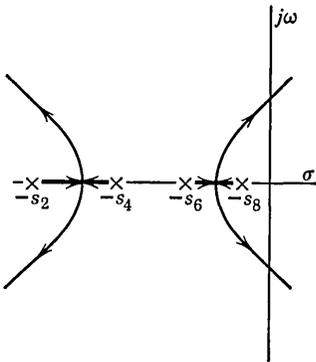


FIG. 82. Root locus plot for

$$G(s)H(s) = \frac{K}{\left\{ \begin{array}{l} (s + s_2)(s + s_4)(s + s_6) \\ \times (s + s_8) \end{array} \right\}}$$

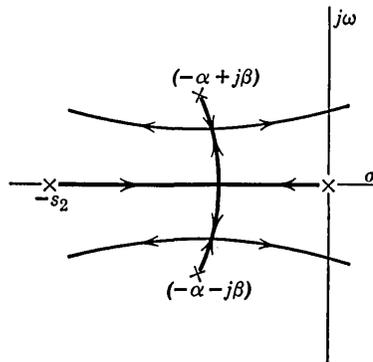


FIG. 83. Root locus plot for

$$G(s)H(s) = \frac{K}{s(s + s_2)(s + \alpha + j\beta)(s + \alpha - j\beta)}$$

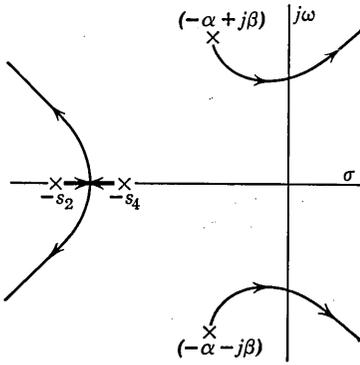


FIG. 84. Root locus plot for  $G(s)H(s)$

$$= \frac{K}{\left\{ \begin{array}{l} (s + s_2)(s + s_4)(s + \alpha + j\beta) \\ \times (s + \alpha - j\beta) \end{array} \right\}}$$

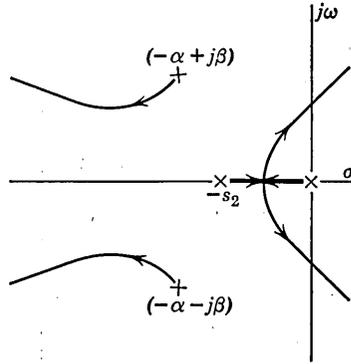


FIG. 85. Root locus plot for  $G(s)H(s)$

$$= \frac{K}{\left\{ \begin{array}{l} s(s + s_2)(s + \alpha + j\beta) \\ \times (s + \alpha - j\beta) \end{array} \right\}}$$

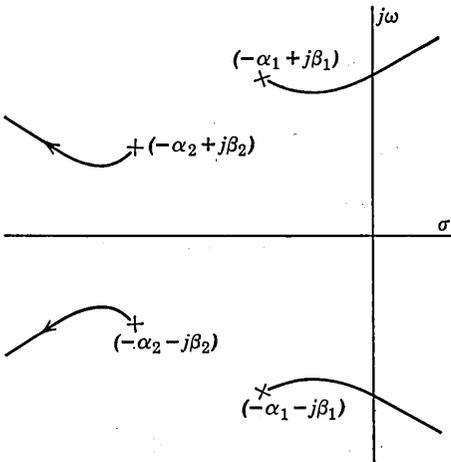


FIG. 86. Root locus plot for  $G(s)H(s)$

$$= \frac{K}{\left\{ \begin{array}{l} (s + \alpha_1 + j\beta_1)(s + \alpha_1 - j\beta_1) \\ \times (s + \alpha_2 + j\beta_2)(s + \alpha_2 - j\beta_2) \end{array} \right\}}$$

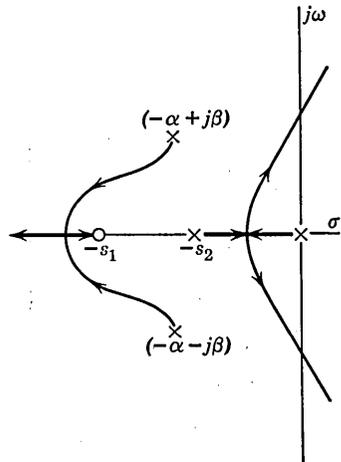


FIG. 87. Root locus plot for  $G(s)H(s)$

$$= \frac{K(s + s_1)}{\left\{ \begin{array}{l} s(s + s_2)(s + \alpha + j\beta) \\ \times (s + \alpha - j\beta) \end{array} \right\}}$$

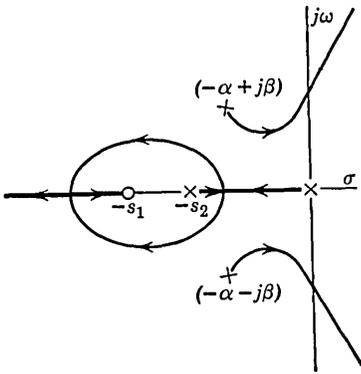


FIG. 88. Root locus plot for

$$G(s)H(s) = \frac{K(s + s_1)}{\left\{ \begin{array}{l} s(s + s_2)(s + \alpha + j\beta) \\ \times (s + \alpha - j\beta) \end{array} \right\}}$$

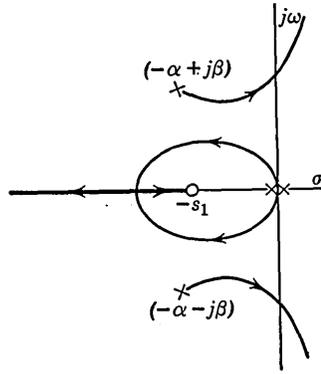


FIG. 89. Root locus plot for

$$G(s)H(s) = \frac{K(s + s_1)}{s^2(s + \alpha + j\beta)(s + \alpha - j\beta)}$$

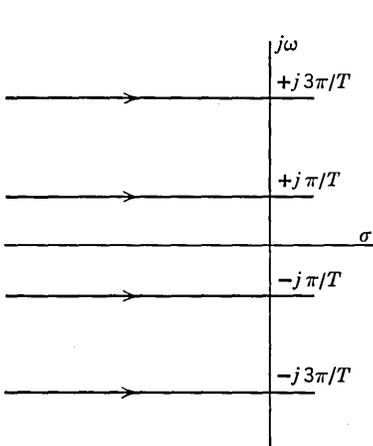


FIG. 90. Root locus plot for

$$G(s)H(s) = Ke^{-Ts}$$

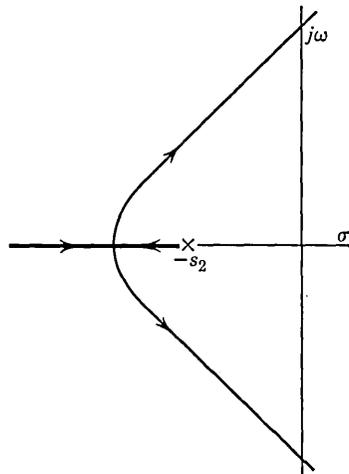


FIG. 91. Root locus plot for

$$G(s)H(s) = \frac{Ke^{-Ts}}{(s + s_2)}$$

### Interpretation of Results

The root locus plot provides a pictorial representation of the roots of the characteristic equation of the closed loop response. The location of these roots determines the modes of the transient response. Figure 92 shows contours of constant characteristics of these modes.

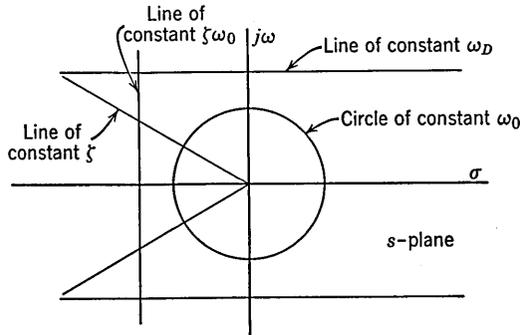


FIG. 92. Contours of constant characteristics of transient response modes.

The  $j\omega$ -axis defines the limit of absolute stability. For the system to be absolutely stable, all roots must lie in the left half  $s$ -plane. Circles concentric with the origin correspond to loci of roots with constant undamped natural frequency. Therefore, for a system prescribed to have a maximum natural frequency mode, all roots must lie within the corresponding prescribed circle. Lines of constant imaginary part correspond to lines of constant damped natural frequency. For prescribed maximum damped natural frequency, all roots must lie within the area bounded

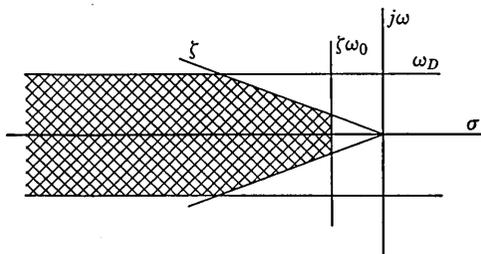


FIG. 93. Location of roots for combined restraints.

by the corresponding prescribed lines of constant imaginary part. Lines of constant real part correspond to lines of constant response time or constant exponential decay factor ( $-\zeta\omega_0$ ). Again for prescribed maximum individual response time, all roots must lie to the left of the corresponding

line of constant negative real part. Radial lines passing through the origin correspond to lines of constant damping ratio ( $\zeta$ ). For prescribed minimum damping ratio, all roots must lie within the area bounded by the corresponding minimum damping ratio lines encompassing the negative real axis. Note that lines of zero damping factor, infinite response time, and absolute stability are the same.

Combined restraints may be imposed on the modes of the transient response by reducing the area of the root location to that area common to the individual areas. For example, with specified maximum response time, maximum damped natural frequency, and minimum damping ratio, the roots would have to lie within the cross-hatched area of Fig. 93.

### Multiloop Systems Analysis

For multiloop systems, the procedure is to reduce the individual inner loops to transfer functions in factored form by use of minor loop root loci. The major root locus is then drawn up as a single loop. A particular advantage of the root locus method of analysis is that, when changes are made in the minor loops, the effect on the overall loop is shown directly.

For example, consider a closed loop voltage regulating system shown in Fig. 94. When a load is imposed upon the system, the gains change

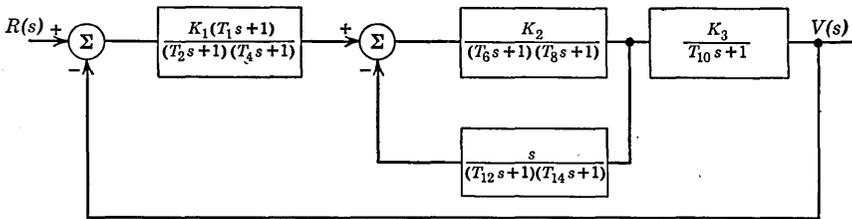


Fig. 94. Multiloop voltage regulating system.

because of nonlinear behavior. The major loop gain decreases whereas the minor loop gain increases.

The minor loop root locus is shown in Fig. 95.

The major loop root is shown in Fig. 96. This figure reveals that the net effect on the overall system of imposition of full load is that the dominant pole pair (those complex roots closest to the origin) shifts to a lower frequency with a slightly higher damping ratio whereas the subdominant pole pair (those complex roots furthest from the origin) shifts to a higher frequency with a lower damping ratio. The conclusion here is that imposition or removal of load does not severely affect the system stability or performance.

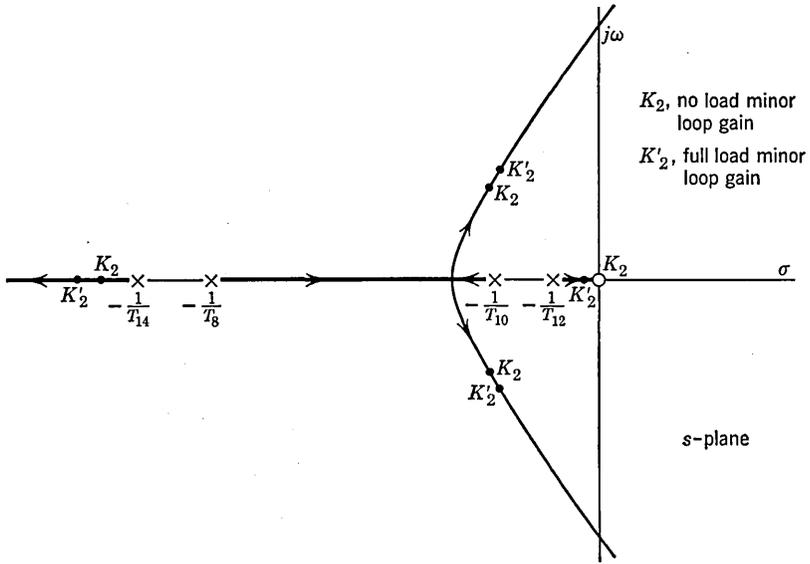


FIG. 95. Minor loop root locus plot

$$G(s)H(s) = \frac{K_2 s}{(T_6 s + 1)(T_8 s + 1)(T_{12} s + 1)(T_{14} s + 1)}$$

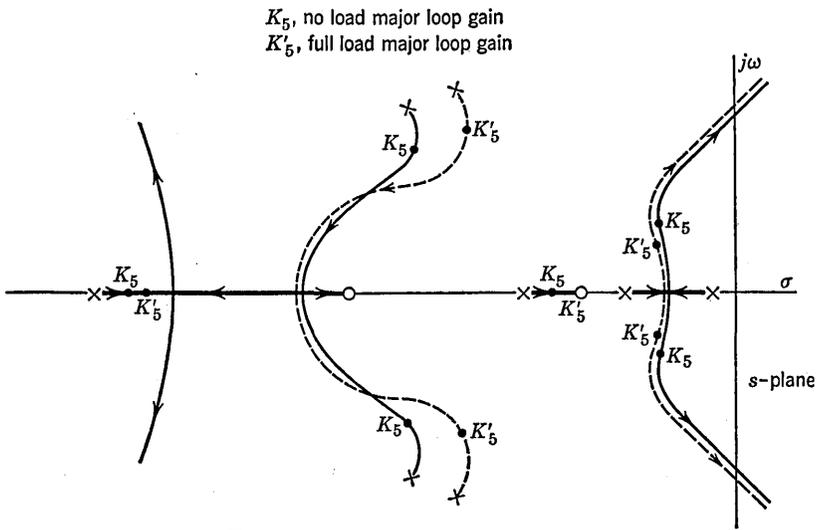


FIG. 96. Major loop root locus plot.

In multiloop systems, desired performance of the overall loop can sometimes be achieved by use of unstable minor loops. In these instances, it must be remembered that if a failure can occur such that the remaining system releases large amounts of uncontrolled energy, the design should be critically reviewed. In practice, systems are usually designed with stable inner loops.

### **System Design**

By nature, synthesis is more complicated than analysis. A few general observations can be made with regard to reshaping the root locus to obtain the required root locations. Inspection of the plots shown in Figs. 60 to 91 shows that poles tend to repel the locus whereas zeros tend to attract it. Also, as the difference between the number of poles and zeros increases, the locus tends to shift toward the right half  $s$ -plane. System synthesis through use of the root locus technique amounts to proper placement of the closed loop poles and zeros. The process is by no means simple, but by use of some of the previously mentioned machinery, a large amount of the labor is circumvented. For a detail of design in terms of root loci, see Chap. 23 and Ref. 16.

### **Relative Merits of Root Locus Method**

This method is theoretically exact and places in evidence the salient features of a closed loop system. Drawing the locus may involve a slight amount of work, but excessive labor is circumvented by use of mechanical aids.

Major advantages of this method are:

- a. The behavior pattern of the entire closed loop can be shown in one simple diagram.
- b. Modes of the transient response are placed directly in evidence.
- c. Effects of variations in system parameters are placed directly in evidence.

This is a relatively new method and is gaining widespread use.

## **7. MISCELLANEOUS STABILITY CRITERIA**

There are many, many methods to perform stability analysis of linear systems. The following is a brief account of some methods not discussed in previous sections. For the interested reader, the references can be consulted for theory and details of operation.

**Hurwitz Criterion** (Refs. 17, 18). This criterion is similar in nature to the Routh criterion and involves use of determinants. It is in general more laborious than the Routh criterion and offers information only with regard to whether or not all roots of the characteristic equation lie in the

left half  $s$ -plane. This method has been used to advantage in deriving other stability criteria such as stability boundary diagrams.

**Dzung's Criterion** (Refs. 19, 20). This stability criterion is very similar to the Nyquist criterion except that it avoids the necessity of determining the location of poles of  $G(s)H(s)$  on the  $j\omega$ -axis. It offers particular advantage when  $G(s)H(s)$  is not known in factored form and Routh's criterion indicates poles of  $G(s)H(s)$  on the  $j\omega$ -axis.

**Wall's Criterion** (Ref. 21). This stability criterion is similar to the Routh criterion and in many cases the computations are somewhat simpler.

**Stability Boundary Theory** (Refs. 22, 23, 24). This method is nice in that some simple arithmetic calculations are made by using the coefficients of the characteristic function, the results are plotted on given charts, and stability is ascertained by inspection. The main disadvantage lies in the large number of charts required for higher order systems.

### Stability Plus Assurance of Margin of Stability (Refs. 25, 26, 27)

By substitution,  $s' = (s + a)$  and/or  $s' = se^{j\theta}$  in the characteristic equation, which corresponds to shift and/or rotation of the axes in the  $s$ -plane, and subsequent analysis of the resulting equation, stability plus assurance of a margin of stability can be ascertained. The substitution may result in an equation with complex coefficients. Analysis may be carried out by use of any of the following.

**Nyquist and Dzung Criteria.** These criteria are general in nature and are applicable.

**Analog of Hurwitz** (Refs. 28, 29), **Wall** (Ref. 4), **Routh Criteria** (Ref. 4). These criteria are similar to the criteria as described previously.

**Leonhard's Criterion** (Ref. 30). This stability criterion is similar to the Nyquist criterion.

**Analog Computer Approach** (Ref. 33). By simulating the equations which describe the physical equipment's behavior, it is possible to study system stability and performance characteristics. An entire part in Vol. 2 is devoted to analog computers.

## 8. CLOSED LOOP RESPONSE FROM OPEN LOOP RESPONSE

As shown by eq. (1), the closed loop response of the general negative feedback system is a function of the forward and feedback transfer functions. Block diagram manipulation of the diagram in Fig. 1 leads to that shown in Fig. 97. That portion of the system shown in the dashed rectangle is a unity feedback system whose closed loop response is given by

$$(71) \quad \frac{C_1(s)}{R(s)} = \frac{G(s)H(s)}{1 + G(s)H(s)}$$

For any value of  $G(s)H(s)$ , the closed loop response can be considered a vector given by

$$(72) \quad \frac{C_1(s)}{R(s)} = Me^{j\alpha}.$$

$M$  is the magnitude of the vector where  $\alpha$  is its direction in radians.

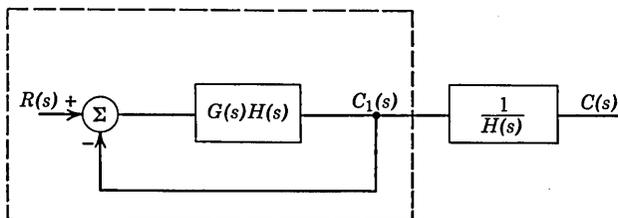


FIG. 97. General negative feedback system.

**Contours of Constant  $M$  and  $\alpha$ .** It can be shown (Ref. 31) that for unity feedback systems, certain curves in the complex plane correspond to loci of constant  $M$  and constant  $\alpha$ .

For the direct polar plot of  $G(s)H(s)$ , the  $M$  loci are circles with

$$\text{Radius} = \left| \frac{M}{M^2 - 1} \right| \quad \text{and center at } -\frac{M^2}{M^2 - 1}$$

on the axis of reals.

The curves are shown in Fig. 98. The  $\alpha$  loci correspond to circles passing through the origin and the  $-1$  point and centers at

$$-\frac{1}{2} + j \frac{1}{2 \tan \alpha}.$$

These  $\alpha$  curves are shown in Fig. 99. These curves of constant  $M$  and  $\alpha$  are useful for many purposes. An important use is that of obtaining the closed loop response from a plot of the open loop response,  $G(s)H(s)$ .  $G(s)H(s)$  is superimposed on curves of constant  $M$  and constant  $\alpha$ , and the closed loop magnitude and phase angles are obtained by inspection of the respective circles at points of intersection with  $G(s)H(s)$ .

For the inverse polar plot of  $G(s)H(s)$  it can be shown (Ref. 31) that the contours of constant  $M$  are circles with center at the  $-1$  point and radius equal to  $1/M$ . See Fig. 100. The contours of constant  $\alpha$  are straight lines passing through the  $-1$  point with slope equal to  $\alpha$ .

The  $M$  and  $\alpha$  contours in the inverse  $G(s)H(s)$ -plane plot are somewhat easier to use because of ease of construction.

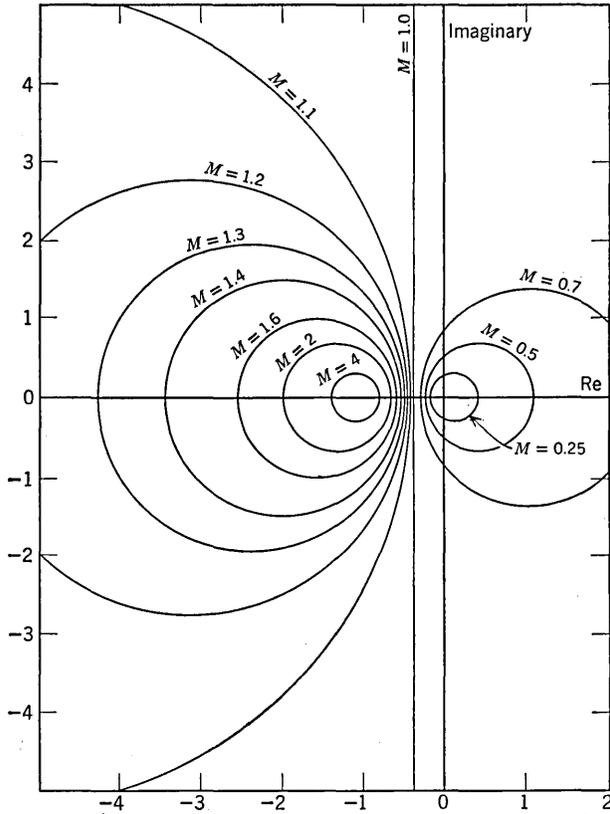


FIG. 98. Circles of constant  $M$  in the  $G(s)H(s)$ -plane.

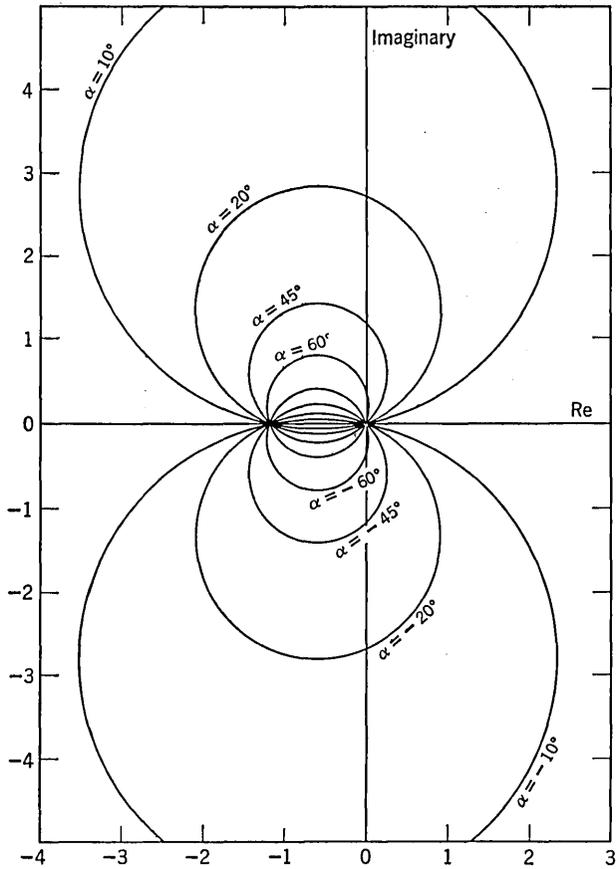


FIG. 99. Circles of constant  $\alpha$  in the  $G(s)H(s)$ -plane.

It is of interest to note that the  $M$  and  $\alpha$  contours are perfectly general curves for unity feedback systems. In other words,  $G(s)H(s)$  is not restricted to those values of  $s$  on the  $j\omega$  axis.

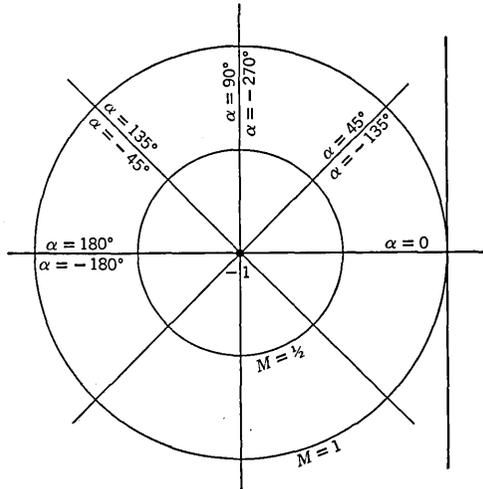


Fig. 100. Contours of constant  $M$  and  $\alpha$  in the inverse  $G(s)H(s)$ -plane.

### Nichols Charts

The information contained in the  $M$  and  $\alpha$  circles, when plotted in terms of decibels and phase angle as shown in Fig. 101, are commonly referred to as Nichols charts because of the fundamental work first done by N. B. Nichols (Ref. 32). The curves shown in the figure have a mirror image about the  $-180^\circ$  ordinate. The total curves correspond to the principal value of the logarithm given by eq. (28). Since the logarithm of a complex number is multivalued, eq. (27) the curves repeat as shown in Fig. 102.

**Stability Analysis on the Nichols Chart.** Note that the  $-1$  point in the  $G(s)H(s)$ -plane corresponds to the 0-db,  $-180^\circ$  point in Fig. 101. For well-behaved, minimum phase  $G(s)H(s)$ , the system is stable if  $G(j\omega)H(j\omega)$  crosses the 0-db line to the right of the  $-1$  point on the Nichols chart. Figure 103 shows a stable system with a phase margin of  $+48^\circ$  at gain crossover and a gain margin of 6.8 db at phase crossover.

**Exact Closed Loop Response.** The procedure to obtain closed loop response from open loop response is as follows:

a. Manipulation of the general negative feedback system to the form shown in Fig. 97.

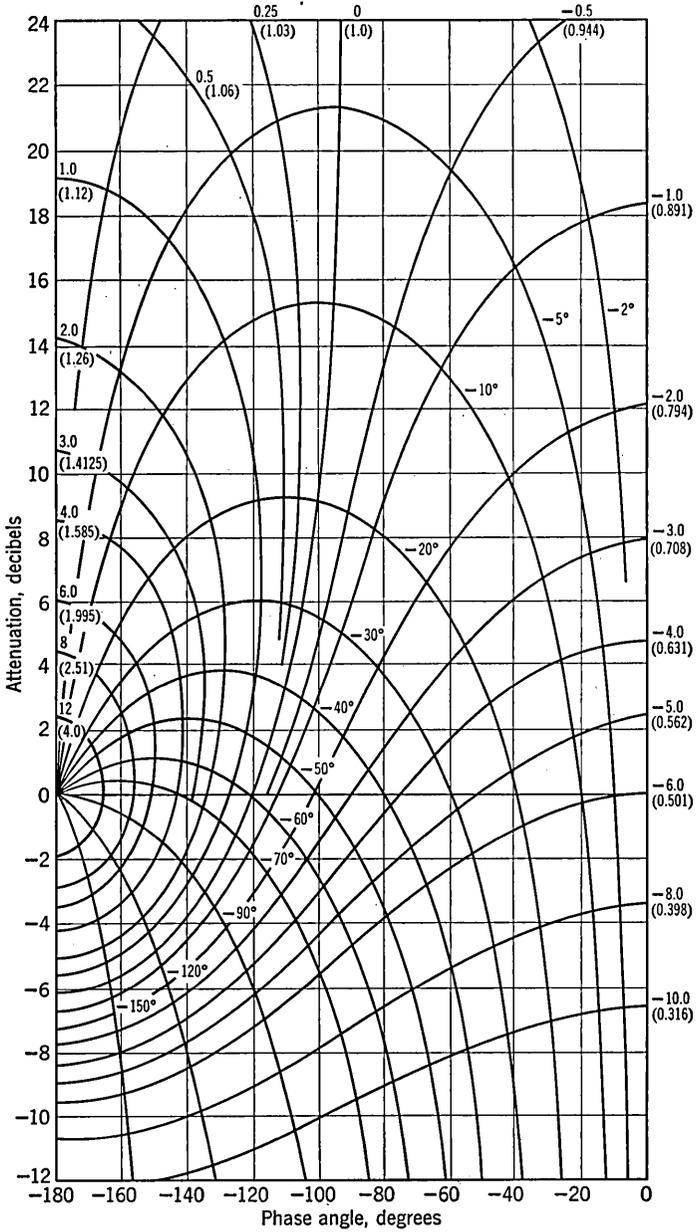


FIG. 101. Nichols chart.

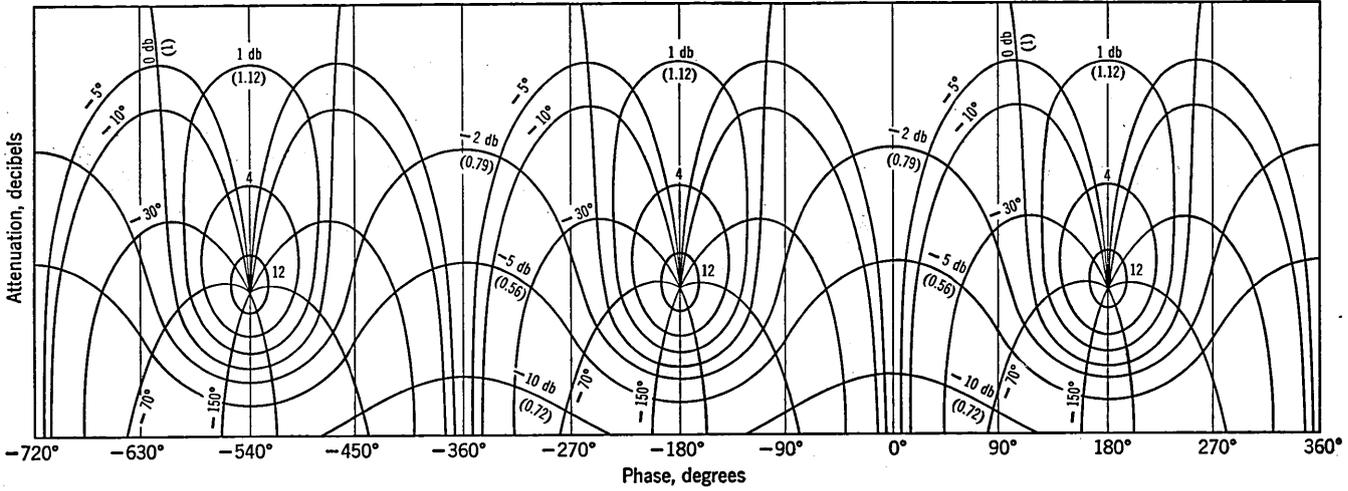


FIG. 102. Multiple Nichols charts.

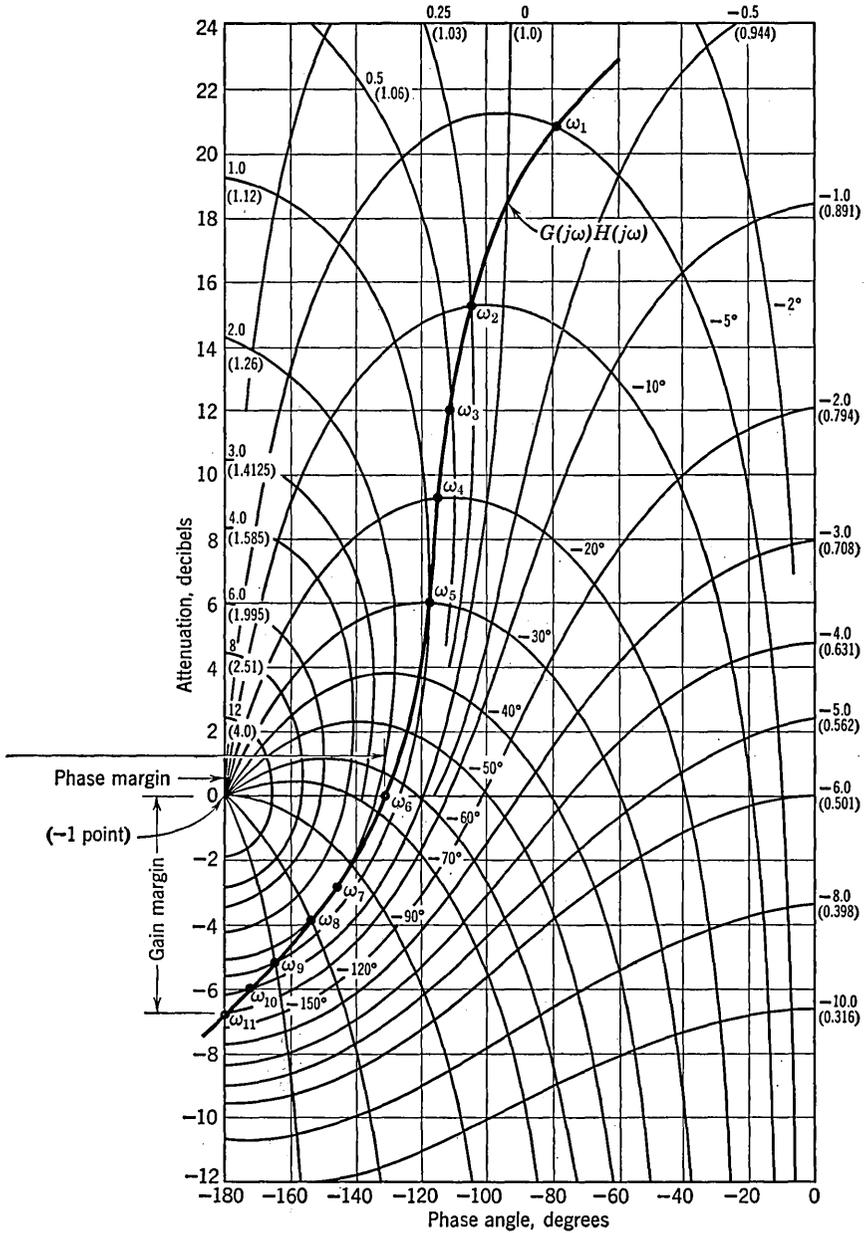


FIG. 103.  $G(s)H(s)$  plotted on Nichols chart.

b. Plot  $G(s)H(s)$  directly or inversely with the corresponding  $M$  and  $\alpha$  loci.  $G(s)H(s)$  on the Nichols chart is shown in Fig. 103.

c. Obtain the closed loop  $M$  and  $\alpha$  at points of intersection of  $G(s)H(s)$  with the  $M$  and  $\alpha$  loci.

d. Modify this response by  $1/H(s)$  to obtain the overall closed loop response.

**Approximate Closed Loop Response.** The approximate closed loop response can be obtained by plotting the Bode diagram for  $G(s)$  and

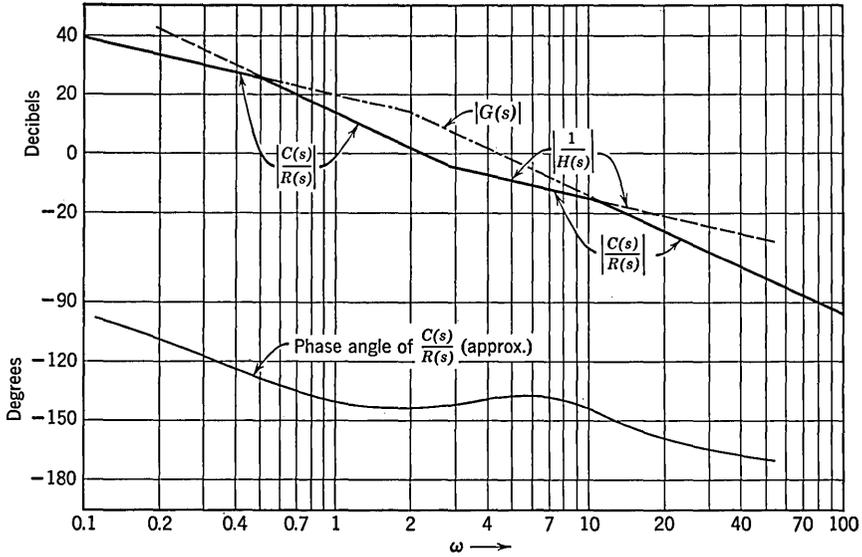


FIG. 104. Approximate closed loop response.

$[H(s)]^{-1}$  on the same sheet. The closed loop response is approximately equal to the lower of the two curves at any given frequency (see Fig. 104).

$$(73) \quad \frac{C(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)} = \frac{1}{[1/G(s)] + H(s)},$$

when  $H(s)$  is much smaller than  $G(s)$  and  $G(s)H(s)$  is greater than 1.

$$\frac{1}{G(s)} + H(s) \approx H(s)$$

and

$$\frac{C(s)}{R(s)} \approx \frac{1}{H(s)},$$

when  $G(s)$  is much smaller than  $H(s)$  and  $G(s)H(s)$  is less than 1.

$$\frac{1}{G(s)} + H(s) \approx \frac{1}{G(s)}$$

and

$$\frac{C(s)}{R(s)} \approx G(s).$$

The approach is to approximate the closed loop response by the lowest portions of  $G(s)$  and  $[H(s)]^{-1}$ . Assume all the breaks are simple and of multiplicity one or more. The phase diagram is drawn assuming the simple breaks of a minimum phase nature.

The approximation is worst in the region of  $\omega$  where  $G(s) = [H(s)]^{-1}$ . If necessary, the exact closed loop response can be obtained in this region by using the preceding exact method and the Nichols charts.

EXAMPLE.

$$G(s) = \frac{10}{s(0.5s + 1)}, \quad H(s) = \frac{0.2s^2}{(0.33s + 1)}.$$

In Fig. 104 are plotted  $G(s)$  and  $[H(s)]^{-1}$ . The approximate closed loop response is shown as the heavy line and is approximated by the equation

$$(74) \quad \frac{C(s)}{R(s)} \approx \frac{10(0.33s + 1)}{s(2s + 1)(0.09s + 1)}.$$

The phase angle curve is that corresponding to eq. (74).

*Note.* There are many ways to investigate stability of linear closed loop systems. If used properly, they should all obtain the same result.

## REFERENCES

1. E. J. Routh, Stability of a dynamical system with two independent motions, *Proc. London Math. Soc.*, ser. 1, 5, 97-99 (1874).
2. E. J. Routh, *A Treatise on the Stability of a Given State of Motion*, Cambridge University Press, Cambridge, England, 1877.
3. E. J. Routh, *Advanced Part of a Treatise on Advanced Rigid Dynamics*, 6th edition, pp. 210-231, Cambridge University Press, Cambridge, England, 1930.
4. T. J. Higgins, Epitomization of the Basic Concepts Underlying the Theory of "The Stability" of Servomechanisms, *Advanced Servomechanisms and Automatic Control Theory*, Class Notes EE 216, University of Wisconsin, Ronald, New York, 1955.
5. E. A. Guillemin, *The Mathematics of Circuit Analysis*, Technology Press and Wiley, New York, 1950.
6. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. 1, Wiley, New York, 1951.
7. H. W. Bode, *Network Analysis and Feedback Amplifier Design*, Van Nostrand, Princeton, N. J., 1945.

8. H. W. Bode, Amplifiers, Patent 2,123,178 (1938).
9. N. Balbiani and W. R. Lepage, What is a minimum phase network? *Trans. Am. Inst. Elec. Engrs.*, Pt. 1, No. 22, January 1956.
10. G. A. Biernson, Estimating transient responses from open-loop frequency response, *Trans. Am. Inst. Elec. Engrs.*, 74, 388-402, Pt. 2, January 1956.
11. W. R. Evans, Graphical analysis of control systems, *Trans. Am. Inst. Elec. Engrs.*, 67, 547-551 (1948).
12. W. R. Evans, Control system synthesis by root locus method, *Trans. Am. Inst. Elec. Engrs.*, 69, Pt. 1, 67-69 (1950).
13. W. R. Evans, *Control Systems Dynamics*, McGraw-Hill, New York, 1954.
14. A. H. Harris, *A Simple Instrument for Summing Angles in the Root Locus Method of Solving Ordinary Equations and Stability Problems*, University of California, UCRL-2269, Berkeley, July 10, 1953.
15. The Complex Plane Analyzer, The Technology Instrument Corporation, CPA type 250-A, Acton, Mass.
16. J. G. Truxal, *Automatic Feedback Control System Synthesis*, McGraw-Hill, New York, 1955.
17. A. Hurwitz, Ueber die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen realen Theilen besitzt, *Math. Ann.*, 46, 273-284 (1895).
18. L. Orlando, Sul problema di Hurwitz, *Rendiconte Accademia Lincei*, ser. 5, Vol. 19, pp. 801-805, Rome, 1910.
19. L. S. Dzung, The Stability Criterion, in *Automatic and Manual Control*, Butterworths, London, 1952 (Proceedings of the 1951 Cranfield Conference, pp. 13-23).
20. L. S. Dzung, Das Stabilitätskriterium nach Nyquist, *Regelungstechnik*, 1, 143-145 (1953).
21. H. S. Wall, Polynomials whose zeros have negative real parts, *Am. Math. Monthly*, 52, 308-322 (1945).
22. E. Sponder, On the representation of the stability region on oscillation problems with the aid of Hurwitz determinants, NACA Technical Memorandum 1348, August 1952, A Translation of E. Sponder, Zur Darstellung des Stabilitätsgebietes bei Schwingungsaufgaben mit Hilfe der Hurwitz-Determinanten, *Schweiz. Arch.*, 16, 93-96 (1950).
23. J. F. Koenig, On the zeros of polynomials and the degree of stability of linear systems, *J. Appl. Phys.*, 24, 476-482 (1953).
24. T. J. Higgins and J. G. Levinthal, Stability limits for third-order servomechanisms, *Trans. Am. Inst. Elec. Engrs.*, 71, Pt. 2, 459-467 (1952).
25. J. F. Koenig, A relative damping criterion for linear systems, *Trans. Am. Inst. Elec. Engrs.*, 72, Pt. 2, 291-295 (1953).
26. A. Vazsong, A generalization of Nyquist's stability criteria, *J. Appl. Phys.*, 20, 863-867 (1949).
27. A. Leonhard, Relative Damping as Criterion for Stability and as an Aid in Finding the Roots of a Hurwitz Polynomial, in *Automatic and Manual Control*, Butterworths, London, 1952 (Proceedings of the 1951 Cranfield Conference, pp. 25-43).
28. E. Frank, On the zeros of polynomials with complex coefficients, *Bull. Am. Math. Soc.*, 52, 144-157 (1946).
29. H. Bilharz, Bemerkung zu einem Satze von Hurwitz, *Z. angew. Math. u. Mech.*, 24, 77-82 (1944).
30. A. Leonhard, Ueber Selbsterregung elektrischer Maschinen, *Arch. Elektrotech.*, 40, 343-346 (1952).
31. G. S. Brown and D. P. Campbell, *Principles of Servomechanisms*, Wiley, New York, 1948.

32. H. M. James, N. B. Nichols, and R. S. Phillips, *Theory of Servomechanisms*, McGraw-Hill, New York, 1947.
33. C. L. Johnson, *Analog Computer Techniques*, McGraw-Hill, New York, 1956.
34. Robert Donahue, unpublished, M.I.T. Flight Control Laboratory.
35. General Electric Company, unpublished, Schenectady, New York.



## Relation between Transient and Frequency Response

*C. E. Bradford and M. W. DeMerit*

1. Introduction	22-01
2. Response Characteristics Defined	22-02
3. Relation between Transient Response and Location of Roots of Characteristic Equation	22-03
4. Relation between Closed Loop and Open Loop Roots	22-15
5. Design Charts Relating Open Loop Frequency Response and Transient Response	22-18
6. Approximate Relations—Rules of Thumb	22-43
7. Numerical and Graphical Techniques of Relating Transient and Frequency Response	22-43
References	22-61

### 1. INTRODUCTION

The frequency response technique of analyzing servo systems is used to facilitate both the analysis and synthesis operations (Chaps. 20 and 21). Often it is desirable to transform the results of the frequency response analysis into transient response form in order to interpret them more readily. Conversely, it is often desirable to transform the transient response performance requirements into frequency response form for synthesis purposes. These operations can be performed exactly by rigorous mathematical techniques; however, the operations are time consuming

and tedious, so *it is often profitable to use less accurate but more easily applied techniques*. The purpose of this section is to present some of the more useful techniques for relating the transient response to the frequency response and the inverse relations between frequency and transient response.

**2. RESPONSE CHARACTERISTICS DEFINED**

**Transient Response.** System response is often specified and interpreted in terms of the characteristics of the transient response to a step

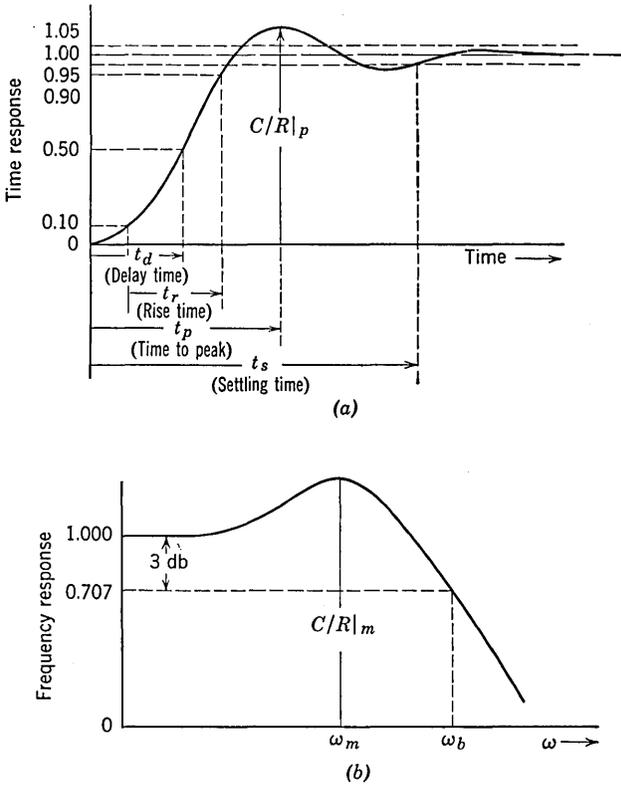


FIG. 1. (a) Representative system response to unit step input. (b) Representative system frequency response (closed loop).

input. The parameters which are most often used to describe the transient response are:

- $C/R|_p$ , the peak value of the transient including any overshoot;
- $t_p$ , the time to the first peak if the response is underdamped and thus has an overshoot;

$t_s$ , the settling time, measured from the initiation of the step input to the time at which the system output no longer deviates more than a certain percentage, often 5 or 2 per cent from its final value;

$N$ , the number of oscillations it takes the system to "settle" or reach  $t_s$ .

Other parameters sometimes used to describe the transient response are:

$t_d$ , the delay time, measured from the initiation of the step input to the time at which the response has reached half the final value;

$t_r$ , the rise time, which is the difference between the time at which the response has reached 10 per cent of the final value and the time at which 90 per cent of the final value is reached. Rise time is also sometimes defined as the time from 5 to 95 per cent, and also as the reciprocal of the slope at the instant the response is 50 per cent of the final value.

Figure 1a illustrates the definitions of these transient response parameters.

**Frequency Response.** System response is also often described in terms of certain frequency response characteristics. Chief of these are:

$M_m$ , the maximum amplitude ratio of the closed-loop frequency response, which is sometimes designated  $C/R|_m$ ;

$\omega_m$ , the frequency at which  $M_m$  occurs;

$\omega_b$ , the bandpass frequency which is generally defined as the frequency at which the closed loop response is down 3 db from the nominal steady-state gain value. Figure 1b illustrates the definitions of these terms.

### 3. RELATION BETWEEN TRANSIENT RESPONSE AND LOCATION OF ROOTS OF CHARACTERISTIC EQUATION

**Mathematical Relation.** The open loop frequency response may be represented by the open loop response function in terms of its poles and zeros, roots of the denominator and numerator respectively, of the forward and feedback portions of the control system,

$$(1) \quad G(s)H(s) = \frac{N_1(s)N_2(s)}{D_1(s)D_2(s)} \\ = \frac{K(s + z_{11})(s + z_{12}) \cdots (s + z_{21})(s + z_{22}) \cdots}{s^m(s + p_{11})(s + p_{12}) \cdots (s + p_{21})(s + p_{22}) \cdots}$$

The closed loop frequency response function can be obtained as follows:

$$(2) \quad \frac{C(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)} = \frac{N_1(s)/D_1(s)}{1 + N_1(s)N_2(s)/D_1(s)D_2(s)} \\ = \frac{N_1(s)D_2(s)}{D_1(s)D_2(s) + N_1(s)N_2(s)}$$

The closed loop poles can be found by factoring  $[D_1(s)D_2(s) + N_1(s)N_2(s)]$ . Substituting the proper Laplace transform for  $R(s)$ ,  $C(s)$  may be represented by a sum of terms such as

$$(3) \quad C(s) = \frac{A_1}{s} + \frac{A_2}{s + a_2} + \frac{A_3}{s + a_3} + \dots,$$

where  $z_{1n}$  are the roots of  $N_1$ ,  $z_{2n}$  are the roots of  $N_2$ ,  $p_{1n}$  are the roots of  $D_1$ ,  $p_{2n}$  are the roots of  $D_2$ , and where the constants  $A_1, A_2, A_3$ , etc., are found by partial fraction expansion.

The time response function is then found by performing the inverse Laplace transformation to get

$$(4) \quad c(t) = A_1 + A_2 \exp(-a_2 t) + A_3 \exp(-a_3 t) + \dots$$

In general, this function must be plotted to determine such parameters as peak overshoot and settling time which are often of prime importance. The straight mathematical approach is impractical for any but simple systems because of the amount of tedious work involved and the fact that it does not lend itself to system synthesis.

### Approximate Approach

The time response can be estimated quite accurately by noting the location of certain predominate closed loop poles and zeros in the complex frequency plane ( $s$ -plane). The closed loop pole-zero configuration may consist of one or more pairs of complex poles and several real axis poles and zeros, and perhaps complex zeros. Ordinarily, one pair of complex poles will be of primary importance because of its frequency or damping ratio. For *example*, if a system contains two pairs of complex poles which have natural frequencies that differ by as much as 10 to 1, the designer may ordinarily consider either pair as dominant and perform an analysis in two parts, considering first one pair and then the other.

*For many cases it is reasonably accurate to neglect all but one complex pole pair and to consider the transient response to be made up of the dominant pair of complex poles and various groupings of real axis poles and zeros.* This assumption is made in the following discussion.

Only *underdamped systems* are to be considered here. Overdamped systems may generally be analyzed quite easily by the normal mathematical techniques.

**Dominant Pair of Complex Poles.** To determine the effect of the  $s$ -plane pole-zero configuration in the system transient response, it is convenient to first consider the relation between a single pair of complex poles on the  $s$ -plane and the characterizing parameters of the time response.

The additional effect of the real axis poles and zeros will be considered later.

The expression for the closed loop frequency response function containing one pair of complex roots is

$$\begin{aligned}
 (5) \quad \frac{C(s)}{R(s)} &= \frac{G(s)}{1 + G(s)} = \frac{K\omega_0^2}{s^2 + 2\zeta\omega_0s + \omega_0^2} \\
 &= \frac{K\omega_0^2}{(s + \sigma_0 + j\omega_d)(s + \sigma_0 - j\omega_d)}
 \end{aligned}$$

where  $\omega_0$  = natural frequency,

$\zeta$  = damping ratio,

$\sigma_0 = \zeta\omega_0$  = damping exponent,

$\omega_d = \omega_0\sqrt{1 - \zeta^2}$  = natural damped frequency, or oscillation frequency.

The parameters  $\omega_0$ ,  $\zeta$ ,  $\sigma_0$  and  $\omega_d$  are shown on the  $s$ -plane in Fig. 2.

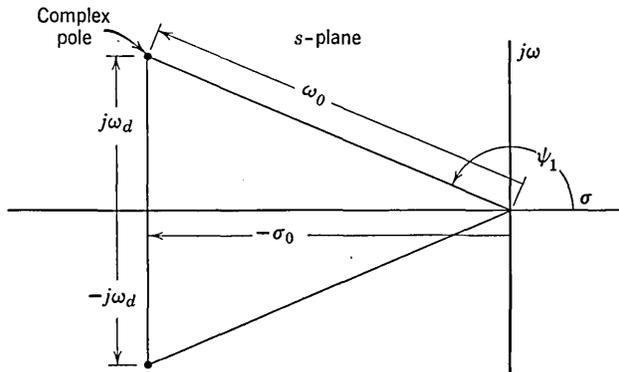


FIG. 2. One pair of complex roots and significant related parameters,  $\zeta = \cos \psi_1$ .

Figure 3 illustrates that for constant values of natural frequency,  $\omega_0$ , the complex roots or poles of eq. (5) generate circles on the  $s$ -plane as the damping ratio  $\zeta$  is varied. Radial lines from the origin are generated by holding  $\zeta$  constant and varying  $\omega_0$ .

Figure 4 illustrates that holding  $\sigma_0$ , the exponential damping factor, constant forms lines parallel to the imaginary ( $j\omega$ )-axis on the  $s$ -plane. Similarly, maintaining constant values for the damped frequency,  $\omega_d$ , forms lines parallel to the real ( $\sigma$ ) axis.

The expression for the time response to a unit step input is

$$(6) \quad c(t) = 1 + (\omega_0/\omega_d) \exp(-\sigma_0 t) \sin(\omega_d t - \psi_1),$$

where  $\psi_1 = \arctan \omega_d / -\sigma_0$

$$= \arctan \sqrt{1 - \zeta^2} / \zeta.$$

Figure 3 also illustrates that constant values of  $\psi_1$  correspond to constant

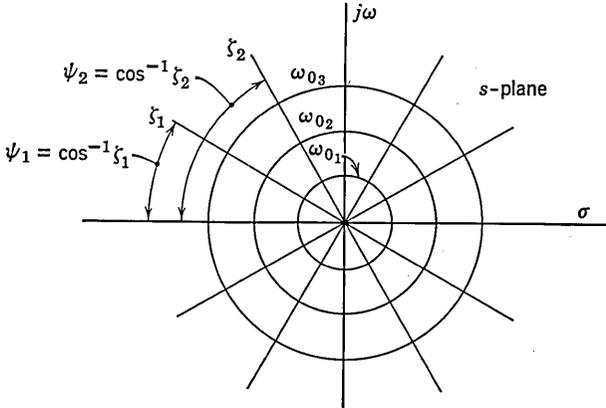


FIG. 3. Constant  $\omega_0$  is a circle; constant  $\zeta$  is a radial line.

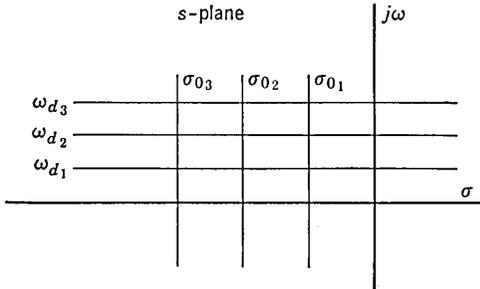


FIG. 4. Illustrating lines of constant  $\omega_d$  and  $\sigma_0$ .

values of  $\zeta$ . From eq. (6) the characterizing parameters of the transient response can be derived. The equations for the more important ones are:

$$(7) \quad t_p = \pi/\omega_d = \pi/\omega_0 \sqrt{1 - \zeta^2}, \quad \text{time required to reach first peak.}$$

$$(8) \quad t_s = 3/\sigma_0 = 3/\zeta\omega_0, \quad \text{time required to settle to within 5\% of final value (=4/\zeta\omega_0 for 2\%).}$$

- (9)  $C/R|_p = 1 + \exp(-\pi\zeta/\sqrt{1-\zeta^2})$ , the peak value of the ratio of output to input.  
 $= 1 + \exp(-\pi\sigma_0/\omega_d)$
- (10)  $N = t_s/(2\pi/\omega_d)$ , number of oscillations to settle to within 5% of final value.  
 $= 3\sqrt{1-\zeta^2}/2\pi\zeta$

Equations (7) through (10) relate the position of the dominant pair of complex poles on the s-plane to certain transient response parameters.

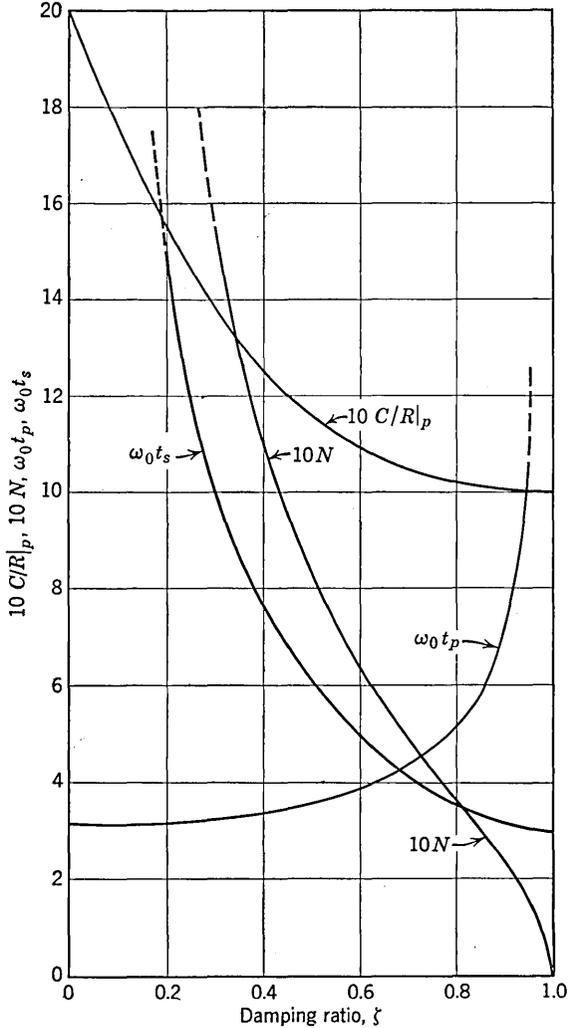


FIG. 5.  $C/R|_p, N, \omega_0 t_p, \omega_0 t_s$  versus  $\zeta$  for system composed of two complex poles.

The relationships between these parameters and the damping ratio,  $\zeta$ , are plotted in Fig. 5.

Similarly certain closed loop frequency response parameters can be related to the position of the poles.

(11)  $M_m = 1/(2\zeta\sqrt{1 - \zeta^2})$ ,  $0 < \zeta < 0.707$ , maximum frequency response ratio of output to input (also defined as  $M_m$ ).  
 $= 1$ ,  $0.707 < \zeta < 1$

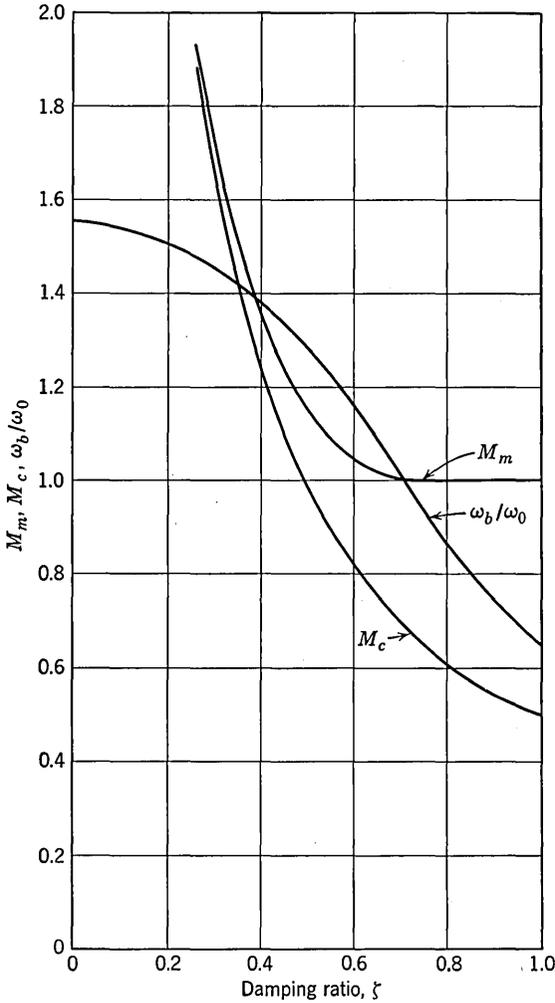


FIG. 6.  $M_m$ ,  $M_c$ ,  $\omega_b/\omega_0$ , versus  $\zeta$  for system composed of two complex poles.

(12)  $M_c = 1/(2\zeta)$ , response ratio at the frequency corresponding to the natural frequency, or corner frequency.

(13)  $\omega_b = \omega_0 \sqrt{1 - 2\zeta^2 + \sqrt{2 - 4\zeta^2 + 4\zeta^4}}$ , bandpass frequency, at which response ratio is 0.707.

Figure 6 shows graphically the relationship between these parameters and damping ratio.

**Effect of Real Axis Poles and Zeros.** The mathematical expression for a system whose dynamic characteristics can be described by one pair of complex roots, one real pole, and one real zero is

$$(14) \quad \frac{C(s)}{R(s)} = \frac{\omega_0^2 p(s + z)}{z(s + p)(s^2 + 2\zeta\omega_0 s + \omega_0^2)}$$

The expression for the transient response to a unit step input may be written as

$$(15) \quad c(t) = 1 - \frac{\omega_0^2(z - p)}{z(\overline{pp_d})^2} \exp(-pt) + \left(\frac{z\overline{p_d}}{z}\right) \left(\frac{p}{\overline{pp_d}}\right) \left(\frac{\omega_0}{\omega_d}\right) \exp(-\sigma_0 t) \sin(\omega_d t - \psi_1 + \psi_3 - \psi_4)$$

where  $\overline{pp_d}$  = distance from  $p$  to  $p_d$ ,

$\overline{zp_d}$  = distance from  $z$  to  $p_d$ .

A graphical representation of this system is contained in Fig. 7.

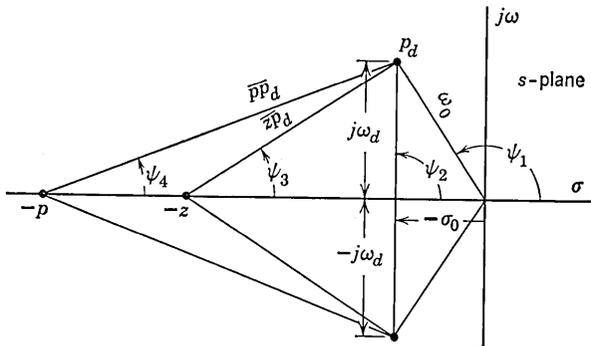


FIG. 7. Illustrating a system with one pair of complex poles, one real pole, and one real zero.

The term  $[\omega_0^2(z-p)/z(\overline{pp_d})^2] \exp(-pt)$  in eq. (15) is neglected in determining the following expressions for the characteristic parameters of the transient response. This approximation is valid when  $p$  is much larger than  $\sigma_0$  (at least 3 times as large). The expressions for  $C/R|_p$ ,  $t_p$ , and  $t_s$  are:

$$(16) \quad \frac{C}{R} \Big|_p = \left( \frac{\overline{zp_d}}{z} \right) \left( \frac{p}{\overline{pp_d}} \right) \exp \left[ \frac{-\sigma_0(\pi - \psi_3 + \psi_4)}{\omega_d} \right],$$

$$(17) \quad t_p = (\pi - \psi_3 + \psi_4)/\omega_d,$$

$$(18) \quad t_s = 3/\zeta\omega_0.$$

The settling time,  $t_s$ , remains the same as before since the exponential pole term in eq. (15) has been neglected. Thus,  $N$  also remains unchanged. For multiple poles and zeros the expressions for  $C/R|_p$  and  $t_p$  in eqs. (16) and (17) become:

$$(19) \quad \frac{C}{R} \Big|_p = \left( \prod_q \frac{p_q}{p_q p_d} \right) \left( \prod_q \frac{\overline{z_q p_d}}{z_q} \right) \exp \left[ \frac{-\sigma_0(\pi - \Sigma\psi_3 + \Sigma\psi_4)}{\omega_d} \right],$$

$$(20) \quad t_p = \frac{\pi - \Sigma\psi_3 + \Sigma\psi_4}{\omega_d},$$

where  $\Sigma\psi_3$  = sum of all angles between the real axis zeros, and the dominant pole,

$\Sigma\psi_4$  = sum of all angles between the real axis poles, and the dominant pole,

$\prod_q \frac{p_q}{p_q p_d}$  = product of the ratios of the poles to the distances from the poles to complex pole at point  $p_d$ ,

$\prod_q \frac{\overline{z_q p_d}}{z_q}$  = product of the ratios of the distances of the zeros to point  $p_d$ , to the zeros.

As noted before, eqs. (19) and (20) are approximate, based on the assumption that  $p$  is large compared to  $\sigma_0$ , which is realistic for many practical systems. *Conclusions reached from a study of eqs. (19) and (20) are:*

1. The time to peak,  $t_p$ , is inversely proportional to  $\omega_d$ , the damped natural frequency.

2. The addition of a pole increases  $t_p$  and decreases  $C/R|_p$ , the magnitude of the peak.

3. The addition of a zero decreases  $t_p$  and increases  $C/R|_p$ .

4. If a pole and zero are close together (dipole), their net effect on  $t_p$  and  $C/R|_p$  is negligible.

5. Poles and zeros far out on the real axis have little effect on  $t_p$  and  $C/R|_p$ .

If the assumption of the real poles and zeros being large compared with the damping factor ( $p \gg \sigma_0$ ) is not valid, the values of  $C/R|_p$  and  $t_p$  can still be estimated though not by the simple use of eqs. (19) and (20).

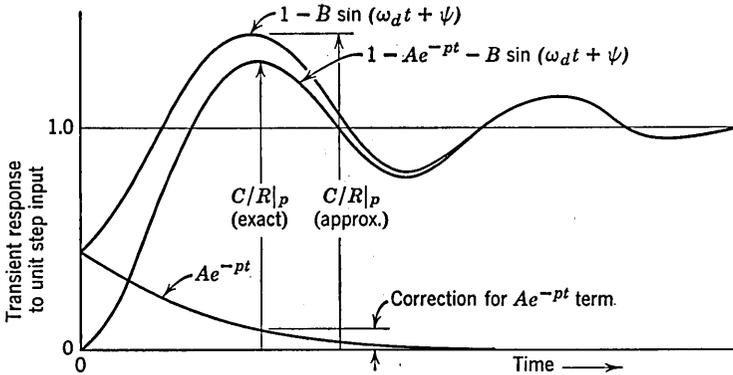


FIG. 8. Illustrating the effect of a significant real pole on  $C/R|_p$  approximations.

The magnitude of the coefficients of the exponential terms such as the one in eq. (15) can be calculated, as outlined in the next section, and then the effect of these terms on  $C/R|_p$  and  $t_p$  can be estimated. By referring to Fig. 8 as an *example*, it is apparent that the value of the simple exponential term at time,  $t_p$ , must be subtracted from the approximate curve to give a more exact value of  $C/R|_p$ . In other cases this correction might have to be added. Of course this process becomes more difficult as the number of significant poles and zeros increases.

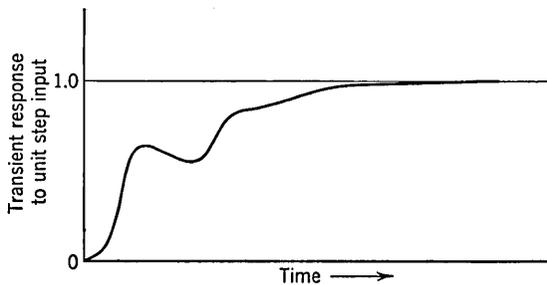


FIG. 9. Response with a dominant real pole and a pair of complex poles.

The addition of poles and zeros to the closed loop response function of a control system may result in a response function whose dominant charac-

teristic is that of the real pole rather than the complex pair. For such a case the system response to a unit step input might be as illustrated in Fig. 9.

### Coefficients of Transient Response Terms

Frequently, it is desired, as soon as the closed loop poles are found by graphical or other means, to determine the exact expression for the time response. This may be especially true when it is obvious that two pairs of complex poles are significant, i.e., they are both located about the same distance from the origin.

The coefficients of the terms in the equation describing the transient response of the system may be calculated by a formula developed in Laplace transform theory. This formula may be interpreted in terms of the pole-zero configuration of the root locus plot for the system.

To *illustrate* this a system described by the following equation is assumed.

$$(21) \quad \frac{C(s)}{R(s)} = \frac{K(s+z)}{(s+p_1)[(s+\sigma_0)^2 + \omega_d^2]}.$$

If the appropriate Laplace transform for  $R(s)$  is substituted into eq. (21), the expression for  $C(s)$  can be written. Assuming a unit step input in this case ( $R(s) = 1/s$ ), then

$$(22) \quad C(s) = \frac{K(s+z)}{s(s+p_1)[(s+\sigma_0)^2 + \omega_d^2]}.$$

In terms of partial fractions eq. (22) can be written as (see Chap. 20)

$$(23) \quad C(s) = K \left[ \frac{A_0}{s} + \frac{A_1}{s+p_1} + \frac{A_2}{s+(\sigma_0+j\omega_d)} + \frac{A_3}{s+(\sigma_0-j\omega_d)} \right].$$

The transient response equation for this system is

$$(24) \quad c(t) = K[A_0 + A_1 \exp(-p_1 t) + A_2 \exp(\sigma_0 + j\omega_d)t + A_3 \exp(\sigma_0 - j\omega_d)t].$$

**Determining the Coefficients.** The formulas for the coefficients  $A_0$ ,  $A_1$ , and  $A_3$  are:

$$(25) \quad A_0 = sC(s) \Big|_{s=0} = \frac{z}{p_1(\sigma_0^2 + \omega_d^2)},$$

$$(26) \quad A_1 = (s+p_1)C(s) \Big|_{s=-p_1} = \frac{z-p_1}{-p_1[(\sigma_0-p_1)^2 + \omega_d^2]},$$

$$\begin{aligned}
 (27) \quad A_2 &= [s + (\sigma_0 + j\omega_d)]C(s) \Big|_{s=-(\sigma_0+j\omega_d)} \\
 &= \frac{z - (\sigma_0 + j\omega_d)}{- (\sigma_0 + j\omega_d)[p_1 - (\sigma_0 + j\omega_d)][(\sigma_0 - j\omega_d) - (\sigma_0 + j\omega_d)]} \\
 &= \frac{z - (\sigma_0 + j\omega_d)}{2j\omega_d(\sigma_0 + j\omega_d)(p_1 - \sigma_0 - j\omega_d)}.
 \end{aligned}$$

$$\begin{aligned}
 (28) \quad A_3 &= [s + (\sigma_0 - j\omega_d)]C(s) \Big|_{s=-(\sigma_0-j\omega_d)} \\
 &= \frac{z - (\sigma_0 - j\omega_d)}{- (\sigma_0 - j\omega_d)[p_1 - (\sigma_0 - j\omega_d)][(\sigma_0 + j\omega_d) - (\sigma_0 - j\omega_d)]} \\
 &= \frac{z - (\sigma_0 - j\omega_d)}{-2j\omega_d(\sigma_0 - j\omega_d)(p_1 - \sigma_0 + j\omega_d)}.
 \end{aligned}$$

Note that these coefficients are the ratios of vectors in the root locus plot. For *example*, consider Fig. 10 which illustrates the pole-zero con-

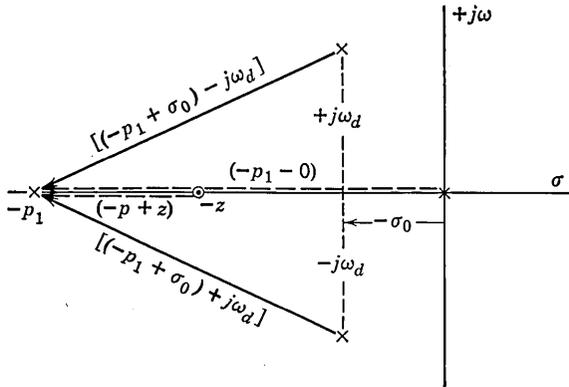


FIG. 10. Vectors for determining  $A_1$  coefficient.

figuration of the system under consideration. With this plot the value of any coefficient can be determined by drawing vectors from all other poles and zeros to the pole or root which corresponds to the coefficient. The coefficient is then the ratio of vectors from the zeros to those from the poles. Figure 10 shows the vectors for the calculation of  $A_1$ . From this

$$\begin{aligned}
 (29) \quad A_1 &= \frac{-p_1 + z}{-p_1[(-p_1 + \sigma_0) + j\omega_d][(-p_1 + \sigma_0) - j\omega_d]} \\
 &= \frac{z - p_1}{-p_1[(\sigma_0 - p_1)^2 + \omega_d^2]}
 \end{aligned}$$

which agrees with eq. (26).

Similarly, the other coefficients may be determined from the root locus plot. Figure 11 shows the same system with the vectors oriented for determination of the  $A_3$  coefficient for one of the complex roots. From this

$$(30) \quad A_3 = \frac{\lambda_3}{\lambda_1 \lambda_2 \lambda_4} = \frac{|\lambda_3| / \psi_3}{(|\lambda_1| / \psi_1)(|\lambda_2| / \psi_2)(|\lambda_4| / \psi_4)}$$

$$= \frac{|\lambda_3|}{j|\lambda_1| |\lambda_2| |\lambda_4|} \psi_3 - \psi_1 - \psi_4,$$

since  $\psi_2 = +j$  or  $+90^\circ$ .

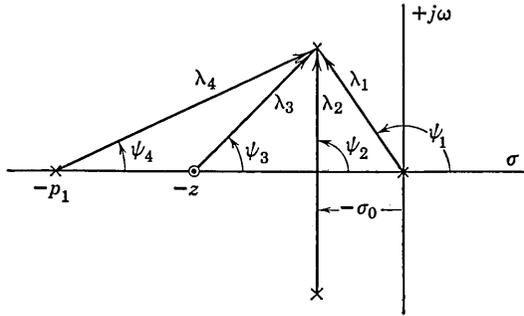


FIG. 11. Vectors for determining  $A_3$  coefficient.

In like manner the  $A_2$  coefficient can be determined as

$$(31) \quad A_2 = \frac{|\lambda_3| \angle -\psi_3}{(|\lambda_1| \angle -\psi_1)(|\lambda_2| \angle -\psi_2)(|\lambda_4| \angle -\psi_4)}$$

$$= \frac{|\lambda_3|}{-j|\lambda_1| |\lambda_2| |\lambda_4|} \angle -(\psi_3 - \psi_1 - \psi_4).$$

The  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$ , and  $\lambda_4$  vectors can be expressed in terms of the pole-zero locations in the root locus plot. When this is done eqs. (25) to (28) are the results. These coefficients can be evaluated conveniently by use of the Spirule, although a ruler and protractor will suffice (Ref. 1).

The *time response* of a system such as the one being considered here is usually expressed in equation form with a sine or cosine term instead of the complex exponents. This is illustrated in the following equations.

$$(32) \quad c(t) = A_0 + A_1 \exp(-p_1 t) + A_2 \exp(-\sigma_0 - j\omega_d)t$$

$$+ A_3 \exp(-\sigma_0 + j\omega_d)t.$$

From eqs. (30) and (31) the coefficients  $A_2$  and  $A_3$  may be expressed as

$$(33) \quad A_3 = \frac{|A_3|}{j} \angle A_3,$$

$$(34) \quad A_2 = -\frac{|A_3|}{j} \angle -A_3.$$

Combining these equations with eq. (32) yields

$$(35) \quad c(t) = A_0 + A_1 \exp(-p_1 t) + \exp(-\sigma_0 t) \times \\ \left[ \frac{|A_3|}{j} \exp(-j\omega_d t - j\angle A_3) + \frac{|A_3|}{j} \exp(j\omega_d t + j\angle A_3) \right] \\ = A_0 + A_1 \exp(-p_1 t) + 2 |A_3| \exp(-\sigma_0 t) \sin(\omega_d t + \angle A_3),$$

where  $|A_3| = \frac{|\lambda_3|}{|\lambda_1| |\lambda_2| |\lambda_4|}$   
 $\angle A_3 = \angle \psi_3 - \psi_1 - \psi_4$

and these vector lengths and angles are shown in Fig. 11.

#### 4. RELATION BETWEEN CLOSED LOOP AND OPEN LOOP ROOTS

**Mathematical Relationship.** The closed loop frequency response function may be readily written in terms of the open loop function as

$$(36) \quad \frac{C(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)},$$

where  $G(s)$  = open loop transfer function of forward element  
 $H(s)$  = feedback element.

If these are written as

$$(37) \quad G(s) = N_1(s)/D_1(s); \quad H(s) = N_2(s)/D_2(s)$$

then

$$(38) \quad \frac{C(s)}{R(s)} = \frac{N_1(s)/D_1(s)}{1 + N_1(s)N_2(s)} = \frac{N_1(s)D_2(s)}{D_1(s)D_2(s) + N_1(s)N_2(s)}.$$

The closed loop poles are thus the roots of the denominator of eq. (38). This is generally a high order polynomial and constitutes a tedious task if it is to be factored. For this reason methods of estimating closed loop roots from open loop roots are useful.

**Graphical Method of Determining Roots.** A convenient way to obtain the closed loop poles {roots of  $[N_1(s) N_2(s) + D_1(s) D_2(s)]$ } is to use the root locus technique of graphically plotting loci of the closed loop poles as functions of the open loop system gain (see Chap. 21). These loci may be plotted from the open loop pole-zero configurations on the complex frequency plane as shown in Chap. 21. The plotting is simplified considerably by use of the Spirule plotting tool.

The selection of the open loop gain to give the proper system response may be determined either by use of frequency response or root locus synthesis techniques. In either case the configuration may be examined for its transient response characteristics.

A method of performing the reverse operation, working from closed loop to open loop roots, is presented in Chap. 23.

**An Iterative Process of Determining Closed Loop Roots.** The closed loop poles {roots of  $[N_1(s) N_2(s) + D_1(s) D_2(s)]$ } can be found more easily by mathematical techniques if known roots can be factored out and leave a simpler polynomial. A technique is available (Ref. 2) which allows closed loop poles to be found after the open loop response characteristics of the system, including the open loop gain, have been determined.

*The rules for determining closed loop poles by this method are the following:*

1. An open loop zero located at a frequency lower than the crossover frequency,  $\omega_c$ , is approximately equal to a closed loop pole.
2. An open loop pole located at a frequency higher than the crossover frequency,  $\omega_c$ , is approximately equal to a closed loop pole.

These rules are used to make the first approximations for the closed loop poles. These approximate values may be refined by *iteration* using the following expressions:

*For open loop zeros much smaller than  $\omega_c$ ,*

$$(39) \quad (p_i + z_1)^n \approx - \left. \frac{(s + z_1)^n}{H(s)G(s)} \right|_{s=p_{i-1}},$$

where  $p_i$  = closed loop pole,

$z_1$  = open loop zero much smaller than  $\omega_c$ ,

$p_{i-1}$  = value of  $p_i$  found by previous iteration (equals  $z_1$  for first iteration),

$n$  = order of open loop zero.

*For open loop poles much larger than  $\omega_c$ ,*

$$(40) \quad (p_i + p_1)^n \approx - (s + p_1)^n H(s)G(s) \Big|_{s=p_{i-1}},$$

where  $p_i$  = closed loop pole,

- $p_1$  = open loop pole much larger than  $\omega_c$ ,  
 $p_{i-1}$  = value of  $p_i$  found by previous iteration (equals  $p_1$  for first iteration),  
 $n$  = order of open loop pole.

If  $n$  is greater than unity,  $n$  values for  $p_i$  will result with each iteration. Further refinement should continue only on those values of  $p_i$  remaining far from  $\omega_c$ . If the value of  $p_i$  approaches  $\omega_c$  the accuracy of the technique is poor.

After these closed loop poles are found with sufficient accuracy by iteration, they may be factored from the closed loop polynomial characteristic equation for the system. The resulting lower order polynomial may then be more easily factored.

In general if the open loop poles and zeros are larger or smaller than  $\omega_c$  by a ratio of 3 to 1 or greater, two iterations will result in sufficient accuracy for finding the first closed loop poles.

The coefficients of these transient response terms may be calculated as indicated in the previous section.

**EXAMPLE. Determining Roots.** Assume the open loop transfer function

$$(41) \quad G(s)H(s) = \frac{400(s+1)}{s(s+2)(s+10)^2}.$$

As previously stated: open loop zeros less than  $\omega_c$  are approximate closed loop poles. The crossover frequency,  $\omega_c$ , is 4 rad per second, as may be easily determined from a graphical plot of eq. (41). Therefore,

$$(42) \quad p_{i-1} = -1.0.$$

To refine this approximation

$$(43) \quad \begin{aligned} (p_i + 1) &\approx \left. \frac{-s(s+2)(s+10)^2}{400} \right|_{s=-1} \\ &\approx - \left[ \frac{-1(-1+2)(-1+10)^2}{400} \right] \\ &\approx 0.20. \end{aligned}$$

$$(44) \quad p_i \approx -1 + 0.20 = -0.80.$$

This may be repeated,

$$(45) \quad \begin{aligned} (p_i + 1) &\approx - \left[ \frac{-0.8(-0.8+2)(-0.8+10)^2}{400} \right] \\ &\approx 0.20. \end{aligned}$$

$$(46) \quad p_i \approx -0.80.$$

Similarly, open loop poles larger than  $\omega_c$  are approximate closed loop poles, so

$$(47) \quad p_i = -10, -10,$$

$$(48) \quad (p_i + 10)^2 = \left. \frac{-400(s+1)}{s(s+2)} \right|_{s=-10} \\ \approx + \frac{400}{10} = +40.$$

$$(49) \quad p_i \approx -3.7, -16.3.$$

Since the  $\omega_c$  is 4, only the larger root can be expected to be useful. With it, repeating the process gives

$$(50) \quad (p_i + 10)^2 = \frac{-400(-16.3 + 1)}{-16.3(-16.3 + 2)} = 26.3,$$

$$(51) \quad p_i = -4.9, -15.1.$$

Again,

$$(52) \quad (p_i + 10)^2 = \frac{-400(-15.1 + 1)}{-15.1(-15.1 + 2)},$$

$$(53) \quad p_i = -15.35.$$

The two closed loop poles thus determined are  $s = -0.80, -15.35$ , and hence they may be factored out of the expression

$$(54) \quad N_1(s)N_2(s) + D_1(s)D_2(s) = 400(s+1) + s(s+2)(s+10)^2 \\ = s^4 + 22s^3 + 140s^2 + 600s + 400$$

to give the closed loop poles near  $\omega_c$  as:

$$(55) \quad s^2 + 5.85s + 33.3 = (s - 2.93 + j5.34)(s - 2.93 - j5.34).$$

Thus, the closed loop poles are:

$$(56) \quad s = -0.80, -15.35, (-2.93 + j5.34), (-2.93 - j5.34).$$

*In this example if  $H(s)$  is other than unity these roots are not the roots of  $C(s)/R(s)$ , but are the roots of  $H(s)[C(s)/R(s)]$ .*

## 5. DESIGN CHARTS RELATING OPEN LOOP FREQUENCY RESPONSE AND TRANSIENT RESPONSE

An approximate method of relating steady-state frequency response characteristics and transient response characteristics has been described (Ref. 3). It makes use of a series of charts which indicate the type of open loop attenuation curves required to produce desired closed-loop responses.

If a servo system falls within the group considered in the charts, this method enables the designer to take a set of specifications setting forth steady-state frequency and/or transient response requirements and quickly estimate the necessary open loop characteristics. The charts also permit the designer to estimate the effect of changing various system parameters to give him a better understanding of the system.

**Description of Charts.** The symbols used on the charts are defined below and illustrated in Fig. 12.

$\left. \frac{C}{R} \right _m$	maximum ratio of the closed loop frequency response ( $M_m$ )
$\left. \frac{C}{R} \right _p$	peak value of the ratio of controlled variable to input for a step function input
$\frac{\omega_m}{\omega_c}$	ratio of the frequency $\omega_m$ at which $C/R _m$ occurs to the frequency $\omega_c$ at which the straight-line approximation of the open loop response is 0 db
$\frac{\omega_l}{\omega_c}$	ratio of $\omega_l$ , the lowest frequency of oscillation for a step input, to $\omega_c$ , the frequency at which the straight-line approximation of the open loop response is 0 db
$\omega_c t_p$	the frequency, $\omega_c$ , at which the straight-line approximation of the open loop response is 0 db times the response time $t_p$ measured from the start of the step function until $C/R _p$ occurs
$\omega_c t_s$	the frequency, $\omega_c$ , at which the straight-line approximation of the open loop response is 0 db times the settling time, $t_s$ , from the start of the step function until the output continues to differ from the input by less than 5 per cent

Indicated in Figs. 12*a*, *b*, and *c* are these various characteristics in terms of the familiar curves of the open loop transfer function, the frequency response, and the transient response to a step input. The charts, Figs. 13 to 30, were prepared for a system with an initial open loop attenuation, Fig. 12*a*, of 20 db per decade. However, the shape of the curve near 0 db is of greatest importance, so the curves may also be used for systems with initial attenuation slopes of 0 to 40 db per decade.

**Limitations.** Of necessity the charts may be used for the analysis and synthesis of a somewhat restricted class of servomechanisms. Their use is restricted to:

(a) Linear systems, or those which may be considered linear for a restricted range of operation.

FEEDBACK CONTROL

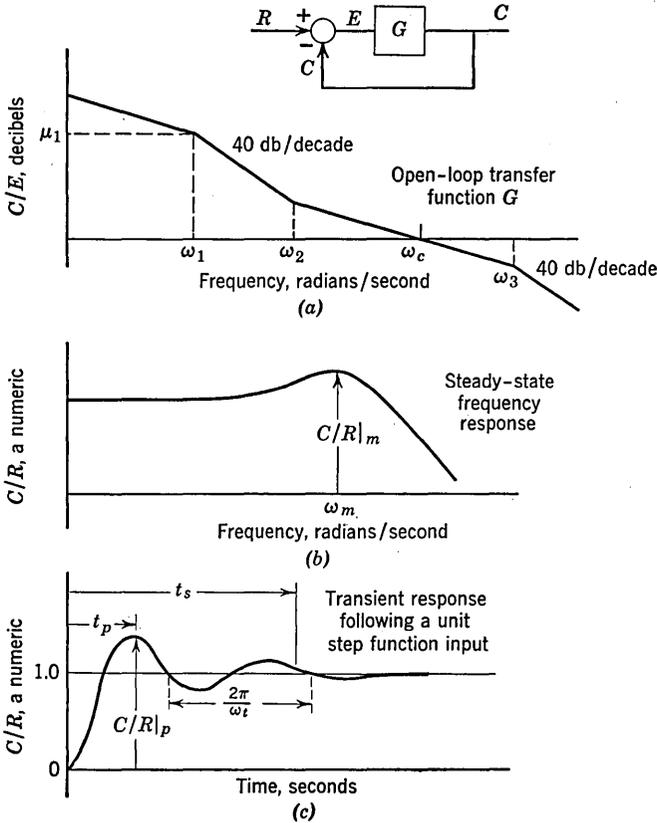


FIG. 12. Sketches showing nomenclature used in the design charts of Figs. 13 to 30.

(b) Single loop, unity feedback systems containing only series elements; of course a multiloop system may be considered if the inner loops are reduced to equivalent series elements.

(c) Systems whose open loop characteristics fall into the category of servomechanisms described by Fig. 12. However, systems which ostensibly are not of this type may often be approximated by some which are, especially if the required approximations occur at an appreciable distance from the crossover frequency,  $\omega_c$ .

(d) A step function as the form of the input signal producing the transient response.

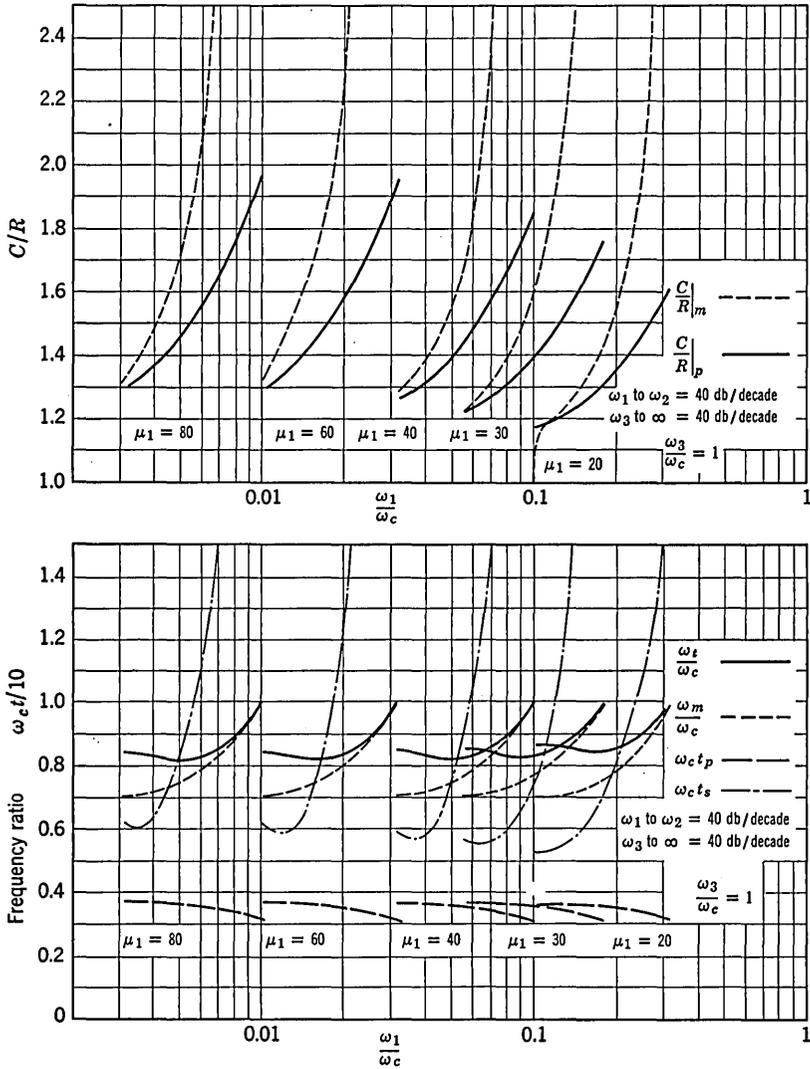


FIG. 13. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

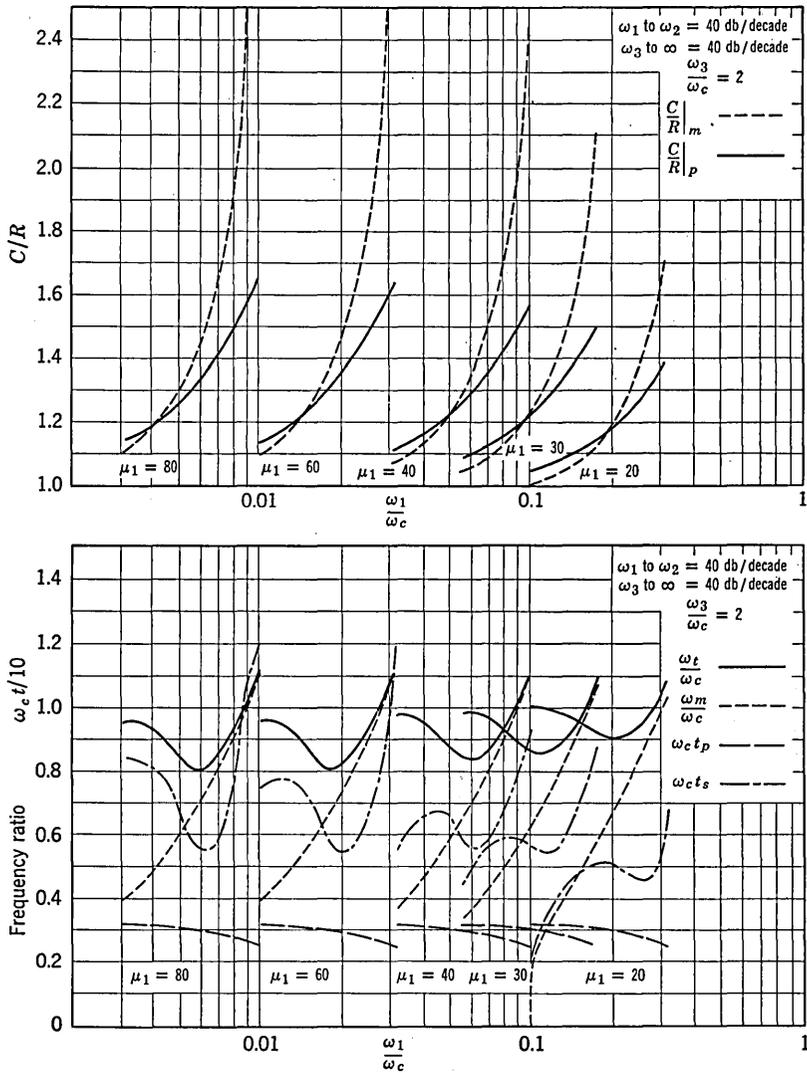


FIG. 14. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

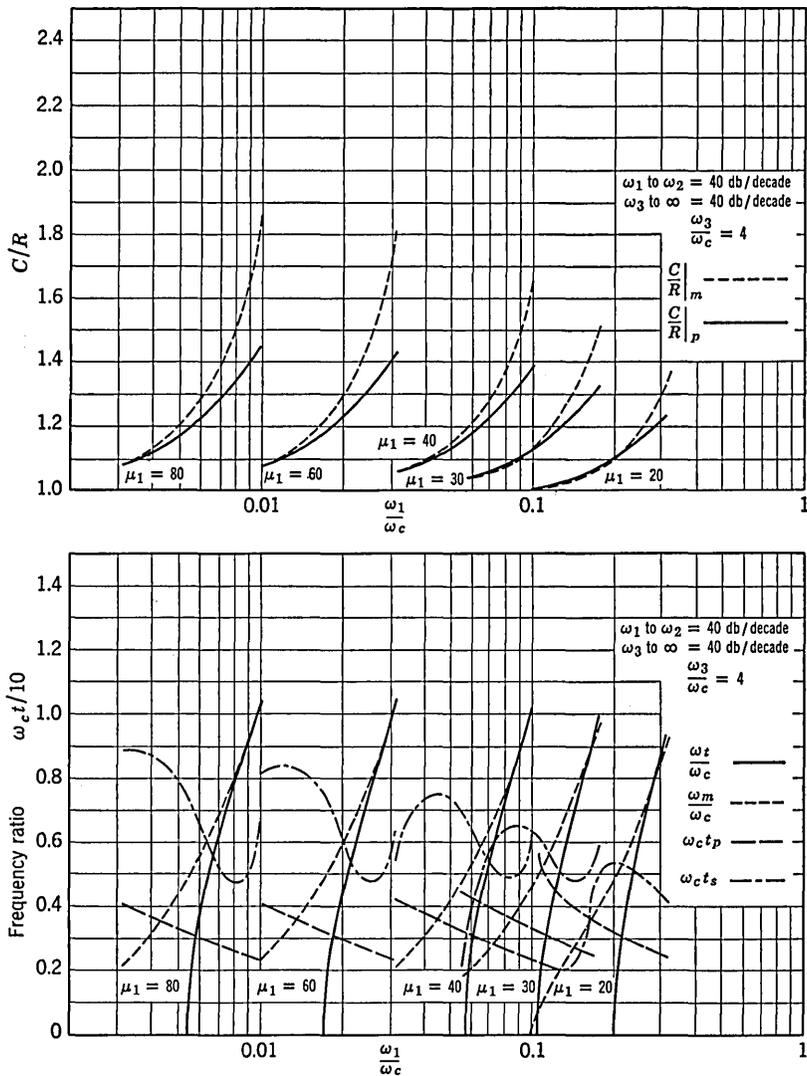


FIG. 15. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

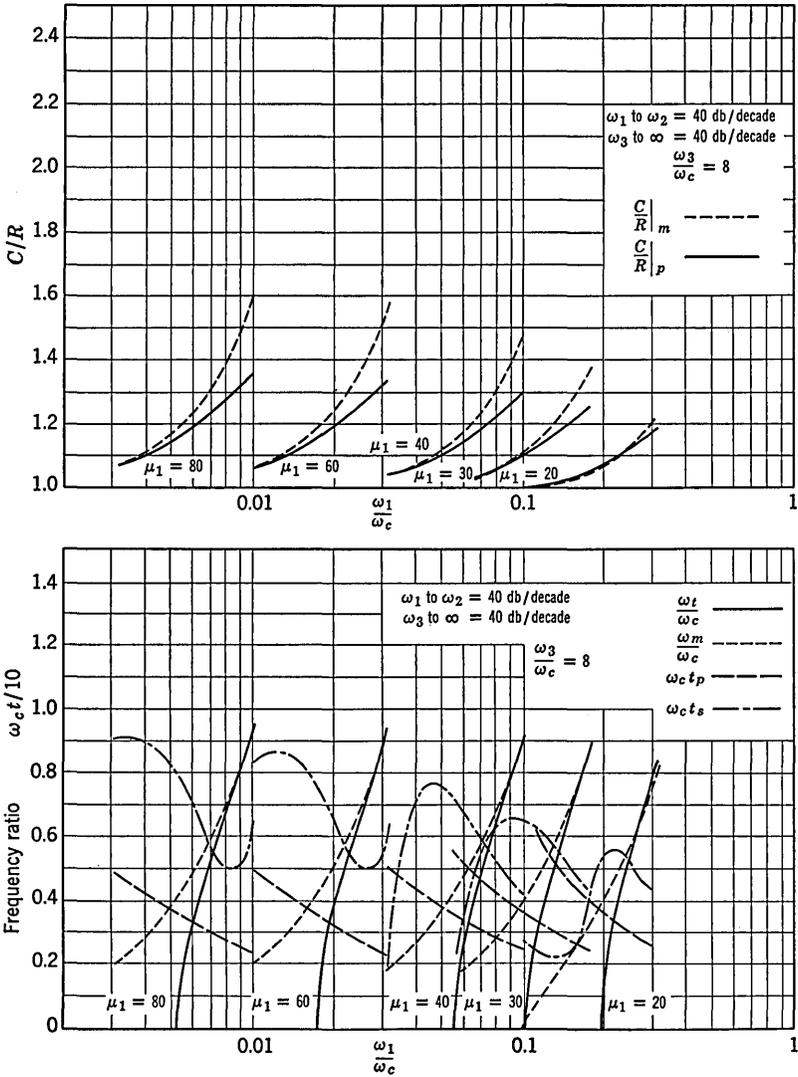


Fig. 16. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

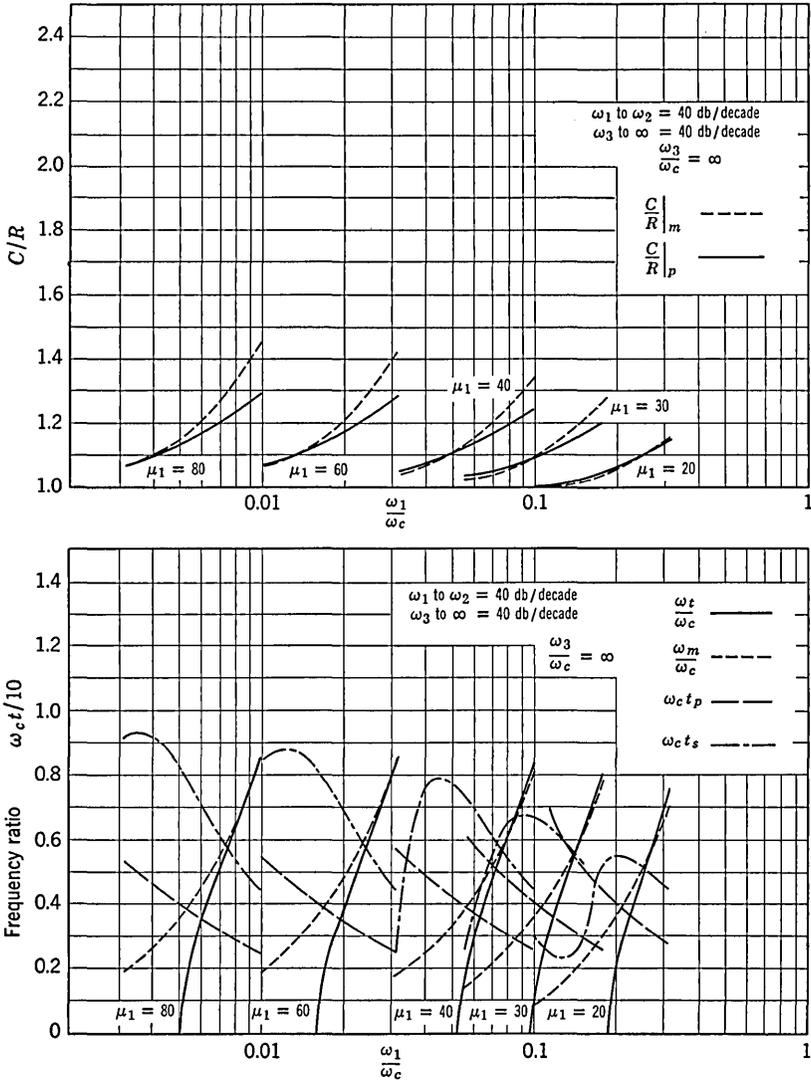


FIG. 17. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

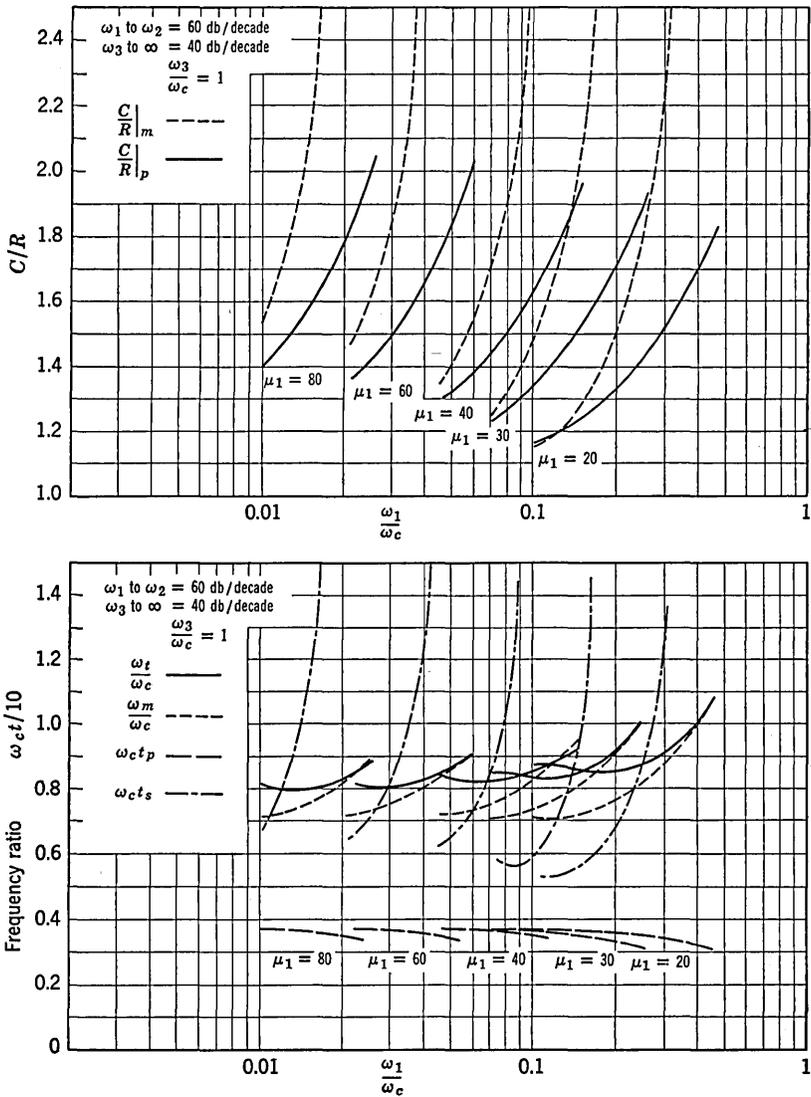


Fig. 18. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

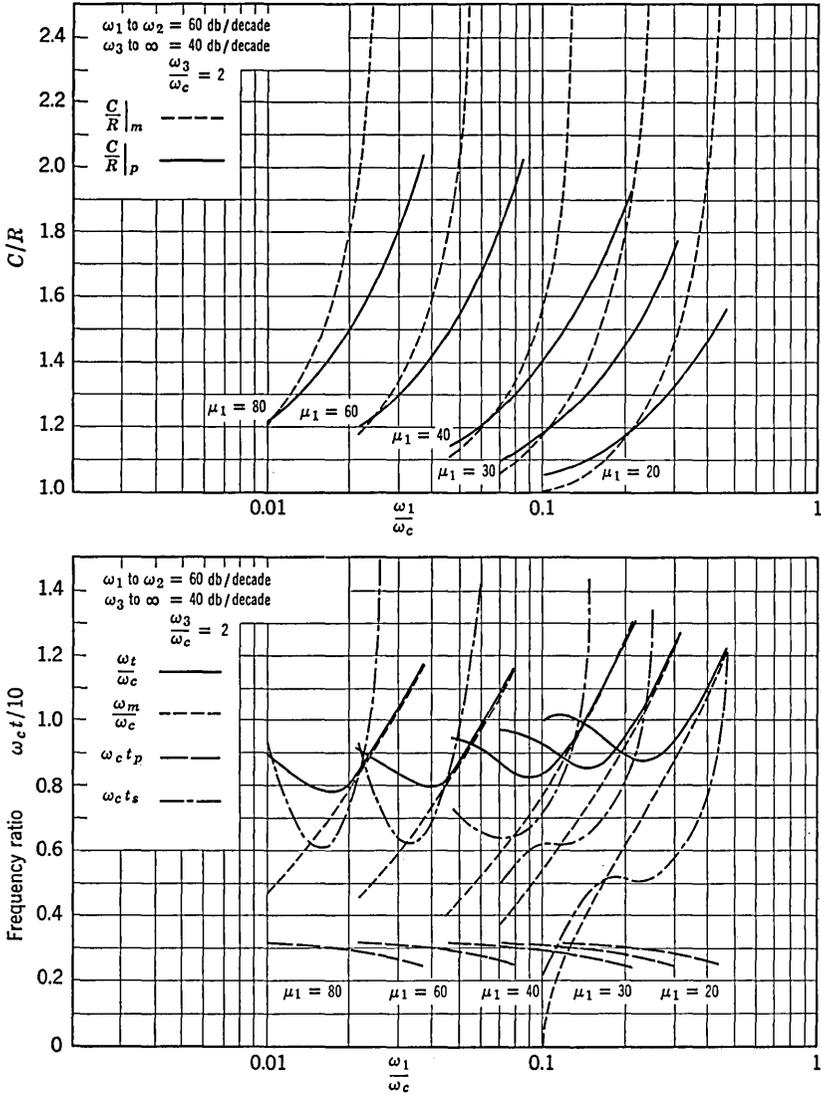


FIG. 19. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

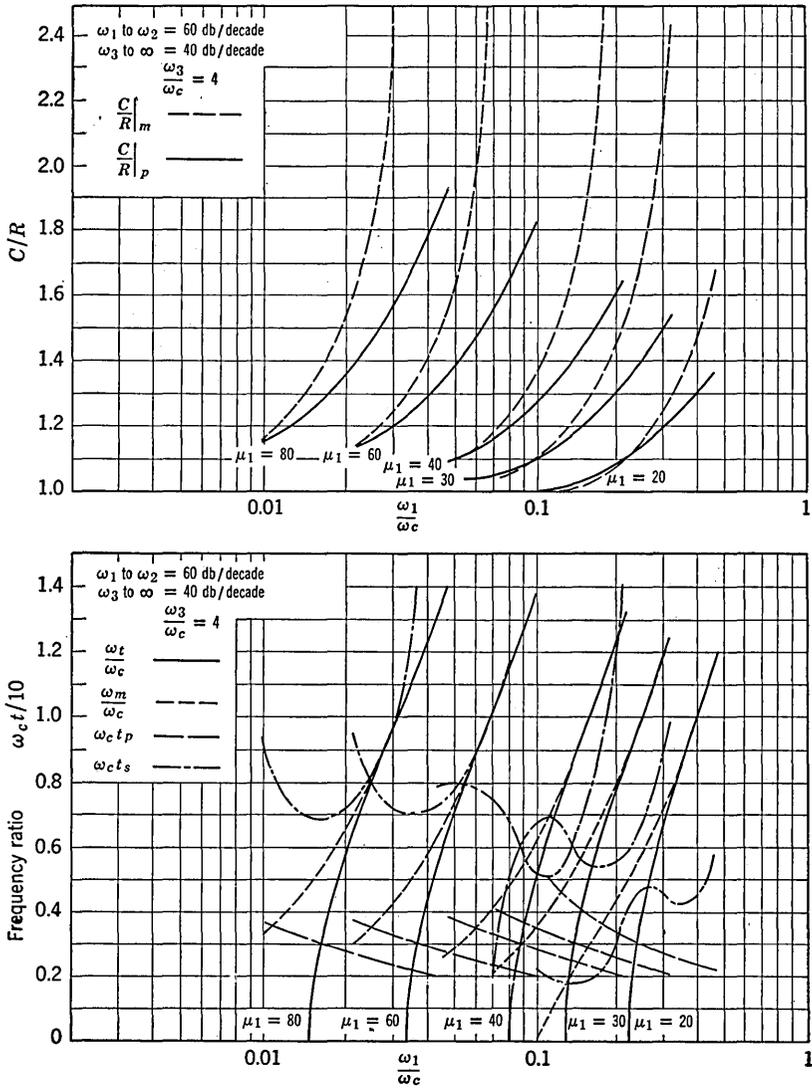


FIG. 20. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

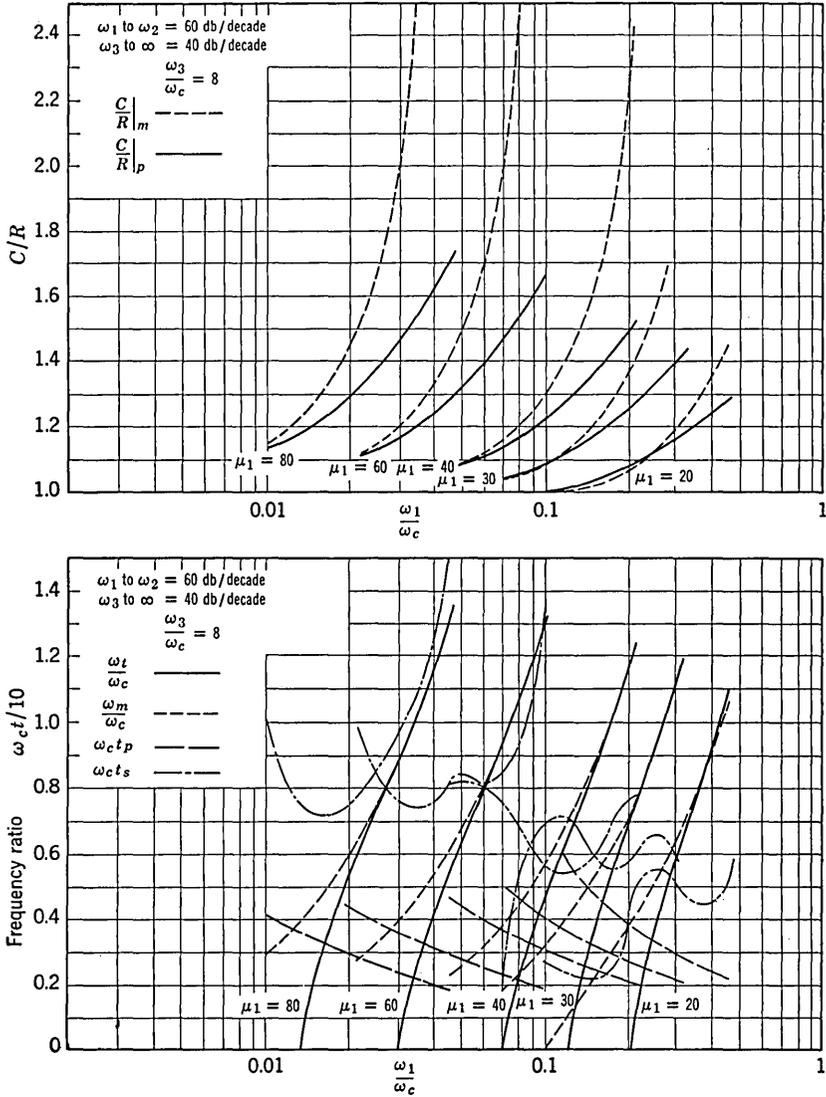


FIG. 21. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

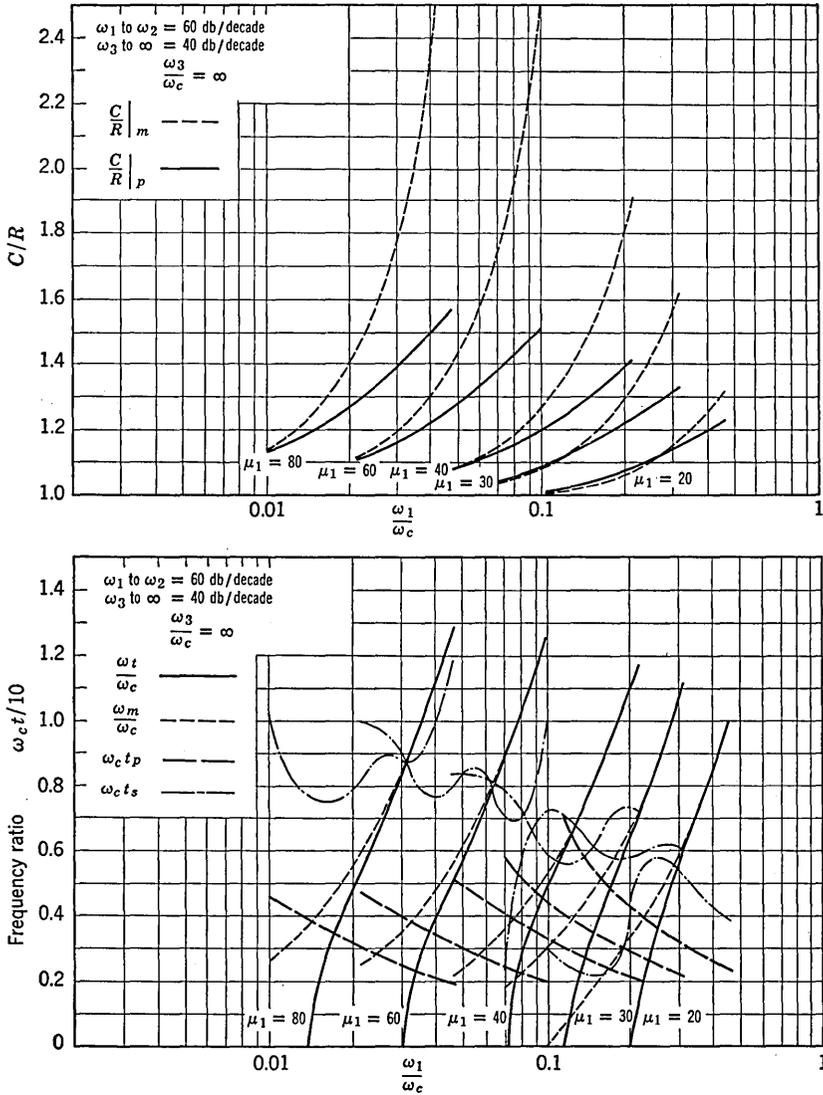


FIG. 22. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

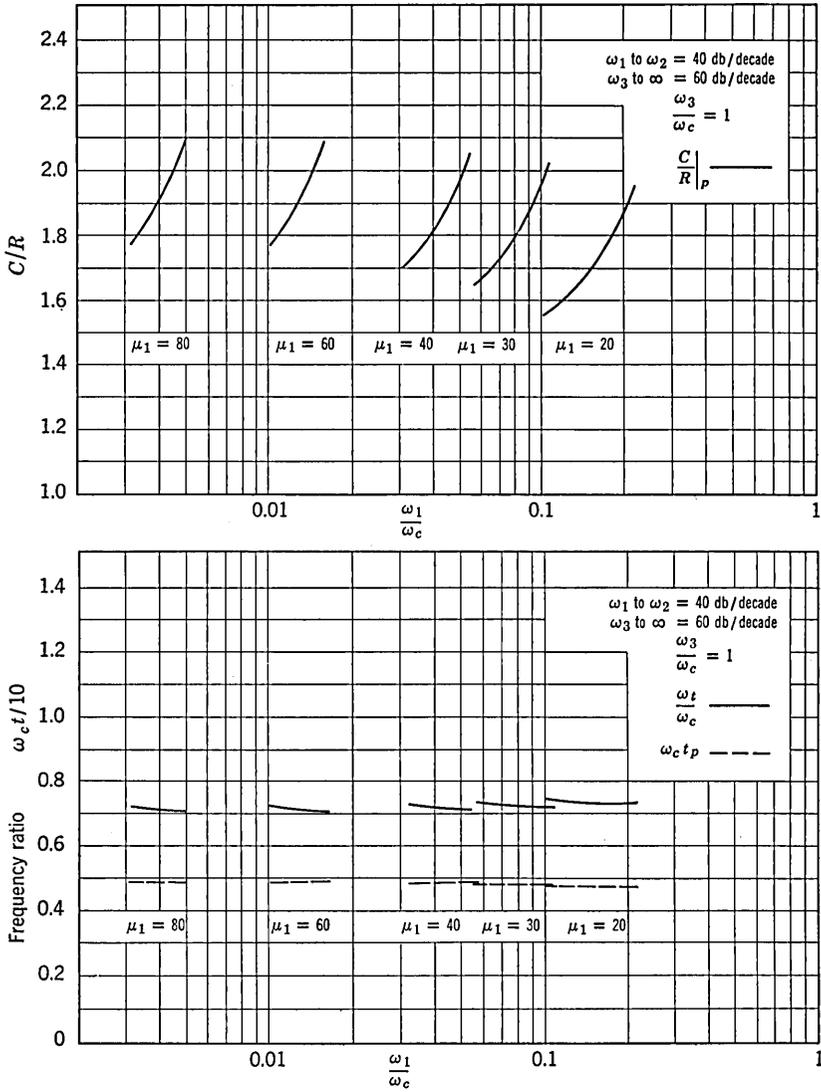


FIG. 23. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

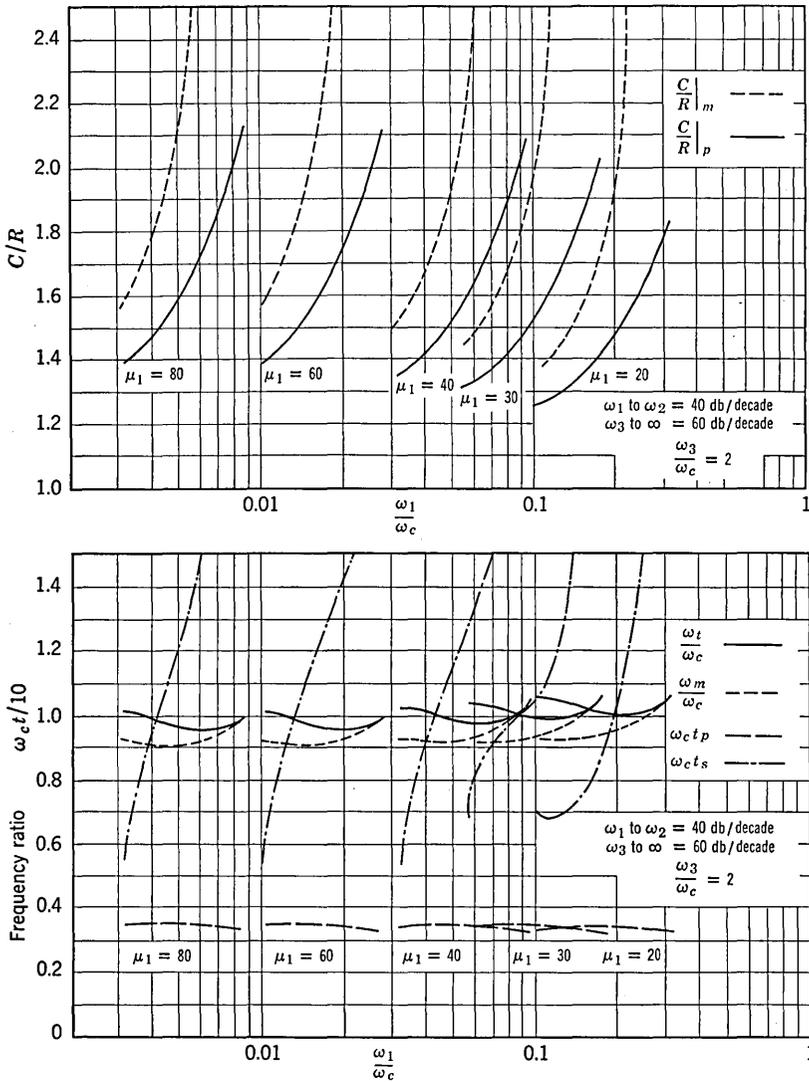


FIG. 24. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

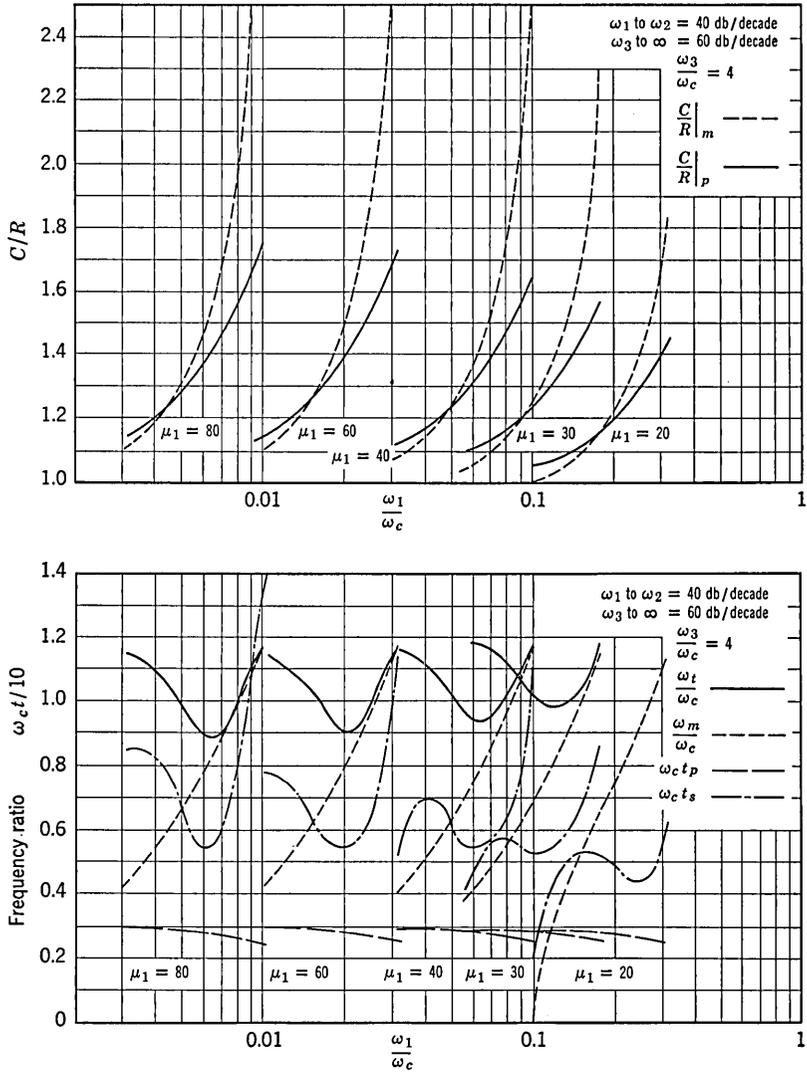


FIG. 25. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

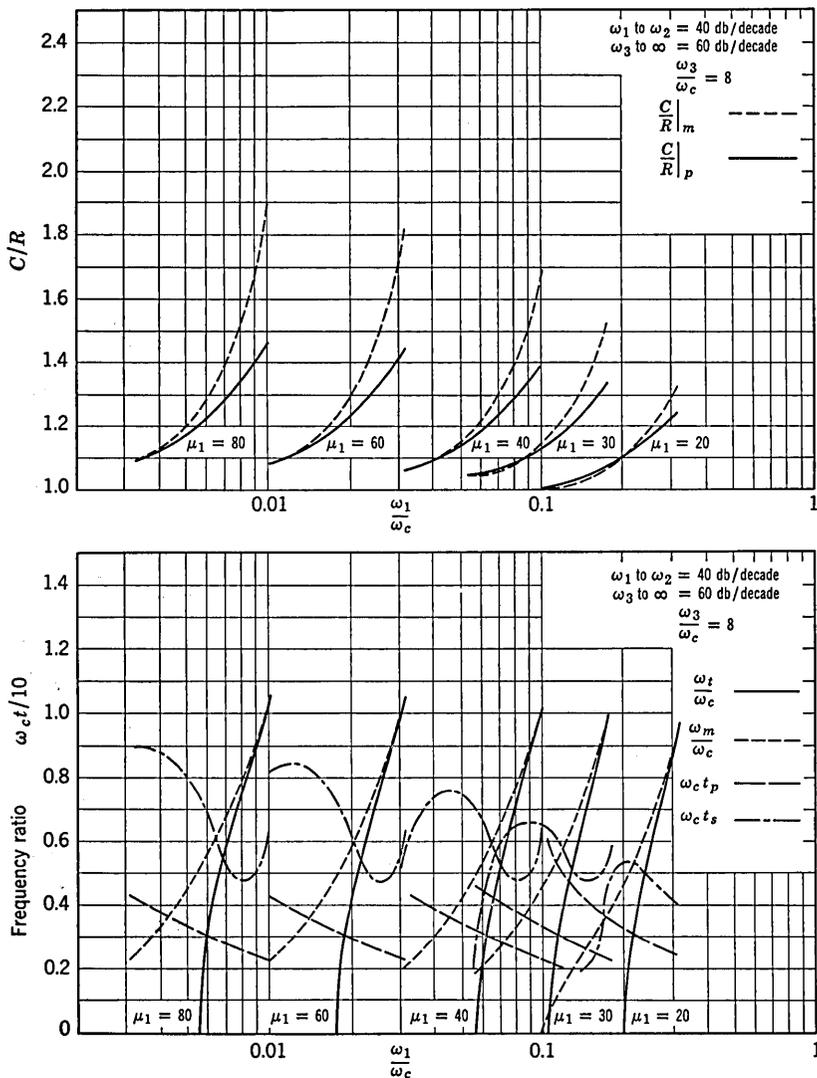


FIG. 26. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

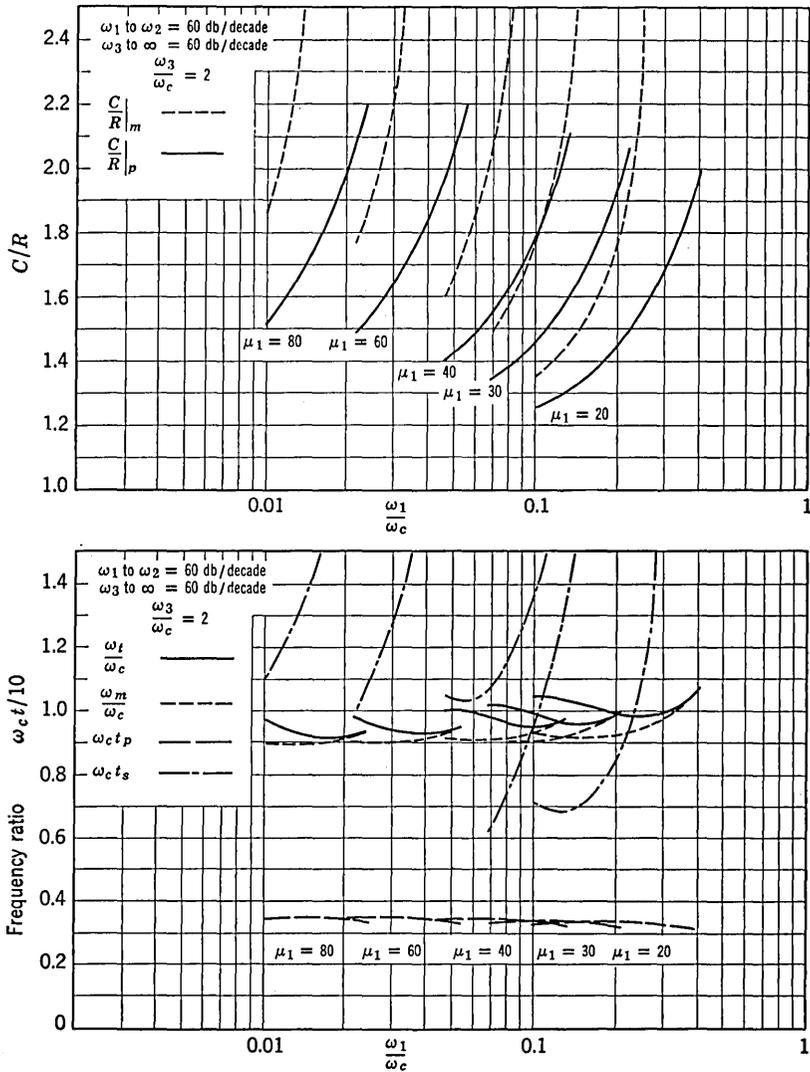


Fig. 27. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

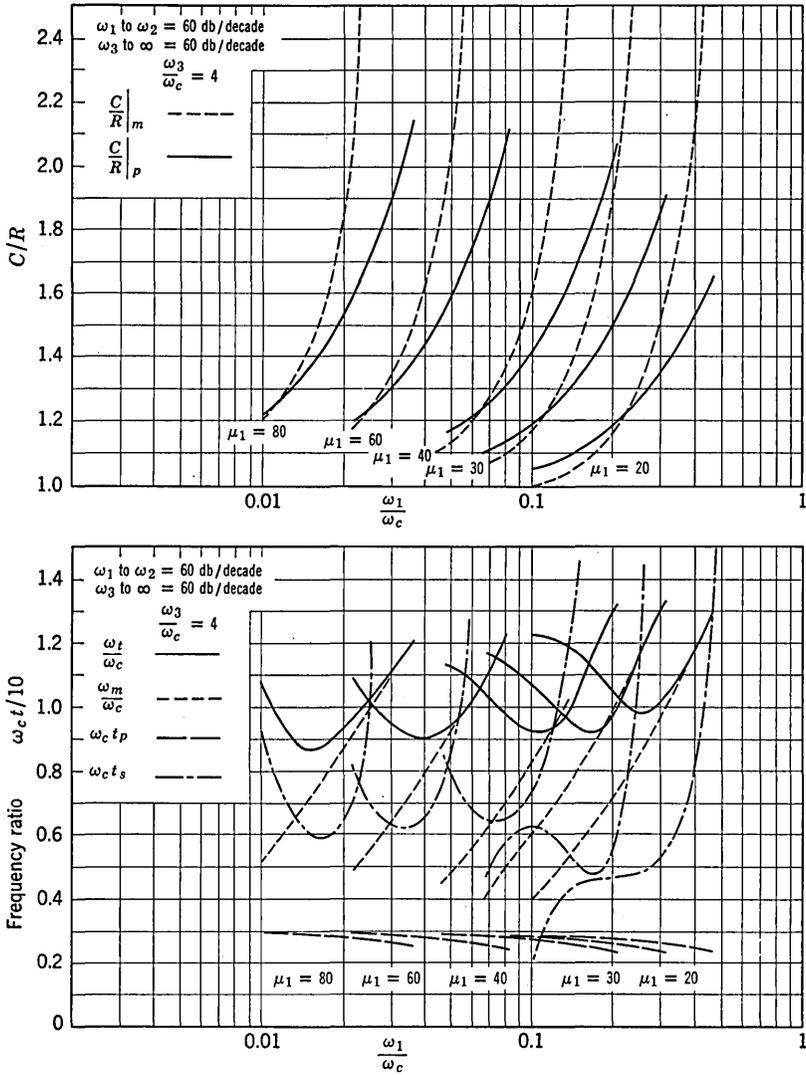


FIG. 28. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

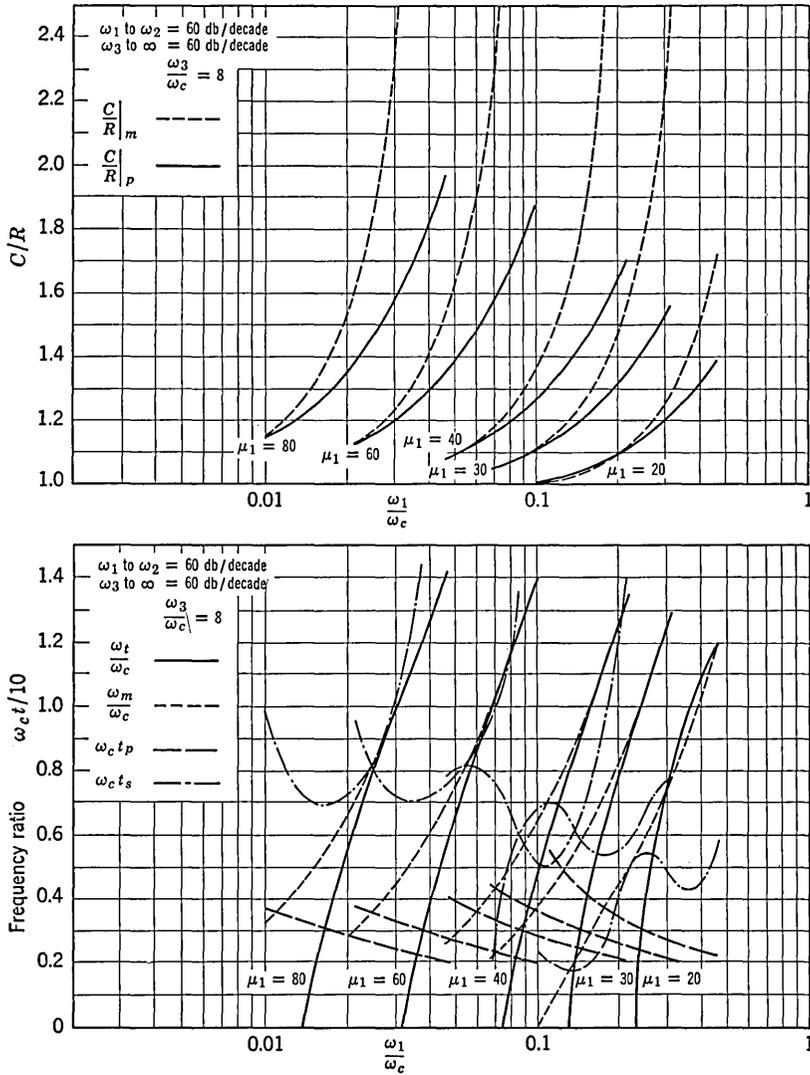


FIG. 29. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3).

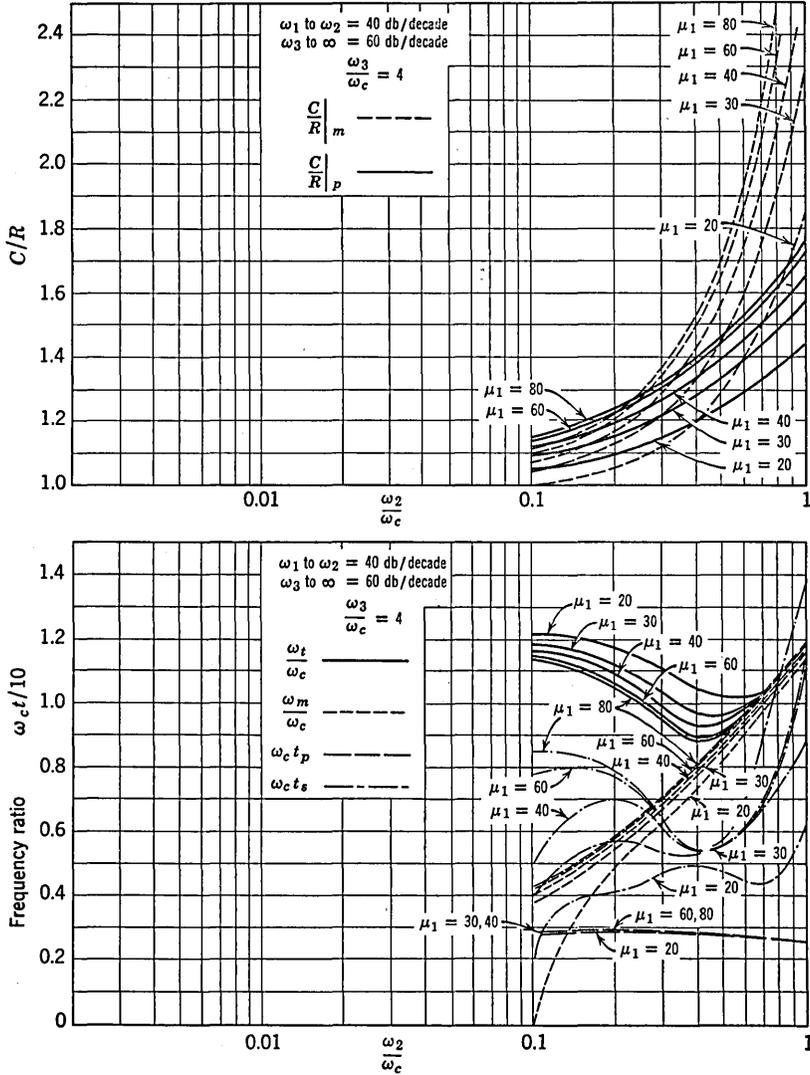


FIG. 30. Charts giving comparison of steady-state frequency response and transient response following a step. See Fig. 12 for definition of nomenclature (Ref. 3). Note that  $\omega_2/\omega_c$  is the abscissa for these charts.

**Uses of Charts.** Figures 13 through 30 are comparisons of steady-state frequency response characteristics and transient response following a step function of input as a function of  $\omega_1/\omega_c$  (Ref. 3). The information presented in Figs. 13 to 30 is useful for analysis, that is, determining the response of systems already designed, or for synthesis, that is, determining what sort of system will be required to do a specified job. Typical examples of each are presented in the following.

**EXAMPLE 1. Analysis.** Determine, approximately, the value of  $C/R|_m$  and the frequency at which it occurs, and the magnitude of the peak over-

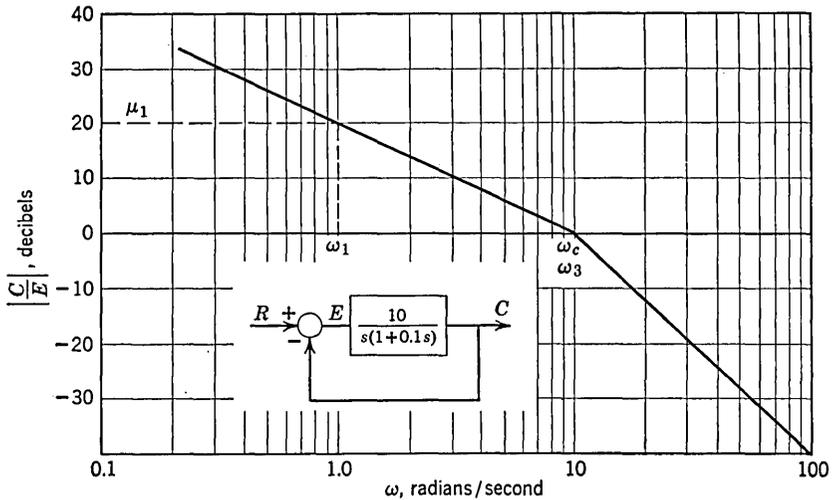


FIG. 31. Gain plot for example problem.

shoot to a step function input and the time when it occurs for a system having the open loop transfer function,

$$(57) \quad C/E = 10/s(1 + 0.1s),$$

which is drawn in Fig. 31. From this,

$$(58) \quad \begin{aligned} \omega_3 &= \omega_c = 10, \\ \omega_1 &= \omega_2 = 1.0, \\ \mu_1 &= 20 \text{ db.} \end{aligned}$$

The values for  $\omega_1$  and  $\mu_1$  were arbitrarily selected; of course, choosing a value for one fixes the value of the other. Since there is no segment between  $\omega_1$  and  $\omega_2$  with either 40 or 60 db per decade slope the chart with

the lower attenuation rate 40 will be used. By entering the chart in Fig. 13 with the following parameters:

$$(59) \quad \begin{aligned} \mu_1 &= 20, & \omega_1 \text{ to } \omega_2 &= 40 \text{ db/decade,} \\ \omega_1/\omega_c &= 0.1, & \omega_3 \text{ to } \infty &= 40 \text{ db/decade,} \\ \omega_3/\omega_c &= 1.0, \end{aligned}$$

the desired information may be obtained.

$$(60) \quad \begin{aligned} C/R|_m &\approx 1.16, & C/R|_p &\approx 1.18, \\ \omega_m/\omega_c &\approx 0.7, & \frac{\omega_c t_p}{10} &\approx 0.36, \\ \omega_m &\approx (0.7)(10) = 7.0, & t_p &\approx 0.36 \text{ sec.} \end{aligned}$$

**EXAMPLE 2. Synthesis.** The requirements of a position control system are assumed to be set forth in the following set of specifications:

1.  $C/R|_m = 1.3$  or less.
2.  $\omega_m = 2$  cycles per second or more.
3. Velocity error coefficient ( $K_v$ ) is  $200 \text{ sec}^{-1}$ .
4. The attenuation rate of the open loop control will be 60 db per decade for frequencies greater than 120 rad per second.

The problem is to determine the open loop transfer function of a suitable control for this application.

The frequency,  $\omega_c$ , must be considerably greater than  $\omega_m$ , so as a first assumption assume that  $\omega_c = 30$  rad per second, and that  $\omega_3 = 120$  rad per second. For the specifications given there are many solutions to the problem. Figure 28, for which  $\omega_3/\omega_c = 4$ , shows that for  $\mu_1 = 40$  db and  $\omega_1/\omega_c = 2/30 = 0.067$ .

$$(61) \quad \begin{aligned} C/R|_m &= 1.25, \\ \omega_m/\omega_c &= 0.6, \\ \omega_m &= (0.6)(30) = 18 \text{ rad/sec.} \end{aligned}$$

Thus, an open loop transfer function satisfying the specifications is

$$(62) \quad \frac{C(s)}{E(s)} = \frac{200(1 + 0.2s)^2}{s(1 + 0.5s)^2(1 + 0.00833s)^2}$$

Synthesis by means of the charts is basically a trial-and-error process—assuming the solution and checking it.

TABLE 1. RULE-OF-THUMB APPROXIMATIONS

Parameter	Approximation	Remarks
Time to peak	$t_p \approx \pi/\omega_c$ where $t_p$ = time from step input to peak value of response transient, seconds $\omega_c$ = open loop crossover frequency, radians/second	In Chestnut and Mayer's charts it is evident that for this general class of servomechanisms, those with a dominant complex pair of closed loop poles, the open loop crossover frequency, $\omega_c$ , times the time to peak, $t_p$ , is about 3 or $\pi$ . In other words, the time to peak is about half the period corresponding to the open loop crossover frequency.
Peak overshoot	$C/R _p \approx 0.85M_m$ where $C/R _p$ = peak value of transient response to a step input $M_m$ = maximum value of closed loop frequency response	The peak value of the transient response, $C/R _p$ , to a unit step input is generally less than the maximum steady-state value, $M_m$ , of the closed loop frequency response. The maximum value of $C/R _p$ generally approaches 2.0 while the maximum value of $M_m$ approaches infinity. For many applications "good" servos are those with the values of $M_m$ between 1.3 and 1.5. For these servos $M_m$ is generally 10 to 20% greater than $C/R _p$ .
Damping ratio	$\zeta = 1/(2M_c)$ where $\zeta$ = damping ratio $M_c$ = value of closed loop frequency response at the corner frequency	The damping ratio may be approximated from the value of the closed loop frequency response of the system at the corner frequency, $\omega_c$ (the frequency at which the lines asymptotic to the log magnitude curve intersect). This is exact for a second order system. Of course this relationship may also be used to estimate $M_c$ , knowing the damping ratio. In addition, $M_c$ is approximately equal to $M_m$ for systems with low damping ratios.
Settling time	$t_s(5\%) \approx 3\sqrt{1 - \zeta^2}/\zeta\omega_d$ $t_s(2\%) \approx 5\sqrt{1 - \zeta^2}/\zeta\omega_d$ $t_s(5\%) \approx 3T_{eq}$	The settling time, $t_s$ , is generally defined as the time for the system to settle to within 5 or

TABLE 1. RULE-OF-THUMB APPROXIMATIONS (*Continued*)

Parameter	Approximation	Remarks
Settling time ( <i>continued</i> )	<p>where <math>t_s</math> = time for response to step input to settle to within some per cent of final value, seconds</p> <p><math>T_{eq}</math> = time for response to reach 63% of final value</p> <p><math>\omega_d</math> = damped natural frequency, radians/second</p> <p><math>\zeta</math> = damping ratio</p>	<p>sometimes 2% of the final value. In either case it is quite difficult to predict <math>t_s</math> for an underdamped system because it is subject to fluctuations of about one-half the period of oscillation for only small changes in system parameters. However, approximations (see eq. 18) can be made. The last approximation is for an overdamped system.</p>
Equivalent time constant	<p><math>T_{eq} \approx 1/\omega_c</math></p> <p>where <math>T_{eq}</math> = time for response to step input to reach 63% of final value, seconds</p> <p><math>\omega_c</math> = gain crossover frequency, radians/second</p>	<p>This relationship is exact for a simple single time constant system, but also quite good for the general case (Ref. 4).</p>
Oscillation frequency	<p><math>\omega_t \approx \omega_m \approx 0.75\omega_c</math></p> <p>where <math>\omega_t</math> = oscillation frequency of transient response, radians/second</p> <p><math>\omega_m</math> = frequency at which <math>M_m</math> occurs, radians/second</p> <p><math>\omega_c</math> = open loop gain crossover frequency, radians/second</p>	<p>The frequency of oscillation of the transient response, <math>\omega_t</math>, is generally about equal to the frequency, <math>\omega_m</math>, at which the frequency response peak, <math>M_m</math>, occurs. Both <math>\omega_m</math> and <math>\omega_t</math> are usually less than <math>\omega_c</math>, the open loop crossover frequency. For the "good" servos with <math>M_m = 1.3</math> to 1.5 an approximate relationship is as indicated. In this approximation <math>\omega_t</math> is used to mean essentially the same thing as <math>\omega_d</math>, the damped natural frequency, previously defined for a system with a dominant complex pair of poles. The use of <math>\omega_t</math> places no restriction on the system characteristics; however, generally, there is no significant difference between <math>\omega_t</math> and <math>\omega_d</math>.</p>

TABLE 1. RULE-OF-THUMB APPROXIMATIONS (*Continued*)

Parameter	Approximation	Remarks
Rise time	$t_r \omega_t \approx t_r \omega_m \approx 1.3$ where $t_r$ = rise time (10 to 90%) $\omega_t$ = (defined above) $\omega_m$ = (defined above)	The system's rise time, $t_r$ , which is here considered to be the time for the response to a step input to go from 10 to 90% of its final value may be approximated as indicated for systems with a $M_m$ value of about 1.3 to 1.5.
Phase margin at crossover frequency	$\gamma_c \geq 40^\circ$ where $\gamma_c$ = open loop phase margin at the crossover frequency	A phase margin of $40^\circ$ at the unity gain (crossover) frequency generally corresponds to a $M_m$ ratio of approximately 1.5. Since this value of $M_m$ is the maximum ordinarily considered feasible, the phase margin should be $40^\circ$ or greater.

## 6. APPROXIMATE RELATIONS—RULES OF THUMB

There are several approximations or rules of thumb which can be quite useful when time or facilities are not available for a more exact analysis. They may also be used as rough checks on the results of a more extensive analysis. The more common of these rules of thumb are presented in Table 1. They must be used with caution because, being approximations, they cannot apply with equal validity to all servo systems; and *the approximations for transient response are applicable only for step inputs.*

## 7. NUMERICAL AND GRAPHICAL TECHNIQUES OF RELATING TRANSIENT AND FREQUENCY RESPONSE

The numerical techniques presented involve only routine calculations and provide a point by point determination of the related response without the need of obtaining the closed loop poles or other intermediate quantities.

The methods presented require the following *assumptions*:

- (a) The system is linear.
- (b) The system frequency response approaches zero as the frequency approaches infinity.
- (c) The system's transient response begins with the system initially at rest.
- (d) The system is stable.

These requirements are satisfied by most servo systems. Even a non-linear system may generally be considered linear over a restricted operating range.

**Determining Transient Response from Frequency Response.** A relatively simple method for obtaining the time response to an impulse function input, knowing the frequency response, was developed by Floyd (Ref. 5). He derives the exact inverse transformation and then presents a method for numerically performing the necessary integration. The exact transformation is

$$(63) \quad c(t) = (2/\pi) \int_0^{\infty} \{\text{Re} [G(j\omega)] \cos t\omega\} d\omega,$$

where  $G(j\omega)$  is the closed loop frequency response of the system considered.

Floyd's procedure for evaluating this integral is to plot the real part of the closed loop frequency response, and then approximate the curve by a series of straight-line segments. This approximation is then treated as a summation of trapezoids. Equation (63) is applied to each trapezoid and the resulting time functions are added to obtain  $c(t)$ .

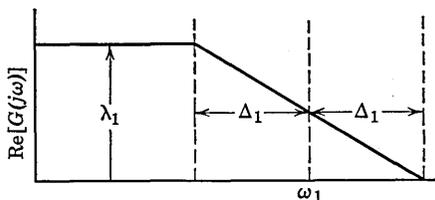


FIG. 32. Geometry of a trapezoid for approximating the real part of response function.

Each particular trapezoid is defined as indicated in Fig. 32. Performing the integration indicated in eq. (63) the value for the integral is

$$(64) \quad \frac{2A_1}{\pi} \left( \frac{\sin \omega_1 t}{\omega_1 t} \right) \left( \frac{\sin \Delta_1 t}{\Delta_1 t} \right),$$

where  $A_1 = \lambda_1 \omega_1$ , the area of the trapezoid, and  $\omega_1$  and  $\Delta_1$  are defined by the figure. The value of  $c(t)$  is then the summation for all the trapezoids.

$$(65) \quad c(t) = \sum_{n=1}^k \frac{2A_n}{\pi} \left( \frac{\sin \omega_n t}{\omega_n t} \right) \left( \frac{\sin \Delta_n t}{\Delta_n t} \right).$$

EXAMPLE. Assume the closed loop frequency response,  $G(j\omega)$ , of the system to be expressed mathematically as

$$(66) \quad G(j\omega) = \frac{18.72}{[(j\omega + 1)^2 + 1][(j\omega + 0.6)^2 + 9]}$$

From this the real part of  $G(j\omega)$  is calculated and plotted as shown in Fig. 33. The values used for  $\omega$ ,  $\Delta$ , and  $A$  of the series in eq. (65) are:

$$\begin{aligned} \omega_1 &= \frac{1.2 + 0.5}{2} & \omega_2 &= \frac{2.0 + 1.2}{2} & \omega_3 &= \frac{3.5 - 2.6}{2} \\ \Delta_1 &= \frac{1.2 - 0.5}{2} & \Delta_2 &= \frac{2.0 - 1.2}{2} & \Delta_3 &= \frac{3.5 - 2.6}{2} \\ A_1 &= 1 \times 0.85 & A_2 &= 0.66 \times 1.6 & A_3 &= 0.66 \times 3.05 \\ \omega_4 &= \frac{7.2 + 3.6}{2} & \omega_5 &= \frac{3.6 + 3.5}{2} \\ \Delta_4 &= \frac{7.2 - 3.6}{2} & \Delta_5 &= \frac{3.6 - 3.5}{2} \\ A_4 &= 0.07 \times 5.4 & A_5 &= 0.07 \times 3.55 \end{aligned}$$

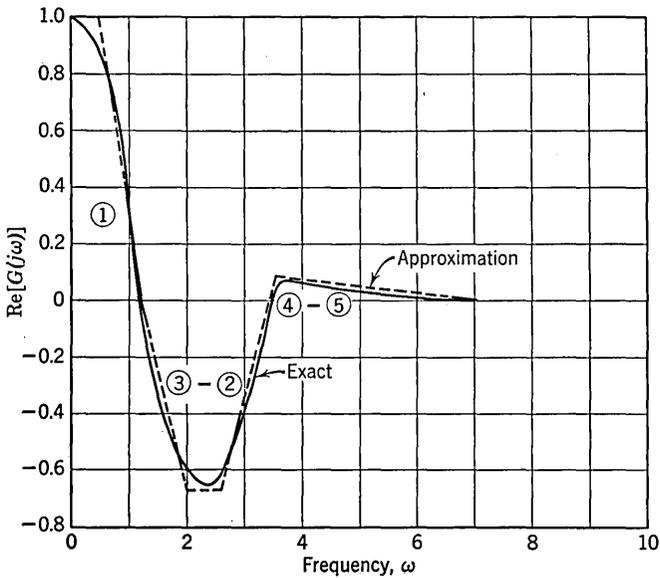


FIG. 33. Real part of response function and approximation.

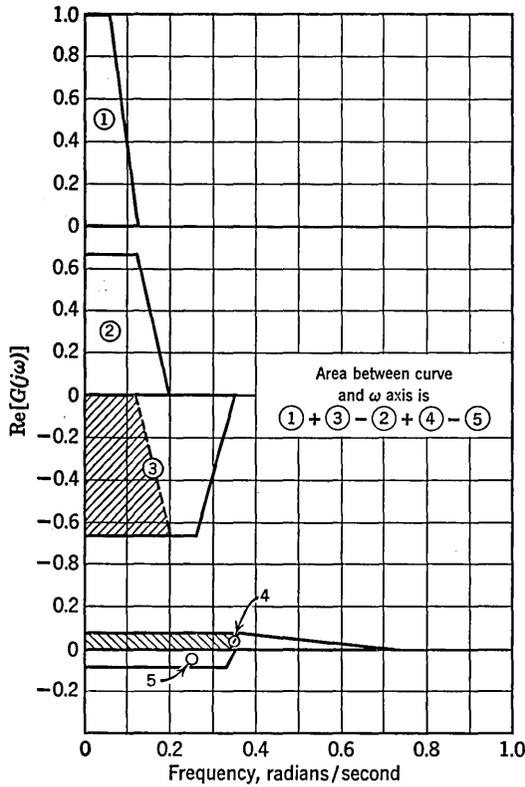


FIG. 34. Illustrating the trapezoids resulting from the straight-line approximation shown in Fig. 33.

Figure 34 illustrates the trapezoidal approximations used for Fig. 33 and the foregoing calculations. The evaluation of eq. (65) then becomes

$$\begin{aligned}
 (67) \quad c(t) = (2/\pi) & \left[ 0.85 \left( \frac{\sin 0.855t}{0.855t} \right) \left( \frac{\sin 0.35t}{0.35t} \right) \right. \\
 & + 1.07 \left( \frac{\sin 1.6t}{1.6t} \right) \left( \frac{\sin 0.4t}{0.4t} \right) - 2.01 \left( \frac{\sin 3.05t}{3.05t} \right) \left( \frac{\sin 0.45t}{0.45t} \right) \\
 & \left. + 0.38 \left( \frac{\sin 5.4t}{5.4t} \right) \left( \frac{\sin 1.8t}{1.8t} \right) - 0.25 \left( \frac{\sin 3.55t}{3.55t} \right) \left( \frac{\sin 0.05t}{0.05t} \right) \right].
 \end{aligned}$$

The  $\sin x/x$  tables (see Table 3) may be used to facilitate the evaluation of this equation at various values of  $t$ .

The exact solution, obtained by the inverse Laplace transformation, gives this result:

(68)

$$c(t) = 2.28 \exp(-t) \sin(t + 5.6^\circ) - 0.761 \exp(-0.6t) \sin(3t + 17^\circ).$$

For comparison both eqs. (67) and (68) are plotted in Fig. 35. This is the system time response to a unit impulse function. If the response to

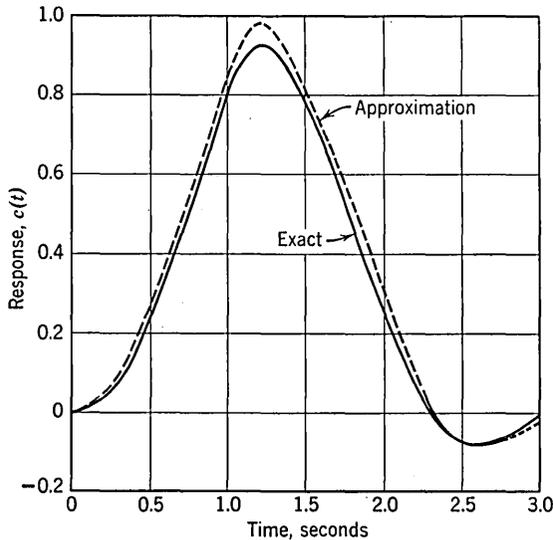


FIG. 35. Transient response for illustrative problem.

a step function is desired instead, the graphical integration of the curve for the impulse response provides it (Ref. 5).

### Determining Frequency Response from Transient Response

Quite often the frequency response characteristics of a component or system need to be known but it is difficult to introduce a sinusoidal signal or to measure magnitude and phase shift of the output. In many cases it is much simpler to introduce an impulse or step input; and since time and frequency responses are uniquely related, it is possible to obtain the frequency response from the transient response.

There are several approximate methods which have been developed for accomplishing this. Floyd's trapezoidal approximation method may be used but it yields only the real part of  $G(j\omega)$ . To obtain the total vector magnitude and phase shift a set of curves such as those presented by Bode (Ref. 6) must be used. Other methods have been developed by

Bedford and Fredendall, by Teasdale, Brooks and German, and by Samulon (Refs. 7, 8, and 9).

**Samulon's Method.** While the approaches vary somewhat the results are the same with the exception that Samulon's final equation has a "correction" term which makes it more accurate than the others. His procedure is presented here. Its basis is:

**SHANNON'S SAMPLING THEOREM.** *If a function  $c(t)$  contains no frequencies higher than  $f_{co}$  cycles per second it is completely determined by giving its ordinate at a series of points spaced  $1/(2f_{co})$  seconds apart.* Nearly any transient response curve will have some limiting value for its frequency spectrum, either due to the properties of the system or the test equipment itself. *Example.* The bandpass of the oscillograph might be the limiting item.

Shannon has also pointed out that such a function, with limited frequency components, can be exactly synthesized by a sum of  $\sin x/x$  func-

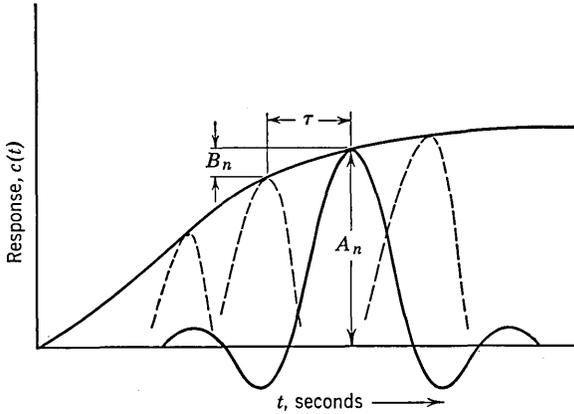


FIG. 36. Use of  $\sin x/x$  function to approximate transient response.

tions in a manner indicated in Fig. 36. The equation resulting from this approach is

$$(69) \quad G(j\omega) = \frac{\left(\frac{\pi}{2}\right) \left(\frac{\omega}{\omega_{co}}\right)}{\sin\left(\frac{\pi}{2}\right) \left(\frac{\omega}{\omega_{co}}\right)} \exp\left(j\omega \frac{\tau}{2}\right) \sum_{n=0}^{\infty} B_n \exp(-j\omega n\tau)$$

where  $B_n$  = the increment in the time response curve for a step function input,

$\omega$  = the frequency of interest, radians/second,

$\omega_{co}$  = the cutoff frequency for the system, radians/second,

$\tau$  = the sampling interval is equal to  $\pi/\omega_{co}$ .

Equation (69) would be exact if the system response contained no frequency components greater than  $\omega_{co}$ . This will never be absolutely true in a practical system but good results may be obtained nevertheless. In choosing the nominal cutoff frequency,  $\omega_{co}$ , the attempt should be made to estimate the frequency at which the steady-state frequency response is attenuated by at least 20 db. A good estimate of  $\omega_{co}$  is ten times  $\omega_c$ , the crossover frequency, as approximated in Table 1.

The calculated response will be in error at frequencies lower than the  $\omega_{co}$  selected if the true response contains higher frequencies. It is therefore desirable to have a frequency characteristic which attenuates rapidly above  $\omega_{co}$  selected for calculation. If the system or instrumentation does not provide this attenuation a filter may be added. Samulon states *"the amount of error, which will be largest near the nominal cutoff frequency,  $\omega_{co}$ , will be in general smaller than the amplitude response at  $\omega_{co}$ , provided that the response does not rise again above its value at  $\omega_{co}$  for frequencies greater than  $\omega_{co}$ ."* The calculated frequency response will indicate how valid the assumption of the cutoff frequency was. Note that with use of a lower  $\omega_{co}$  fewer points must be calculated.

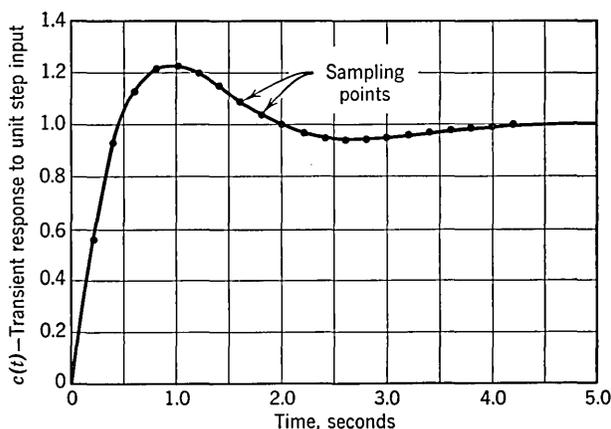


FIG. 37. Transient response for illustrative problem.

EXAMPLE. Assume a system with a time response to a step input as shown in Fig. 37. By assuming a system cutoff frequency of  $\omega_{co}$  of 15.7 rad per second,

$$(70) \quad f_{co} = 15.7/2\pi = 2.5 \text{ cps.}$$

By Shannon's theorem the sampling interval should be

$$(71) \quad \tau = 1/(2f_{co}) = 0.2 \text{ sec.}$$

By reading the ordinates from the curve at the sampling points Table 2 is constructed. For computational convenience the frequency response at  $\omega = \pi/1.6$  will be computed.

TABLE 2. FREQUENCY RESPONSE CALCULATED BY SAMULON'S METHOD FOR  $\omega = \pi/1.6$

$n\tau$	$c(n\tau)$	$B_n$	$B_n \exp(-j\omega n\tau)$			
			Real		Imaginary	
			+	-	+	-
0.2	0.57	0.57	0.527			0.218
0.4	0.93	0.36	0.254			0.254
0.6	1.13	0.20	0.076			0.185
0.8	1.22	0.09				0.090
1.0	1.23	0.01		0.004		0.009
1.2	1.20	-0.03	0.021		0.021	
1.4	1.15	-0.05	0.046		0.019	
1.6	1.09	-0.06	0.060			
1.8	1.04	-0.05	0.046			0.019
2.0	1.00	-0.04	0.028			0.028
2.2	0.97	-0.03	0.011			0.028
2.4	0.95	-0.02				0.020
2.6	0.94	-0.01		0.004		0.009
2.8	0.94					
3.0	0.95	0.01	0.009		0.004	
3.2	0.96	0.01	0.010			
3.4	0.97	0.01	0.009			0.004
3.6	0.98	0.01	0.007			0.007
3.8	0.99	0.01	0.004			0.009
4.0	1.00	0.01				0.010
4.2	1.00	0				
			1.108 - 0.008		0.044 - 0.890	
			+1.10		-j0.846	

$$(72) \quad \sum B_n \exp(-j\omega n\tau) = 1.10 - j0.846 = 1.385/\underline{-37.6^\circ}.$$

The vectors which are shown resolved and added numerically in Table 2 may be added graphically by simply plotting them end to end.

The correction terms in eq. (69) will now be computed.

Magnitude correction:

$$(73) \quad \frac{(\pi/2)(\pi/1.6)(1/5\pi)}{\sin(\pi/2)(\pi/1.6)(1/5\pi)} = \frac{\pi/16}{\sin(\pi/16)} = 1.007.$$

Phase correction:

$$(74) \quad \exp(j\omega\tau/2) = \exp j(\pi/1.6)(0.1) = 1/\underline{11.2^\circ}.$$

The final result is:

$$\begin{aligned}
 G(j\omega) &= \frac{(\pi/2)(\omega/\omega_{co})}{\sin(\pi/2)(\omega/\omega_{co})} \exp(j\omega\tau/2) \Sigma B_n \exp(-j\omega n\tau) \\
 (75) \quad &= (1.007)(1.0/11.2^\circ)(1.385/-37.6^\circ) \\
 &= 1.395/-26.4^\circ.
 \end{aligned}$$

The function chosen as an example is:

$$(76) \quad c(t) = 1 - \exp(-2t) + \exp(-t) \sin 1.5t,$$

$$(77) \quad C(s) = \frac{3.5s^2 + 7s + 6.5}{(s + 2)[(s + 1)^2 + 2.25]}$$

$$(78) \quad G(j\omega) = \frac{(1 - 0.538\omega)^2 + j(1.077\omega)}{(1 - 0.616\omega^2) + j(1.115\omega - 0.154\omega^3)}$$

The computed point is shown plotted on the exact  $G(j\omega)$  curve in Fig. 38.

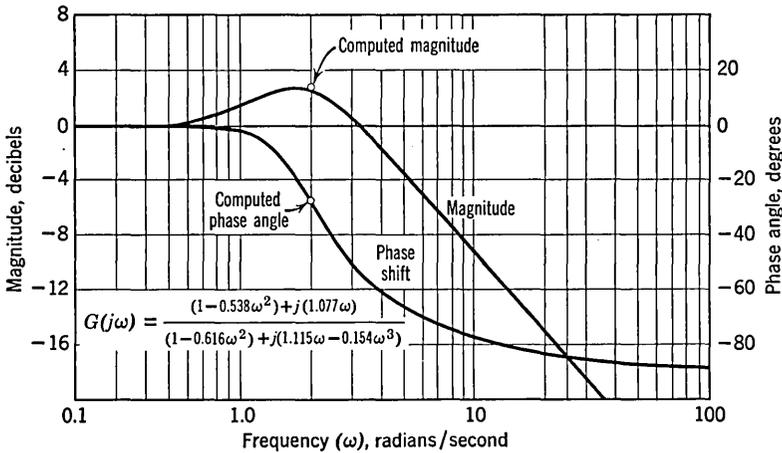


FIG. 38. Exact response curves and calculated points for example problem.

The  $\sin x/x$  values given in Table 3 may be used to aid in computing the magnitude correction. Generally, these correction terms will be negligible; however, if accuracy is important they should be checked. Samulon (Ref. 9) presents a series of tables and nomographs which are useful if extensive work of this kind is to be done.

The previous example illustrates the fact that the number of calculations required makes the whole problem rather tedious. If a great amount of such work is to be done, a special purpose analog computer may be used (Ref. 10).

TABLE 3. A FOUR-PLACE TABLE OF  $\sin x/x$  (Ref. 11)

$x$	0	1	2	3	4	5	6	7	8	9
0.0	+10000	10000	9999	9999	9997	9996	9994	9992	9989	9987
0.1	9983	9980	9976	9972	9967	9963	9957	9952	9946	9940
0.2	9933	9927	9919	9912	9904	9896	9889	9879	9870	9860
0.3	9851	9840	9830	9820	9808	9797	9785	9774	9761	9748
0.4	9735	9722	9709	9695	9680	9666	9651	9636	9620	9605
0.5	+9589	9572	9555	9538	9521	9503	9486	9467	9449	9430
0.6	9411	9391	9372	9351	9331	9311	9290	9269	9247	9225
0.7	9203	9181	9158	9135	9112	9089	9065	9041	9016	8992
0.8	8967	8942	8916	8891	8865	8839	8812	8785	8758	8731
0.9	8704	8676	8648	8620	8591	8562	8533	8504	8474	8445
1.0	+8415	8384	8354	8323	8292	8261	8230	8198	8166	8134
1.1	8102	8069	8037	8004	7970	7937	7903	7870	7836	7801
1.2	7767	7732	7698	7663	7627	7592	7556	7520	7484	7448
1.3	7412	7375	7339	7302	7265	7228	7190	7153	7115	7077
1.4	7039	7001	6962	6924	6885	6846	6807	6768	6729	6690
1.5	+6650	6610	6570	6530	6490	6450	6410	6369	6328	6288
1.6	6247	6206	6165	6124	6083	6042	6000	5959	5917	5875
1.7	5833	5791	5749	5707	5665	5623	5580	5538	5495	5453
1.8	5410	5368	5325	5282	5239	5196	5153	5110	5067	5024
1.9	4981	4937	4894	4851	4807	4764	4720	4677	4634	4590
2.0	+4546	4503	4459	4416	4372	4329	4285	4241	4198	4153
2.1	4111	4067	4023	3980	3936	3893	3849	3805	3762	3718
2.2	3675	3632	3588	3545	3501	3458	3415	3372	3328	3285
2.3	3242	3199	3156	3113	3070	3028	2984	2942	2899	2857
2.4	2814	2772	2730	2687	2645	2603	2561	2519	2477	2436
2.5	+2394	2352	2311	2269	2228	2187	2146	2105	2064	2023
2.6	1983	1942	1902	1861	1821	1781	1741	1702	1662	1622
2.7	1583	1544	1504	1465	1427	1388	1349	1311	1273	1234
2.8	1196	1159	1121	1083	1046	1009	972	935	898	861
2.9	825	789	753	717	681	646	610	575	540	505
3.0	+470	436	402	368	334	300	266	233	200	167
3.1	+134	+102	+69	+37	+5	-27	-58	-90	-121	-152
3.2	-182	213	243	273	303	333	362	392	421	449
3.3	478	506	535	562	590	618	645	672	699	725
3.4	752	778	804	829	855	880	905	930	954	978
$x$	0	1	2	3	4	5	6	7	8	9

TABLE 3. A FOUR-PLACE TABLE OF  $\sin x/x$  (Ref. 11) (Continued)

$x$	0	1	2	3	4	5	6	7	8	9
3.5	-1002	1026	1050	1073	1096	1119	1141	1164	1186	1208
3.6	1229	1251	1272	1293	1313	1334	1354	1374	1393	1413
3.7	1432	1451	1470	1488	1506	1524	1542	1559	1576	1593
3.8	1610	1627	1643	1659	1675	1690	1705	1720	1735	1749
3.9	1764	1777	1791	1805	1818	1831	1844	1856	1868	1880
4.0	-1892	1903	1915	1926	1936	1947	1957	1967	1977	1987
4.1	1996	2005	2014	2022	2030	2039	2046	2054	2061	2068
4.2	2075	2082	2088	2094	2100	2106	2111	2116	2121	2126
4.3	2131	2135	2139	2143	2146	2150	2153	2156	2158	2161
4.4	2163	2165	2166	2168	2169	2170	2171	2172	2172	2172
4.5	-2172	2172	2172	2171	2170	2169	2168	2166	2164	2162
4.6	2160	2158	2155	2152	2150	2146	2143	2139	2136	2132
4.7	2127	2123	2119	2114	2109	2104	2098	2093	2087	2081
4.8	2075	2069	2063	2056	2049	2042	2035	2028	2020	2013
4.9	2005	1997	1989	1981	1972	1963	1955	1946	1937	1927
5.0	-1918	1908	1899	1889	1879	1868	1858	1848	1837	1826
5.1	1815	1804	1793	1782	1770	1759	1747	1735	1723	1711
5.2	1699	1687	1674	1662	1649	1636	1623	1610	1597	1584
5.3	1570	1557	1543	1530	1516	1502	1488	1474	1460	1445
5.4	1431	1417	1402	1387	1373	1358	1343	1328	1313	1298
5.5	-1283	1268	1252	1237	1221	1206	1190	1175	1159	1143
5.6	1127	1111	1095	1079	1063	1047	1031	1015	999	982
5.7	966	950	933	917	900	884	867	851	834	818
5.8	800	784	768	751	734	718	701	684	667	650
5.9	634	617	600	583	567	550	533	516	499	482
6.0	-466	449	432	416	399	382	365	348	332	315
6.1	299	282	265	249	232	216	200	183	167	150
6.2	-134	-118	-102	-85	-69	-53	-37	-21	-5	+11
6.3	+27	43	58	74	90	105	121	136	152	167
6.4	182	197	212	227	242	257	272	287	302	316
6.5	+331	346	360	374	388	403	417	431	445	458
6.6	472	486	499	513	526	539	552	566	579	591
6.7	604	617	630	642	654	667	679	691	703	715
6.8	727	738	750	761	773	784	795	806	817	828
6.9	838	849	859	870	880	890	900	910	919	929
$x$	0	1	2	3	4	5	6	7	8	9

TABLE 3. A FOUR-PLACE TABLE OF  $\sin x/x$  (Ref. 11) (Continued)

$x$	0	1	2	3	4	5	6	7	8	9
7.0	+939	948	957	966	975	984	993	1002	1010	1019
7.1	1027	1035	1043	1051	1058	1066	1074	1081	1088	1095
7.2	1102	1109	1116	1123	1129	1135	1142	1148	1153	1159
7.3	1165	1171	1176	1181	1186	1191	1196	1201	1206	1210
7.4	1214	1219	1223	1227	1231	1234	1238	1241	1244	1248
7.5	+1251	1254	1256	1259	1261	1264	1266	1268	1270	1272
7.6	1274	1275	1277	1278	1279	1280	1281	1282	1282	1283
7.7	1283	1284	1284	1284	1284	1283	1283	1282	1282	1281
7.8	1280	1279	1278	1277	1275	1274	1272	1270	1269	1267
7.9	1264	1262	1259	1257	1255	1252	1249	1246	1243	1240
8.0	+1237	1233	1230	1226	1222	1218	1214	1210	1206	1202
8.1	1197	1193	1188	1183	1179	1174	1169	1163	1158	1153
8.2	1147	1142	1136	1130	1124	1118	1112	1106	1100	1093
8.3	1087	1080	1074	1067	1060	1053	1046	1039	1032	1025
8.4	1017	1010	1002	995	987	979	972	964	956	948
8.5	+939	931	923	915	906	898	889	880	872	863
8.6	854	845	836	827	818	809	800	790	781	771
8.7	762	752	743	733	724	714	704	694	684	675
8.8	665	655	645	635	625	614	604	594	584	573
8.9	563	552	542	532	521	511	500	490	479	469
9.0	+458	447	437	426	415	404	394	383	372	361
9.1	351	340	329	318	307	296	286	275	264	253
9.2	242	231	220	210	199	188	177	166	156	145
9.3	134	123	112	101	91	80	69	58	48	37
9.4	+26	+16	+5	-6	-16	-27	-37	-48	-58	-69
9.5	-79	89	100	110	120	131	141	151	161	172
9.6	182	192	202	212	222	231	241	251	261	271
9.7	280	290	299	309	318	328	337	346	356	365
9.8	374	383	392	401	410	419	428	436	445	454
9.9	462	471	479	487	496	504	512	520	528	536
10.0	-544	552	560	567	575	582	590	597	604	612
10.1	619	626	633	640	647	653	660	667	673	680
10.2	686	692	699	705	711	717	723	728	734	740
10.3	745	751	756	761	767	772	777	782	787	791
10.4	796	801	805	809	814	818	822	826	830	834
$x$	0	1	2	3	4	5	6	7	8	9

TABLE 3. A FOUR-PLACE TABLE OF  $\sin x/x$  (Ref. 11) (Continued)

$x$	0	1	2	3	4	5	6	7	8	9
10.5	-838	842	845	849	852	855	859	862	865	868
10.6	871	873	876	879	881	883	886	888	890	892
10.7	894	896	898	899	901	902	904	905	906	907
10.8	908	909	910	911	911	912	912	913	913	913
10.9	913	913	913	913	913	912	912	911	911	910
11.0	-909	908	907	906	905	904	902	901	899	898
11.1	896	894	892	890	888	886	884	882	879	877
11.2	874	872	869	866	863	860	857	854	851	848
11.3	844	841	837	834	830	826	822	819	815	811
11.4	806	802	798	794	789	785	780	776	771	766
11.5	-761	756	751	746	741	736	731	726	720	715
11.6	709	704	698	693	687	681	675	669	663	657
11.7	651	645	639	633	626	620	614	607	601	594
11.8	588	581	574	568	561	554	547	540	533	526
11.9	519	512	505	498	491	484	476	469	462	454
12.0	-447	440	432	425	417	410	402	395	387	379
12.1	372	364	356	348	341	333	325	317	309	301
12.2	294	286	278	270	262	254	246	238	230	222
12.3	214	206	198	190	182	174	166	158	150	142
12.4	134	125	117	109	101	93	85	77	69	61
12.5	-53	-45	-37	-29	-21	-13	-5	+3	+11	+19
12.6	+27	35	42	50	58	66	74	82	89	97
12.7	105	113	120	128	136	143	151	158	166	173
12.8	181	188	196	203	210	218	225	232	240	247
12.9	254	261	268	275	282	289	296	303	310	316
13.0	+323	330	337	343	350	356	363	369	376	382
13.1	388	395	401	407	413	419	425	431	437	443
13.2	448	454	460	466	471	477	482	488	493	498
13.3	503	509	514	519	524	529	534	538	543	548
13.4	552	557	562	566	570	575	579	583	587	591
13.5	+595	599	603	607	611	614	618	622	625	628
13.6	632	635	638	641	644	647	650	653	656	659
13.7	661	664	666	669	671	673	676	678	680	682
13.8	684	686	688	689	691	692	694	695	697	698
13.9	699	700	702	703	703	704	705	706	706	707
$x$	0	1	2	3	4	5	6	7	8	9

TABLE 3. A FOUR-PLACE TABLE OF  $\sin x/x$  (Ref. 11) (Continued)

$x$	0	1	2	3	4	5	6	7	8	9
14.0	+708	708	708	709	709	709	709	709	709	709
14.1	709	708	708	708	707	707	706	705	705	704
14.2	703	702	701	700	699	697	696	695	693	692
14.3	690	688	687	685	683	681	679	677	675	673
14.4	671	668	666	663	661	658	656	653	650	648
14.5	+645	642	639	636	633	630	626	623	620	616
14.6	613	609	606	602	599	595	591	587	583	579
14.7	575	571	567	563	559	555	550	546	542	537
14.8	533	528	524	519	514	509	505	500	495	490
14.9	485	480	475	470	465	460	455	449	444	439
15.0	+434	428	423	417	412	406	401	395	390	384
15.1	378	373	367	361	355	349	344	338	332	326
15.2	320	314	308	302	296	290	284	278	272	265
15.3	259	253	247	241	234	228	222	216	209	203
15.4	197	190	184	178	171	165	159	152	146	140
15.5	+133	127	120	114	108	101	95	88	82	76
15.6	69	63	56	50	43	37	31	24	18	11
15.7	+5	-1	-8	-14	-20	-27	-33	-39	-46	-52
15.8	58	64	71	77	83	89	95	102	108	114
15.9	120	126	132	138	144	150	156	162	168	174
16.0	-180	186	192	197	203	209	215	220	226	232
16.1	237	243	248	254	259	265	270	276	281	286
16.2	292	297	302	307	312	318	323	328	333	337
16.3	342	347	352	357	362	366	371	376	380	385
16.4	389	393	398	402	407	411	415	419	423	427
16.5	-431	435	439	443	447	451	454	458	462	465
16.6	469	472	476	479	482	486	489	492	495	498
16.7	501	504	507	510	513	515	518	521	523	526
16.8	528	531	533	535	538	540	542	544	546	548
16.9	550	552	553	555	557	558	560	561	563	564
17.0	-566	567	568	569	570	571	572	573	574	575
17.1	575	576	577	577	578	578	579	579	579	579
17.2	580	580	580	580	580	579	579	579	579	578
17.3	578	577	577	576	576	575	574	573	572	571
17.4	570	569	568	567	566	565	563	562	561	559
$x$	0	1	2	3	4	5	6	7	8	9

TABLE 3. A FOUR-PLACE TABLE OF  $\sin x/x$  (Ref. 11) (Continued)

$x$	0	1	2	3	4	5	6	7	8	9
17.5	-557	556	554	553	551	549	547	545	543	541
17.6	539	537	535	533	530	528	526	523	521	518
17.7	516	513	510	508	505	502	499	496	493	490
17.8	487	484	481	478	475	471	468	465	461	458
17.9	454	451	447	444	440	436	433	429	425	421
18.0	-417	413	409	405	401	397	393	389	385	381
18.1	376	372	368	364	359	355	350	346	341	337
18.2	332	328	323	319	314	309	304	300	295	290
18.3	285	281	276	271	266	261	256	251	246	241
18.4	236	231	226	221	216	211	206	201	195	190
18.5	-185	180	175	170	164	159	154	149	143	138
18.6	133	128	122	117	112	106	101	96	90	85
18.7	80	74	69	64	58	53	48	42	37	32
18.8	-26	-21	-16	-10	-5	+0	+6	+11	+16	+21
18.9	+27	32	37	42	48	53	58	63	68	74
19.0	+79	84	89	94	99	104	110	115	120	125
19.1	130	135	140	145	150	155	159	164	169	174
19.2	179	184	188	193	198	202	207	212	216	221
19.3	226	230	235	239	244	248	252	257	261	265
19.4	270	274	278	282	286	290	295	299	303	307
19.5	+311	314	318	322	326	330	333	337	341	344
19.6	348	351	355	358	362	365	369	372	375	378
19.7	382	385	388	391	394	397	400	403	405	408
19.8	411	414	416	419	422	424	427	429	431	434
19.9	436	438	440	443	445	447	449	451	453	455
$x$	0	1	2	3	4	5	6	7	8	9

TABLE 3. A FOUR-PLACE TABLE OF  $\sin x/x$  (Ref. 11) (Continued)

$x$	0	2	4	6	8	$x$	0	2	4	6	8
20.0	+456	460	463	466	469	23.5	-425	425	425	424	424
20.1	472	475	477	479	481	23.6	423	423	422	421	419
20.2	483	485	486	487	488	23.7	418	416	415	413	411
20.6	489	490	490	490	490	23.8	408	406	403	401	398
20.4	490	490	489	488	487	23.9	395	392	388	385	381
20.5	+486	485	483	482	480	24.0	-377	373	369	365	361
20.6	478	475	473	470	467	24.1	356	352	347	342	337
20.7	464	461	458	454	450	24.2	332	327	321	316	310
20.8	447	442	438	434	429	24.3	304	299	293	287	280
20.9	424	420	415	409	404	24.4	274	268	261	255	248
21.0	+398	393	387	381	375	24.5	-241	235	228	221	214
21.1	369	362	356	349	342	24.6	206	199	192	185	177
21.2	335	328	321	314	307	24.7	170	162	155	147	139
21.3	299	292	284	276	268	24.8	132	124	116	108	100
21.4	260	252	244	236	228	24.9	93	85	77	69	61
21.5	+219	211	202	194	185	25.0	-53	45	37	29	21
21.6	176	168	159	150	141	25.1	-13	-5	+3	+11	+19
21.7	132	123	114	105	96	25.2	+27	35	42	50	58
21.8	87	78	69	60	51	25.3	66	74	81	89	96
21.9	42	32	23	14	5	25.4	104	111	119	126	134
22.0	-4	13	22	31	40	25.5	+141	148	155	162	169
22.1	49	58	67	76	85	25.6	176	183	189	196	203
22.2	93	102	111	119	128	25.7	209	215	222	228	234
22.3	136	145	153	161	169	25.8	240	246	251	257	263
22.4	178	185	193	201	209	25.9	268	273	278	284	288
22.5	-217	224	231	239	246	26.0	+293	298	303	307	311
22.6	253	260	267	274	280	26.1	315	320	323	327	331
22.7	287	293	299	305	311	26.2	334	338	341	344	347
22.8	317	323	329	334	339	26.3	350	352	355	357	359
22.9	344	349	354	359	364	26.4	361	363	365	367	368
23.0	-368	372	376	380	384	26.5	+370	371	372	373	373
23.1	388	391	394	397	400	26.6	374	374	375	375	375
23.2	403	406	408	410	413	26.7	375	374	374	373	372
23.3	415	416	418	419	421	26.8	371	370	369	368	366
23.4	422	423	423	424	424	26.9	365	363	361	359	357
$x$	0	2	4	6	8	$x$	0	2	4	6	8

TABLE 3. A FOUR-PLACE TABLE OF  $\sin x/x$  (Ref. 11) (Continued)

$x$	0	2	4	6	8	$x$	0	2	4	6	8
27.0	+354	352	349	346	343	30.5	-260	256	252	247	243
27.1	340	337	334	331	327	30.6	238	233	229	224	219
27.2	323	319	316	312	307	30.7	214	209	204	198	193
27.3	303	299	294	290	285	30.8	188	182	177	171	165
27.4	280	275	270	265	260	30.9	160	154	148	142	136
27.5	+254	249	243	238	232	31.0	-130	124	118	112	106
27.6	226	220	214	208	202	31.1	100	94	87	81	75
27.7	196	190	184	177	171	31.2	69	62	56	50	43
27.8	164	158	151	145	138	31.3	37	31	24	18	11
27.9	131	124	117	111	104	31.4	-5	+1	+8	+14	+20
28.0	+97	90	83	76	69	31.5	+27	33	39	45	52
28.1	62	55	48	41	34	31.6	58	64	70	76	82
28.2	+26	+19	+12	+5	-2	31.7	88	94	100	106	112
28.3	-9	16	23	30	37	31.8	118	124	129	135	140
28.4	44	51	58	65	72	31.9	146	151	157	162	167
28.5	-79	85	92	99	105	32.0	+172	177	182	187	192
28.6	112	118	125	131	138	32.1	197	202	206	211	215
28.7	144	150	156	162	168	32.2	219	224	228	232	236
28.8	174	180	186	192	197	32.3	239	243	247	250	254
28.9	203	208	213	219	224	32.4	257	260	263	266	269
29.0	-229	234	239	243	248	32.5	+272	275	277	280	282
29.1	253	257	261	266	270	32.6	284	286	288	290	292
29.2	274	278	281	285	288	32.7	293	295	296	297	299
29.3	292	295	298	301	304	32.8	300	300	301	302	302
29.4	307	310	312	315	317	32.9	303	303	303	303	303
29.5	-319	321	323	325	326	33.0	+303	303	302	302	301
29.6	328	329	330	331	332	33.1	300	299	298	297	296
29.7	333	334	334	335	335	33.2	294	293	291	290	288
29.8	335	335	335	335	334	33.3	286	284	281	279	277
29.9	334	333	332	332	331	33.4	274	272	269	266	263
30.0	-329	328	327	325	323	33.5	+260	257	254	250	247
30.1	321	320	317	315	313	33.6	243	240	236	232	228
30.2	311	308	305	302	300	33.7	224	220	216	212	208
30.3	296	293	290	287	283	33.8	203	199	194	190	185
30.4	280	276	272	268	264	33.9	180	175	171	166	161
$x$	0	2	4	6	8	$x$	0	2	4	6	8

TABLE 3. A FOUR-PLACE TABLE OF  $\sin x/x$  (Ref. 11) (Continued)

$x$	0	2	4	6	8	$x$	0	2	4	6	8
34.0	+156	151	145	140	135	37.0	-174	170	165	161	156
34.1	130	124	119	113	108	37.1	152	147	143	138	133
34.2	102	97	91	86	80	37.2	129	124	119	114	109
34.3	74	69	63	57	51	37.3	104	99	94	89	84
34.4	46	40	34	28	22	37.4	79	74	68	63	58
34.5	+17	+11	+5	-1	-7	37.5	-53	47	42	37	32
34.6	-12	18	24	30	35	37.6	26	21	16	10	5
34.7	41	47	52	58	63	37.7	+0	6	11	16	21
34.8	69	75	80	85	91	37.8	27	32	37	42	47
34.9	96	102	107	112	117	37.9	53	58	63	68	73
35.0	-122	127	132	137	142	38.0	+78	83	88	93	98
35.1	147	152	157	161	166	38.1	102	107	112	117	121
35.2	170	175	179	183	187	38.2	126	130	135	139	143
35.3	192	196	199	203	207	38.3	148	152	156	160	164
35.4	211	214	218	221	225	38.4	168	172	176	179	183
35.5	-228	231	234	237	240	38.5	+186	190	193	197	200
35.6	243	245	248	250	253	38.6	203	206	209	212	215
35.7	255	257	259	261	263	38.7	218	220	223	225	228
35.8	264	266	268	269	270	38.8	230	232	234	236	238
35.9	271	272	273	274	275	38.9	240	241	243	244	246
36.0	-275	276	276	277	277	39.0	+247	248	249	250	251
36.1	277	277	277	276	276	39.1	252	253	253	254	254
36.2	276	275	274	273	272	39.2	254	255	255	255	255
36.3	271	270	269	268	266	39.3	254	254	254	253	252
36.4	265	263	261	259	257	39.4	252	251	250	249	248
36.5	-255	253	251	248	246	39.5	+246	245	244	242	241
36.6	243	241	238	235	232	39.6	239	237	235	233	231
36.7	229	226	223	220	216	39.7	229	227	224	222	219
36.8	213	209	206	202	198	39.8	217	214	211	208	206
36.9	194	190	186	182	178	39.9	203	199	196	193	190
$x$	0	2	4	6	8	$x$	0	2	4	6	8

## ACKNOWLEDGMENT

Figures 13 to 30 are reproduced with permission from H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. I, Wiley, New York, 1951.

The example in the section on Determining Transient Response from Frequency Response is reprinted with permission from G. S. Brown and D. P. Campbell, *Principles of Servomechanisms*, Wiley, New York, 1948.

---

**REFERENCES**

1. W. R. Evans, *Control System Dynamics*, McGraw-Hill, New York, 1954.
2. G. A. Biernson, Quick methods for evaluating the closed-loop poles of feedback control systems, *Trans. Am. Inst. Elec. Engrs.*, **72**, Pt. 2, 53-70 (1953).
3. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. I, Wiley, New York, 1951.
4. G. A. Biernson, Estimating transient response from open-loop frequency response, *Trans. Am. Inst. Elec. Engrs.*, **74**, 388-403 (1956).
5. G. S. Brown and D. P. Campbell, *Principles of Servomechanisms*, Wiley, New York, 1948.
6. H. W. Bode, *Network Analysis and Feedback Amplifier Design*, Van Nostrand, Princeton, N. J., 1945.
7. A. V. Bedford and G. L. Fredendall, Analysis, synthesis and evaluation of the transient response of television apparatus, *Proc. I.R.E.*, **30**, 440-458 (1942).
8. A. R. Teasdale, Jr., F. E. Brooks, Jr., and J. P. German, System Frequency Response Derived from Transient Response, Am. Inst. Elec. Engrs. District Paper, New York, October 1950.  
A. R. Teasdale, Jr., Get frequency response from transient data by adding vectors, *Control Eng.*, **2**, 56-59 (1955).
9. H. A. Samulon, Spectrum analysis of transient response curves, *Proc. I.R.E.*, **39**, 175-186 (1951).
10. J. B. Reynolds, Jr., Get frequency response from transient data by machine computing, *Control Eng.*, **2**, 60-63 (1955).
11. J. Sherman, *Z. Krist.*, **85**, 404 (1933).



## Feedback System Compensation

P. G. Cushman

1. Design Criteria and Techniques	23-01
2. Compensating Components: D-C Systems	23-18
3. Compensating Networks: A-C Systems	23-48
4. Open-Closed Loop Control	23-54
References	23-56

### 1. DESIGN CRITERIA AND TECHNIQUES

The first step in the design of a feedback control system is the selection of a suitable power element with sufficient torque, or force, speed, and power rating to drive the load. Once the selection of a power element with known characteristics has been made, the signal devices, amplifiers, and stabilizing components have to be chosen with such characteristics that make the entire feedback control system meet system requirements of accuracy, speed of response, and stability. This chapter is devoted to the synthesis of required characteristics of these compensating components and the presentation of characteristics of practical control system components. Section 1 derives feedback control system characteristics from system specifications.

#### Synthesis of Log Magnitude Diagram from System Requirements

**Low-Frequency Portion: Static Error Coefficients.** *Error coefficients* are one of the most common means of specifying control system

performance. These coefficients are *figures of merit*, the higher the coefficient, the smaller the control system error in achieving a required output.

The static error coefficients are defined as the ratio of the constant output required (position, velocity, or acceleration) to the control system error required to achieve that output. The types of control system and the static error coefficient associated with each type are summarized in Chap. 20. (See also Ref. 1, Chap. 8.)

The static error coefficients influence the *log magnitude diagram* in an easily visualized way and lead to a method of control system classifica-

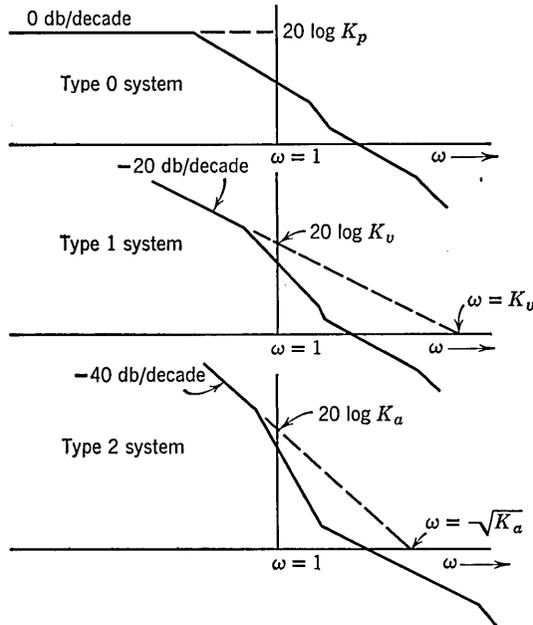


FIG. 1. Sample log magnitude diagrams showing influence of static error coefficients.

tion. For example, a control system with a transfer function that approaches  $K_p$ , a constant, at low frequencies (open loop transfer function,  $G(s)$  approaches  $K_p$  as  $s$  approaches 0) will have a log magnitude diagram which has zero slope at low frequencies. Such a system is called a *type 0 system* (0 slope at low frequencies) and can follow a steady input,  $r_0$ , with an error of  $r_0/(1 + K_p)$ . If  $K_p$  is large, the error will be small. However, if a velocity signal,  $r = v_0 t$ , is applied, the error will continue to increase with time. For a system to follow such a signal with small error, a *type 1 system* is required which has a transfer function at low frequency of

$K_v/j\omega$  (limit  $sG(s)_{s=0} = K_v$ ) and an initial slope on the log magnitude diagram of  $-20$  db per decade. A type 1 system would follow the constant velocity input with an error of only  $v_0/K_v$ . Similarly, a *type 2 system* ( $K_a/(j\omega)^2$  transfer function giving a slope of  $2(-20) = -40$  db per decade at low frequencies) is required to follow a constant acceleration input with moderate error.

The type of system determines the shape of the log magnitude diagram at low frequencies and the gain magnitude of this portion of the diagram is determined by the static error coefficients. The intersection of the extensions of the initial log magnitude diagram slope with the  $\omega = 1$  line is at the value  $20 \log K_p$ ,  $20 \log K_v$ , or  $20 \log K_a$ , as the case may be. The intersection of the extensions of the initial slope with the 0-db axis also has significance as shown in Fig. 1.

**Low-Frequency Portion: Dynamic Error Coefficients.** In addition to the steady-state characteristics, expressible in terms of the static error coefficients, it is often desirable to specify control system errors during a transient by means of *dynamic error coefficients*, defined in Chap. 20. That is,

$$e = \frac{1}{K_0} r + \frac{1}{K_1} \dot{r} + \frac{1}{K_2} \ddot{r} + \frac{1}{K_3} \dddot{r} + \dots,$$

where  $r$ ,  $\dot{r}$ ,  $\ddot{r}$  are successive derivatives of the input time function and  $K_0$ ,  $K_1$ ,  $K_2$ , etc., are the dynamic error coefficients. The above relation is valid during time intervals in a transient which are far displaced in time from a discontinuity in the input function,  $r$ , and its derivatives. The above equation converges quickly to useful values for slowly changing input functions for which the higher order derivatives are small relative to the lower order terms. The coefficients can be evaluated by straightforward Laplace transform techniques, as given in Chap. 20. That is,

$$\frac{1}{K_n} = \frac{1}{n!} \lim_{s \rightarrow 0} \frac{d^n}{ds^n} \left[ \frac{E}{R}(s) \right].$$

Some of the error coefficients, evaluated in this way, will be found identical to the static coefficients of the previous paragraph. However, additional coefficients will also be determined. The composition of these generalized error coefficients can be seen from a general control system transfer function (see Ref. 4).

$$\frac{E(s)}{R(s)} = \frac{n_0 + n_1 s + n_2 s^2 + n_3 s^3 + \dots}{1 + d_1 s + d_2 s^2 + d_3 s^3 + \dots}.$$

The dynamic error coefficients for this system are:

$$\frac{1}{K_0} = n_0,$$

$$\frac{1}{K_1} = n_1 - \frac{1}{K_0} d_1,$$

$$\frac{1}{K_2} = n_2 - \frac{1}{K_1} d_1 - \frac{1}{K_0} d_2,$$

$$\frac{1}{K_k} = n_k - \sum_{j=0}^{j=k-1} \frac{1}{K_j} d_{(k-j)}.$$

The dynamic error coefficients in general are composed of the gain term in combination with various sums and products of the system time constants. These coefficients are readily calculable and are valuable for analysis purposes for a system of known transfer function. However, they are not very useful in synthesizing the log magnitude diagram from system requirements, because each of the coefficients is composed of a number of parameters of the system characteristics. For this synthesis work, a more direct procedure is outlined in the next paragraph.

**Low-Frequency Portion: Transient Curve Fitting Procedure.** A curve fitting procedure (Ref. 2) by which certain system error requirements are transformed directly to log magnitude values which the log magnitude diagram must exceed is useful. In this method, the expected transient input signals are matched by sinusoids. The principle is that if a control system can follow with small error sine wave inputs with amplitude, velocity, and acceleration components as great as those of the transient input, then it can follow the transient with small error. The worst transients that the control system will be expected to follow are presumably known, either in graphical or analytical form, together with the allowable errors during these transients. These transient time functions are plotted and fitted as closely as possible in various places with sine waves as indicated in Fig. 2. The amplitudes of these sine waves are  $A_1$ ,  $A_2$ ,  $A_3$ , etc., with frequencies  $\omega_1$ ,  $\omega_2$ ,  $\omega_3$ .  $A_1/E$ ,  $A_2/E$ ,  $A_3/E$ , etc., are the required gain magnitudes of the log magnitude diagram at  $\omega_1$ ,  $\omega_2$ ,  $\omega_3$  if  $E$  is the allowable control system error. These points are shown in Fig. 3. The required log magnitude diagram must be above these points.

It is important to fit the input transient at several places, such as peaks and maximum slope points, so that broad coverage of requirements is established by several points on the log magnitude diagram. Sometimes it is advantageous to take the derivative of the input transient and fit it

with sine waves of amplitude  $V_1, V_2$ , etc., at frequencies  $\omega_1, \omega_2$ . These fits will establish points on the log magnitude diagram of magnitude  $V_1/\omega_1 E, V_2/\omega_2 E$ , etc. The procedure can be extended to higher derivatives also.

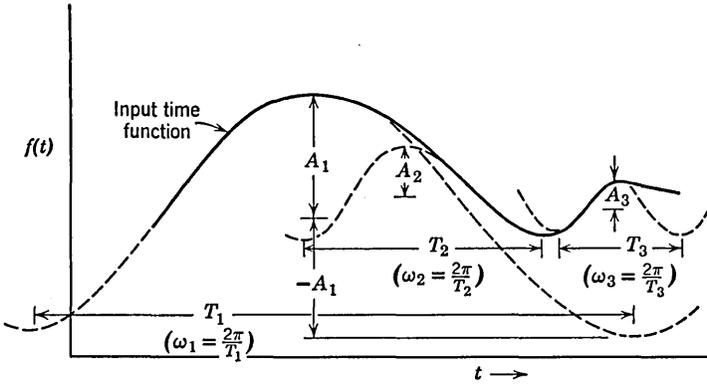


FIG. 2. Construction illustrating curve fitting procedure.

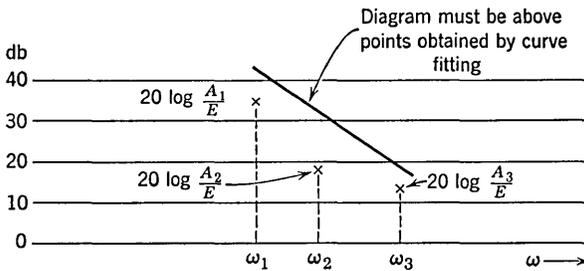


FIG. 3. Log magnitude points obtained from Fig. 2.

**Mid-Frequency Portion of the Log Magnitude Diagram.** Once the gain of the entire control system has been set so as to meet the requirements as outlined in the preceding paragraphs, it is usually desirable to reduce the system gain effective at the higher frequencies in order to reduce the susceptibility of the system to noise and other extraneous signals. However, this reduction in gain has to be achieved in such a manner that the system has the required stability. As explained in Chap. 21, stability may be assured by requiring the log magnitude diagram to have a slope of  $-20$  db per decade in the vicinity of the crossover frequency. To obtain adequate stability, this  $-20$  db per decade slope should extend for a frequency range of a decade or more. The use of the *log magnitude-angle chart* (Nichols chart) provides a measure of stability in terms of the maximum  $M$  of the closed loop frequency response. Such charts are given

in Chap. 21. To indicate approximate magnitudes, Fig. 4 shows the maximum  $M$  that could possibly be obtained for a particular minimum value of phase margin.

Often it is convenient to express the degree of stability, or damping, of a system by means of a *damping factor*. Strictly, a damping factor can be applied only to a system that can be described by a second order linear differential equation with constant coefficients, but it is frequently applied

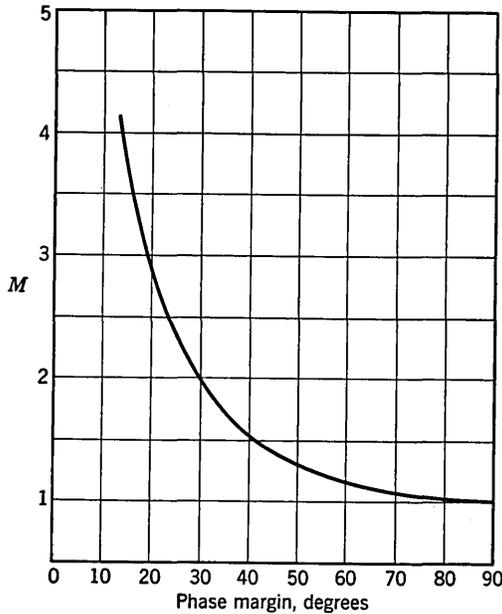


FIG. 4. Maximum peak of frequency response versus minimum phase margin.

to higher order systems. When the response is determined largely by two complex roots, which is fairly common, the closed loop response is characterized by a zero slope region of approximately unity gain at low frequencies followed by a resonant peak in the vicinity of crossover of the open loop. At frequencies above the resonant peak, the slope changes to  $-40$  db per decade and then usually to even greater negative slopes. Thus for the frequency region from zero to somewhat above the resonant peak, many systems have much the same frequency characteristic as a second order system. For such a system the height of the resonant peak, when expressed as a numeric ratio,  $M_m$ , determines the damping factor,  $\zeta$ , by the equation:

$$M_m = \frac{1}{2\zeta\sqrt{1-\zeta^2}}, \quad \text{valid: } 0 < \zeta < 0.707.$$

Frequently it is convenient to measure the magnitude of the frequency response,  $M_c$ , at the corner frequency. For such a measurement, the damping factor is

$$\zeta = \frac{1}{2M_c}$$

The damping of oscillations in a physical system is a function of the damping factor. When a system is excited by a unit step function, the magnitude of the first and successive overshoots is determined by the damping factor as shown in Fig. 5. See also Chap. 20.

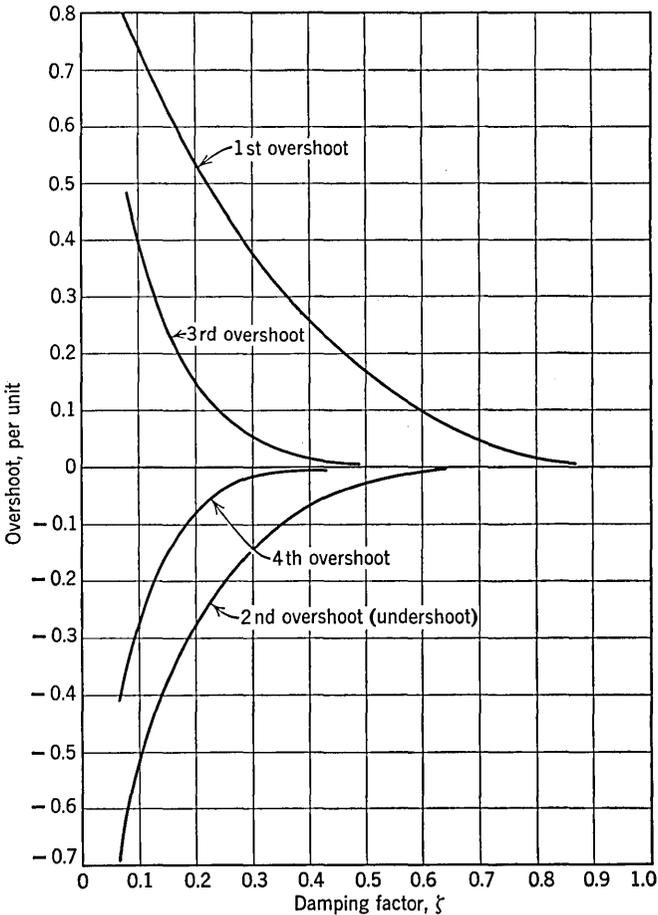


Fig. 5. Overshoot magnitude versus damping factor for unit step function applied to second order system.

**High-Frequency Portion.** At frequencies higher than the crossover frequency, the log magnitude diagram should be attenuated as rapidly as possible, consistent with stability and damping requirements, in order to reduce noise. In systems in which there is a relatively large amount of noise signal present, the noise error,  $E_n$ , in the control system output may be larger than the system errors,  $E_s$ , that were considered in the preceding paragraphs, even when the high-frequency portion of the log magnitude diagram is attenuated. If the noise has significant high-frequency components, it will be profitable to lower  $\omega_c$ , the crossover frequency, and the gain at frequencies below  $\omega_c$  in order to reduce  $E_n$ . This procedure increases  $E_s$ , but the overall root mean square error may be reduced. The usual criterion in such cases is to minimize

$$E_{\text{rms}} = (E_n^2 + E_s^2)^{1/2}.$$

See Chap. 24 for a complete treatment of noise evaluation and system optimization.

Figure 6 shows a sample log magnitude diagram synthesized by the methods described above.

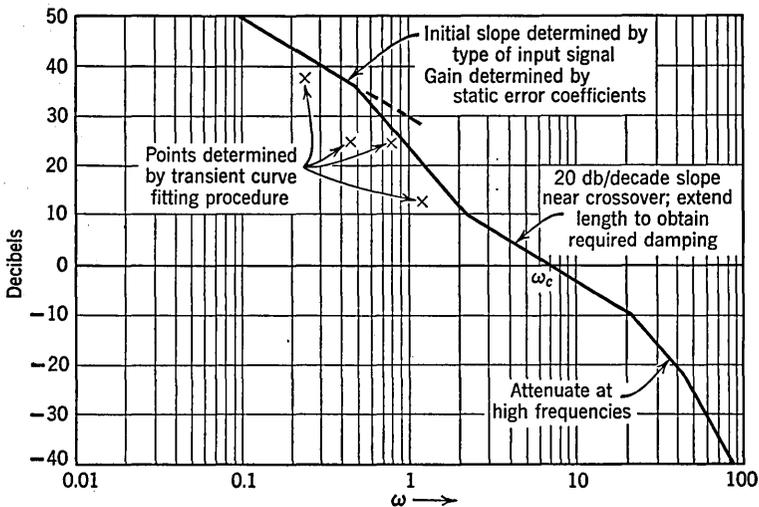


FIG. 6. Sample log magnitude diagram summarizing synthesis procedure.

### Performance Charts Relating Transient and Frequency Response.

The methods of the previous paragraphs will be found straightforward in synthesizing the low- and medium-frequency portions of the log magnitude diagram from system requirements. The degree of stability and the nature of the transient response often can be more easily obtained by means

of the design charts relating frequency response and transient response as given in Chap. 22.

**Load Disturbances.** The previous discussion has been concerned with obtaining the proper overall open loop gain and frequency characteristics. In many practical systems, not only the overall characteristics but also the characteristics and order of the individual elements of the system become important. The latter situation occurs whenever there are additional or *extraneous inputs* acting on the system in addition to the main controlling signal input. *Examples of the extraneous inputs* are (a) torque changes in the load on the power element of a speed control system in which a voltage, proportional to derived speed, is the main input signal and (b) electrical pickup of alternating voltage signals from power supplies in a d-c amplifier channel of a feedback control system. In all cases of this sort it is desirable that the control system be insensitive to these extraneous inputs yet follow the desired input signals properly.

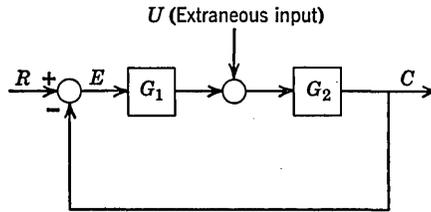


Fig. 7. Block diagram of general feedback control system showing extraneous input.

Figure 7 shows a block diagram of a feedback control system with an extraneous input. The normal closed loop transfer function to the input signal is

$$\frac{C}{R}(s) = \frac{G_1(s)G_2(s)}{1 + G_1(s)G_2(s)},$$

and the gain and frequency characteristic of  $G_1(s) G_2(s)$  could be established as indicated in the previous paragraphs. The transfer function on the extraneous signal is

$$\frac{C}{U}(s) = \frac{G_2(s)}{1 + G_1(s)G_2(s)} = \frac{1}{G_1(s)} \frac{C}{R}(s).$$

Clearly, the influence of  $U$  on  $C$  can be reduced by decreasing  $G_2(s)$  and increasing  $G_1(s)$  (keeping the product  $G_1(s) G_2(s)$  constant), and this procedure has to be followed until  $C(s)/U(s)$  is reduced to acceptable values. This may be reasonably easy to do if the  $u$  disturbance is limited

to a certain frequency bandwidth, so that  $G_2(s)$  need be reduced only in a limited frequency region. Otherwise it may be necessary to make fairly basic changes in the control system to reduce  $G_2(s)$  and increase  $G_1(s)$ . For instance, although an attempt is usually made in good servomechanism design to choose motors and gear ratios to load to minimize moment of inertia, it may be necessary to choose gear ratio and motor to increase greatly the moment of inertia and thus reduce the susceptibility of a system to load torque disturbances.

In some instances, the extraneous inputs may dictate that certain types of systems or components cannot be used. An example of this is the stabilized platform used in long-range inertial navigation systems. The platform is maintained horizontal for short time periods by reference to gyros mounted on the platform. As the vehicle carrying the platform moves, this motion is measured by accelerometers mounted on the platform, and this information is used to precess the gyros (and therefore the platform) to keep the platform horizontal as the vehicle moves around the earth. The accelerometers also sense errors in the horizontal alignment of the table by reference to gravity. The system is shown in the simplified one-channel block diagram of Fig. 8. Angular motions of the vehicle

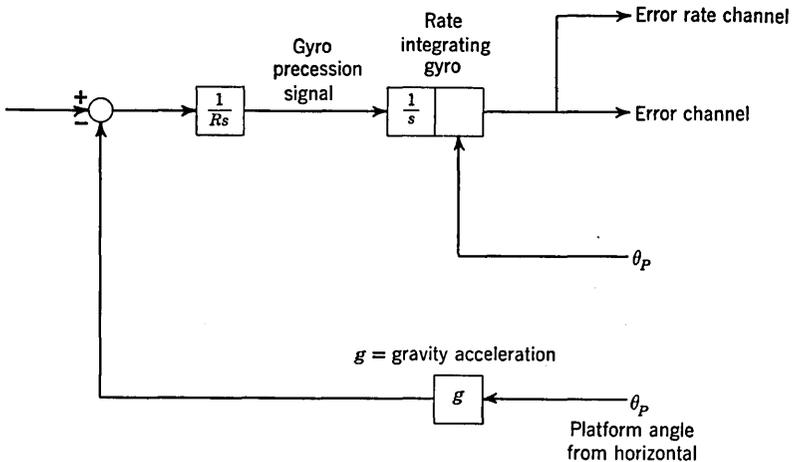


FIG. 8. Block diagram of one channel of inertial navigation system showing influence of tachometer stabilization on extraneous input of vehicle angular motion.

during the journey are an extraneous input which influences the gimbal servo design. Normally, a tachometer, geared to the shaft of a driving motor, is an acceptable way to stabilize a servo. In this case, however,

vehicle angular motion signals are amplified by such a tachometer channel and the stabilizing has to be accomplished by a rate measurement of platform motion alone such as can be obtained by a rate gyro.

### **Synthesis of Feedback Control System with Pole-Zero Location Techniques (Ref. 3)**

Synthesis of the signal channel amplification and stabilizing characteristics can also proceed from an orderly tracing of the influence of poles and zeros of transfer functions, both of the closed loop and of the open loop. In this procedure:

1. System characteristics are first translated into terms of closed loop transfer function poles and zeros.
2. By a combined graphical and numerical process, the open loop transfer function poles and zeros are determined from the closed loop poles and zeros.
3. The known power element transfer function can then be divided into the open loop transfer function to obtain series stabilizing components. A modification of this last step allows the synthesis of internal feedback transfer functions.

This procedure has an advantage over the frequency response approach in that system requirements are more easily and directly translated into the pole-zero values than frequency functions. This is particularly true of requirements concerning transient response. On the other hand, the procedure can easily lead to physically unrealizable, or at least physically impractical, transfer functions unless the proper arrangement of closed loop poles and zeros is chosen. Thus, the method in practice usually is not the direct synthesis approach that it at first appears.

The method has the disadvantage that the work involved increases rapidly with the complexity of the system which is dictated largely by the power element.

**Determination of Closed Loop Poles and Zeros from System Requirements.** The closed loop characteristics of a feedback control system are usually dominated by a pair of complex conjugate poles so that it is customary to express system performance in terms of these *dominant poles* with corrections for the influence of other poles and zeros. In the following, performance characteristics will be given in terms of a system which contains only two poles and then of more complex systems. Many of the transient characteristics associated with closed loop pole-zero configuration have already been discussed in Chap. 22, and they will be only listed here. Derivation of the relationships of Table 1 are given in Ref. 3.

TABLE 1. RELATION BETWEEN SYSTEM CHARACTERISTICS AND CLOSED LOOP POLES FOR SYSTEM CONTAINING ONLY TWO COMPLEX CONJUGATE POLES

$$p = -\zeta\omega_0 \pm j\sqrt{1 - \zeta^2}\omega_0$$

System Characteristics	Symbol	Magnitude
Time to first peak of transient response to a unit step function	$t_p$	$\frac{\pi}{\sqrt{1 - \zeta^2} \cdot \omega_0}$
Magnitude of the first peak of a transient response to a unit step	$M_p$	$\exp\left(-\frac{\pi\zeta}{\sqrt{1 - \zeta^2}}\right) + 1$
Settling time (to 2% of final value) of a transient response to a unit step	$t_s$	$\frac{4}{\zeta\omega_0}$
Number of oscillations until settling time	$N$	$\frac{2\sqrt{1 - \zeta^2}}{\pi\zeta}$ or $\frac{t_s}{2t_p}$
Bandwidth of frequency response (magnitude down 3 db)	$\omega_b$	$\omega_0\sqrt{1 - 2\zeta^2 + \sqrt{2 - 4\zeta^2 + 4\zeta^4}}$
Dynamic velocity error coefficient	$K_v$	$\frac{\omega_0}{2\zeta}$
Dynamic acceleration error coefficient	$K_a$	$\frac{\omega_0^2}{1 - 4\zeta^2}$

The quantities of Table 1 are modified by the presence of additional poles and zeros in the closed loop transfer function, as shown in Table 2.

If only the two dominant poles are involved, it is easy to convert system requirements of error coefficients, bandwidth or transient response into

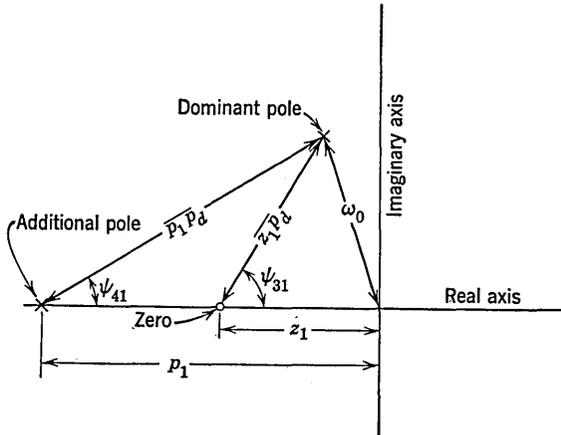


FIG. 9. Angle and distance relations between dominant pole and additional zeros and poles.

TABLE 2. RELATIONS BETWEEN SYSTEM CHARACTERISTICS AND CLOSED LOOP POLES FOR SYSTEM CONTAINING TWO COMPLEX CONJUGATE POLES PLUS ADDITIONAL POLES AND ZEROS

Characteristic	Magnitude <sup>a</sup>
$t_p$	$\frac{1}{\omega_0 \sqrt{1 - \zeta^2}} (\pi - \Sigma\psi_3 + \Sigma\psi_4)$
$M_p$	$\exp \left\{ -\frac{\pi\zeta}{\sqrt{1 - \zeta^2}} \left[ \prod_q \frac{p_q}{p_q p_d} \right] \left[ \prod_q \frac{z_q p_d}{z_q} \right] \right\} + 1$
$\omega_b$	$\omega_0 \sqrt{1 - 2\zeta^2 + (\omega_0^2/z^2)} + \sqrt{2 - 4\zeta^2 + 2(\omega_0^2/z^2) + 4\zeta^4 - 4\zeta^2(\omega_0^2/z^2) + (\omega_0^4/z^4)}$
$\frac{1}{K_v}$	$\frac{2\zeta}{\omega_0} - \Sigma \frac{1}{z_q} + \Sigma \frac{1}{p_q}$
$\frac{1}{K_a}$	$-\frac{1}{K_p^2} - \Sigma \frac{1}{p_q^2} + \Sigma \frac{1}{z_q^2}$
$t_s, N$	$t_s$ and $N$ are substantially influenced by additional poles and zeros, and it is usually not practical to handle these quantities in this way.

<sup>a</sup>  $\Sigma\psi_3$  is the sum of the angles from the additional closed loop real axis zeros to the dominant pole and  $\Sigma\psi_4$  is the sum of the angles from the real axis poles to the dominant pole.

$$\prod_q \frac{p_q}{p_q p_d}$$

is the product of the ratios of the distances from origin to the poles, to the distances from the poles to the dominant pole, and

$$\prod_q \frac{z_q p_d}{z_q}$$

is the product of the ratios of the distances from the zeros to the dominant pole, to the distances from the origin to the zeros. See Fig. 9. (These expressions are somewhat approximate, but the error is only a few per cent for practical and useful pole-zero configurations.)

$\omega_0$  and  $\zeta$ . If additional poles and zeros are involved, the transient response and bandwidth requirements still largely determine  $\omega_0$  and  $\zeta$  and the additional zeros and poles must be chosen to give the required error coefficients, by use of the proper relations listed in Table 2. Additional poles are not inserted because of any system improvement they give but only because of physical necessity. The closed loop transfer function must contain as many poles as the power element transfer function since it is physically impossible to create isolated zeros in the stabilizing transfer function to cancel out poles in the power element transfer function. The

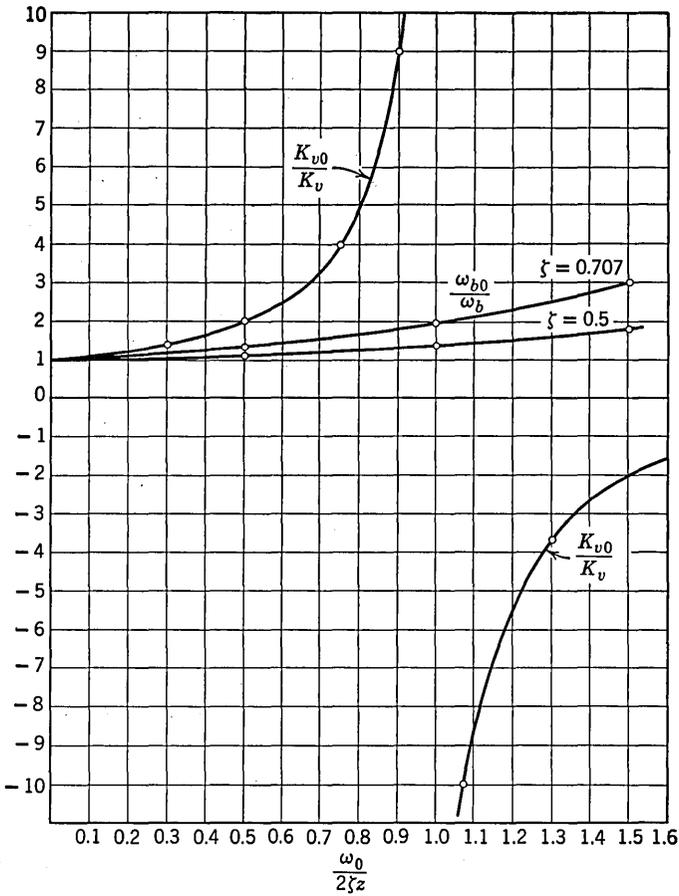


FIG. 10. Influence of zero on velocity constant and bandwidth.

$\frac{K_{v0}}{K_v}$  = ratio of velocity constant with added zero to velocity constant without added zero,

$\frac{\omega_{b0}}{\omega_b}$  = ratio of bandwidth with added zero to bandwidth without added zero,

$\frac{\omega_0}{z}$  = ratio of dominant pole to added zero,

$\zeta$  = damping factor associated with dominant poles.

best that can be done is to select closed loop poles with large negative values so that their influence on the various design requirements is small. The value of these poles should be  $3\omega_0$  or greater to reduce their importance. Zeros (with a companion pole of large magnitude) may be added purposely to the closed loop transfer function to increase the value of the error coefficients. Zeros also increase the bandwidth but not nearly as much as the error coefficients. This is shown in Fig. 10. Dipoles are very useful for increasing the error coefficients. A dipole is composed of a zero and a pole of nearly the same magnitude. The influence of a dipole on transient response and bandwidth may be very small and yet if both pole and zero have small values, the increase in the error coefficients may be considerable. This may be shown by a simple *example*. For a system with a pair of complex poles  $\omega_0 = 10$ ,  $\zeta = 0.5$ ,  $K_v = 10$ ,  $\omega_b = 13.6$ ,  $t_p = 0.3624$  and  $M_p = 1.163$ . Now add a dipole with  $p = 1$ ,  $z = 0.9524$ . The  $\omega_0$ ,  $t_p$ , and  $M_p$  values are changed to 12.72, 1.3628, and 1.149 respectively, but  $K_v$  is increased to 20.

**Determination of Open Loop Poles and Zeros from the Closed Loop Poles and Zeros.** The open loop transfer function

$$\frac{C(s)}{E(s)} = \frac{N(s)}{D(s)}$$

is desired. The closed loop transfer

$$\frac{C(s)}{R(s)} = \frac{N(s)}{D_r(s)} = \frac{N(s)/D(s)}{1 + N(s)/D(s)} = \frac{N(s)}{D(s) + N(s)}$$

is synthesized from requirements.

Clearly the open loop zeros are the same as the closed loop zeros. The open loop poles are indicated by the expression,  $D_r(s) - N(s)$ . This can be solved by a straightforward procedure forming the  $D_r(s)$  polynomial in  $s$  by multiplying the  $(s + p_1)(s + p_2)$ , etc., factors together where  $p_1, p_2$ , etc., are the closed loop poles, similarly forming the  $N(s)$  polynomial from the zeros, performing the indicated subtraction algebraically, and then factoring the resulting expression for the open loop poles.

The last two steps of this procedure may be replaced by a graphical process in which  $D_r(s)$  and  $N(s)$  are plotted as a function of the real variable,  $s = \sigma$ , for negative values. The two plotted functions intersect at values of  $s$ , which are equal to the poles of  $D(s)$ . This graphical method works out easily only if a proper choice of closed loop poles and zeros is made. First of all, except for the dominant complex pair, all the poles should be real. Then the plot of  $D_r(s)$  crosses the axis at the values of these poles,

and the plotting is somewhat simplified. Secondly, there should be only one zero so that  $N(s)$  can be plotted as a straight line. Typical *examples* are shown in Figs. 11 and 12.

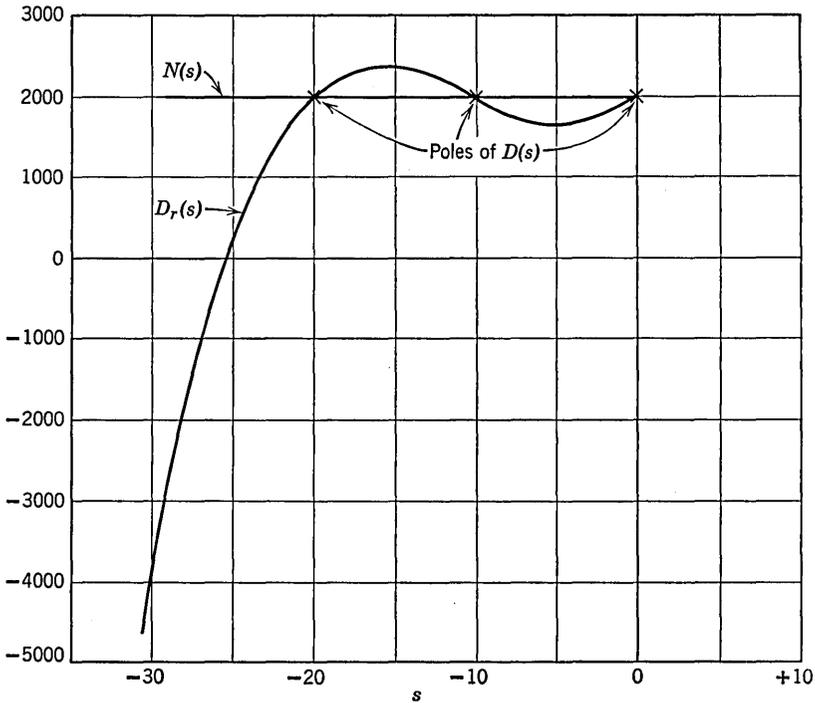


FIG. 11. Example showing graphical solution for  $D(s)$  from plots of  $D_r(s)$  and  $N(s)$ .

$$\begin{aligned}
 N(s) &= 2000 \text{ (no zeros),} \\
 D_r(s) &= s^3 + 30s^2 + 200s + 2000 \\
 &= (s + 25.2)(s + 2.4 + j8.55)(s + 2.4 - j8.55), \\
 D(s) &= s(s + 10)(s + 20).
 \end{aligned}$$

Once the open loop poles and zeros have been determined, and therefore the open loop transfer function, the power element transfer function may be divided into it thus obtaining the required added transfer function. That is, if the required open loop transfer function were found to be  $N(s)/D(s)$  and the power element transfer function were  $N_p(s)/D_p(s)$ , then the added or series compensating transfer function would be  $N(s)/D(s) \times D_p(s)/N_p(s)$ .

In some cases, stabilization by internal feedback may be desired. If  $N(s)/D_r(s)$  is the required closed loop transfer function and  $N_p(s)/D_p(s)$

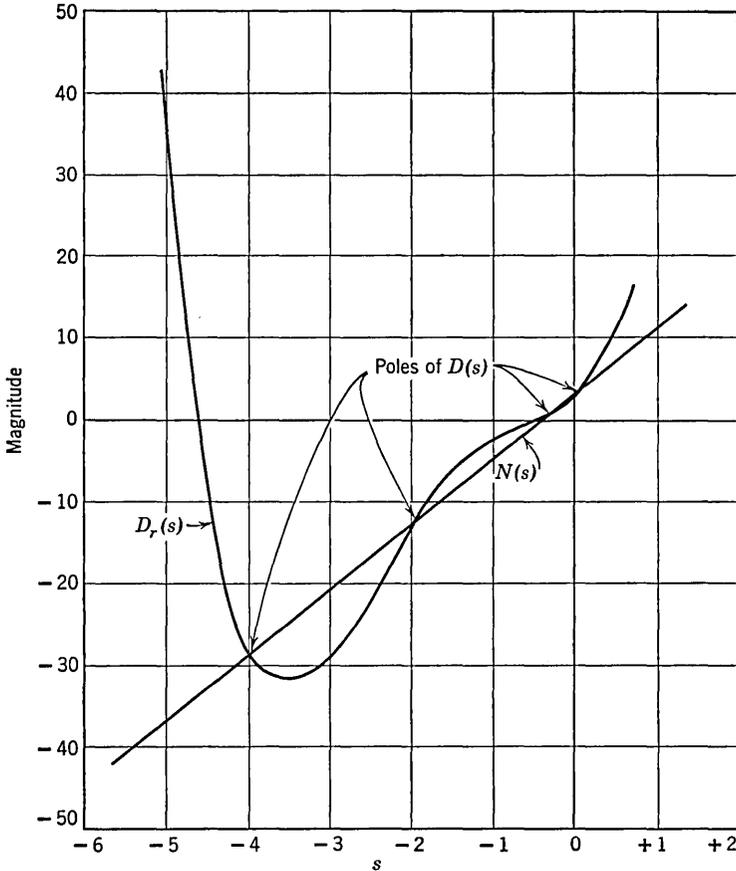


FIG. 12. Example showing graphical solution for  $D(s)$  from plots of  $D_r(s)$  and  $N(s)$ .

$$\begin{aligned}
 N(s) &= 8s + 3.2, \\
 D_r(s) &= s^4 + 6.2s^3 + 9.2s^2 + 9.6s + 3.2 \\
 &= (s + 4.627)(s + 0.497)(s + 0.538 \pm j1.044), \\
 D(s) &= s(s + 2)(s + 2)(s + 4).
 \end{aligned}$$

is the power element transfer function, then the feedback transfer function,

$$H(s) = \frac{N_h(s)}{D_h(s)} = \frac{D_r(s)}{N(s)} - 1 - \frac{D_p(s)}{N_p(s)}.$$

The above equation can be written in terms of the open loop transfer function  $N(s)/D(s)$  as

$$\frac{N_h(s)}{D_h(s)} = \frac{D(s)}{N(s)} - \frac{D_p(s)}{N_p(s)}.$$

The feedback transfer function can be derived in a straightforward manner from either of the above equations. A graphical procedure similar to that used for the series stabilized case is not feasible except for very simple examples.

## 2. COMPENSATING COMPONENTS: D-C SYSTEMS

After the power element, with known log magnitude diagram, has been chosen, the task remains to add control components to the system so that the system log magnitude diagram has the required characteristics, as described in Sect. 1. There are two general ways of adding control components to achieve this purpose:

1. Add series stabilizing and control components in the signal channels which control the power element.
2. Add internal feedback control loops around the power element.

### Series Compensation

Series components may be phase lead, phase lag or a combination. In Figs. 13, 14, and 15, a sample log magnitude diagram for a power element is shown. This sample is typical of an electric motor and many other power elements. This power element is compensated by the three different series components mentioned above.

**Phase Lead Compensation.** Figure 13 shows the log magnitude diagram for a phase lead component having the transfer function

$$\begin{aligned} \frac{M}{E}(s) &= \frac{K(1 + T_1s)}{(1 + T_2s)} \\ &= \frac{K(1 + j\omega/\omega_1)}{(1 + j\omega/\omega_2)} \text{ in frequency function form.} \end{aligned}$$

$K$  is a constant of magnitude that gives the required static error coefficient when multiplied by the gain term in the power element.  $\omega_1$  is chosen so as to make crossover occur within a long section of  $-20$  db per decade slope. This insures adequate phase margin at crossover. Figure 13 also shows the resulting system log magnitude diagram. The phase margin at crossover may be increased or decreased by adjusting the  $\omega_2/\omega_1$  ratio.

**Phase Lag Compensation.** Figure 14 shows the log magnitude diagram for a phase lag component having the transfer function

$$\begin{aligned} \frac{M}{E}(s) &= \frac{K(1 + T_2s)}{(1 + T_1s)} \\ &= \frac{K(1 + j\omega/\omega_2)}{(1 + j\omega/\omega_1)} \text{ in frequency function form.} \end{aligned}$$

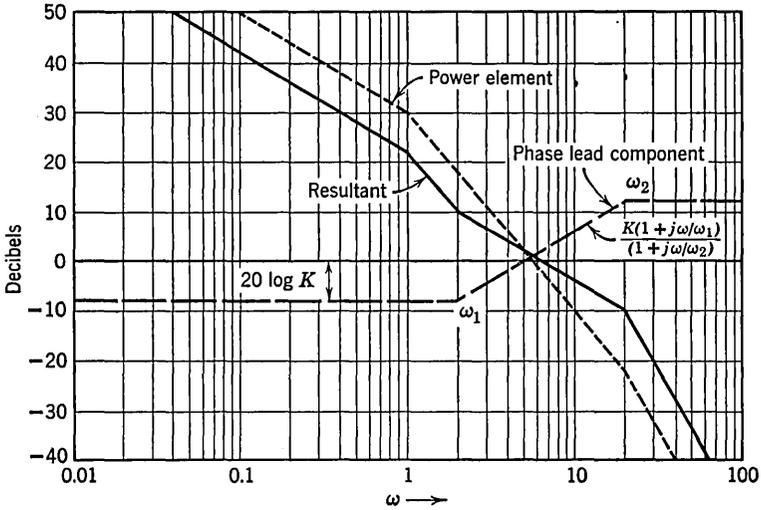


FIG. 13. Phase lead compensation.

The frequencies  $\omega_1$  and  $\omega_2$  must be chosen so that the gain of the system is reduced at a low enough frequency to cause crossover in the  $-20$  db per decade region of the power element. Figure 14 shows the resultant log magnitude diagram.

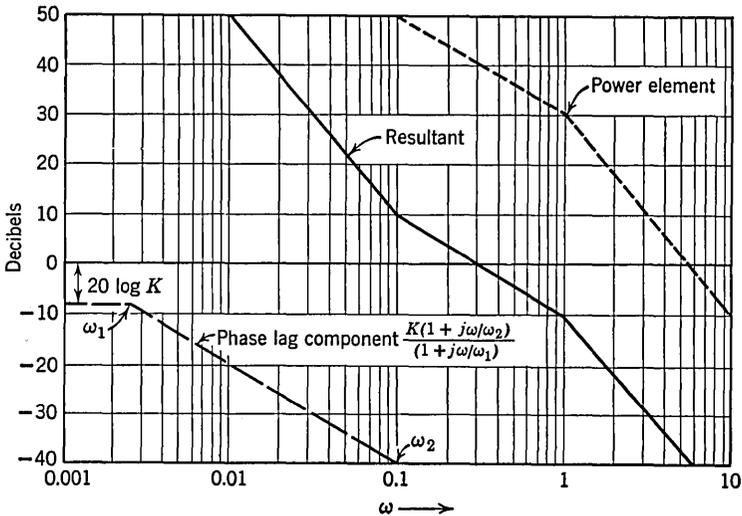


FIG. 14. Phase lag compensation.

**Lag-Lead Compensation.** Figure 15 shows the log magnitude diagram for a lag-lead component having the transfer function

$$\frac{M}{E}(s) = \frac{K(1 + T_2s)(1 + T_3s)}{(1 + T_1s)(1 + T_4s)} = \frac{K(1 + j\omega/\omega_2)(1 + j\omega/\omega_3)}{(1 + j\omega/\omega_1)(1 + j\omega/\omega_4)}$$

Figure 15 also shows the resultant system log magnitude diagram.

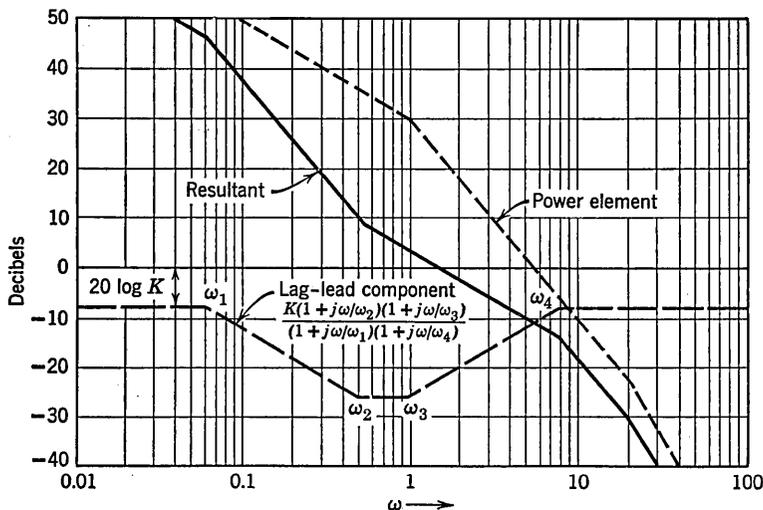


FIG. 15. Lag-lead compensation.

In the examples given above, the simplest components were used. These had log magnitude slopes of +20 or -20 db per decade. Other slopes and characteristics are frequently required and are available in networks as given in a later section.

**Comparison of Networks.** For a given static error coefficient, the phase lead compensation gives the highest crossover frequency. This may make the system susceptible to noise in the signal channel. The phase lag compensation gives the lowest crossover frequency and tends to make the system sluggish and have large errors during transients. The lag-lead compensation gives a situation which is intermediate between these two extremes. By proper adjustment of time constants, considerable flexibility of system characteristics may be obtained.

The lead network has an inherent attenuation at low frequencies which has to be balanced by additional amplifier gain to achieve a desired error coefficient. In such a system there is an added danger of amplifier saturation during synchronizing transients. Saturation can lead to instability unless the lead network is located ahead of the saturating amplifier in the

error channel. On the other hand, a lag network reduces the chance of amplifier saturation, but if saturation does occur in the error channel beyond the lag network, instability may result.

In *type two, or higher, systems, a lag network cannot be used.*

**Feedback Compensation**

This method of compensation uses the principle that the closed loop characteristic is nearly equal to the inverse of the feedback characteristic whenever the open loop gain is much larger than one. In this way, the desired open loop log magnitude characteristics can be obtained by adding

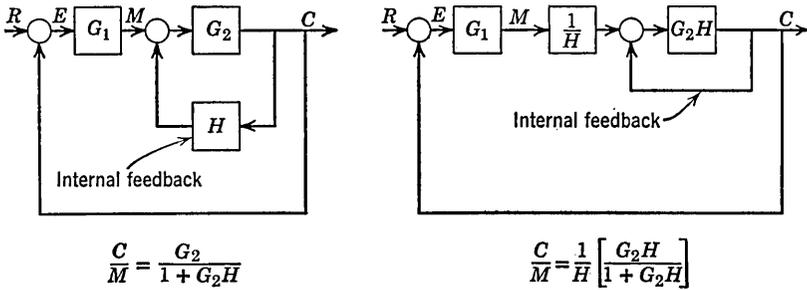


FIG. 16. Block diagram of system using internal feedback.

internal feedback components with log magnitude characteristics equal to the inverse of the desired open loop characteristic. This principle can be demonstrated by the equation of the inner closed loop (see Fig. 16).

$$\frac{C}{M}(s) = \frac{G_2(s)}{(1 + G_2(s)H(s))} = \frac{1}{H(s)} \frac{G_2(s)H(s)}{1 + G_2(s)H(s)}$$

Whenever  $G_2(s)H(s)$  is much larger than unity,  $C/M$  approaches  $1/H(s)$ . Figure 16 shows the equivalent block diagram transformation. The feedback compensation may be in a variety of forms:

**Rate Feedback.** Many different types of rate measuring components for feedback control systems are available. Commonly used types are:

1. Tachometers, which are mechanically connected to the power element of angular position control systems.
2. Stabilizing transformers, which effectively measure the rate of change of voltage in regulating systems.
3. Rate gyros, which measure the rate of change of airplane heading in automatic pilot systems.

Figure 17 shows the log magnitude diagram of the same power element used in the section on series stabilization. It also shows the log magnitude

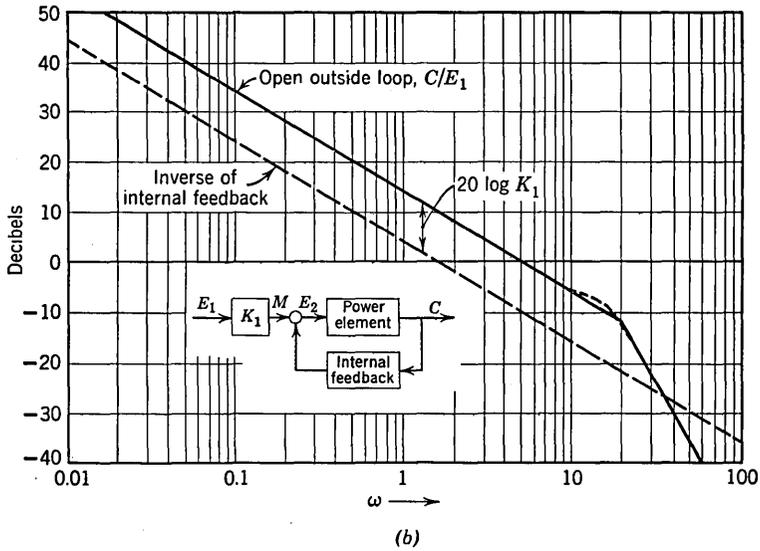
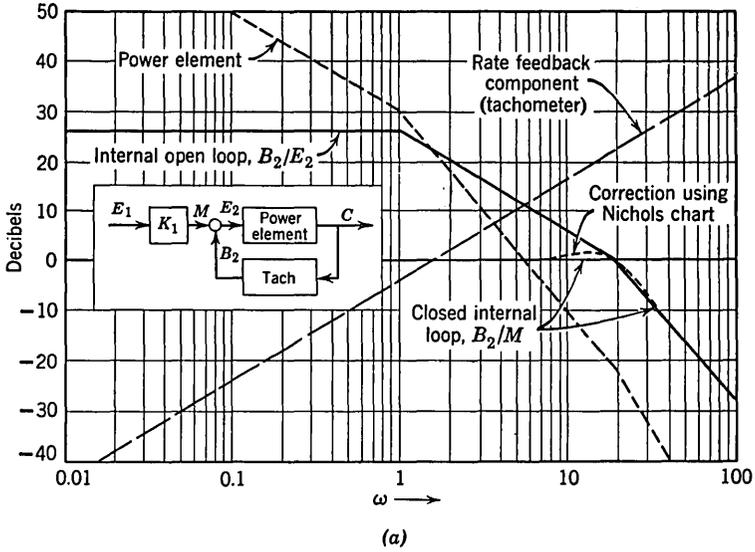


FIG. 17. Internal rate feedback compensation.

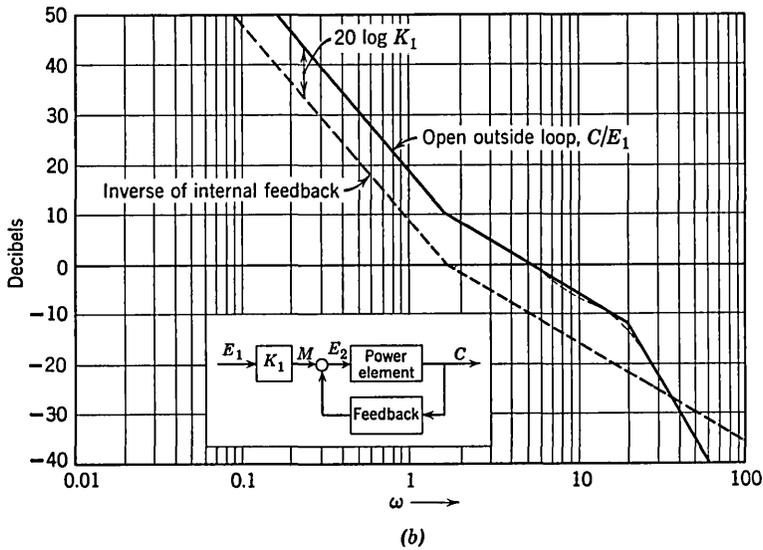
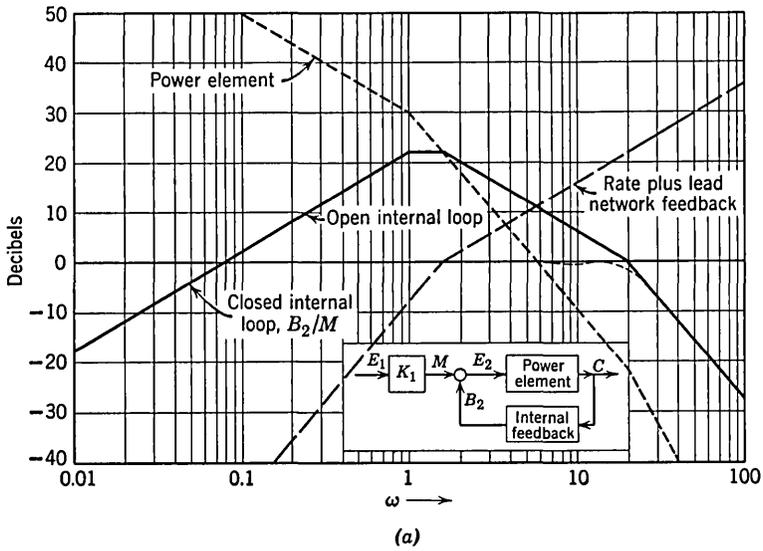


FIG. 18. Internal rate plus lead network feedback.

and phase diagram for an internal rate feedback element, the closed internal loop and the log magnitude diagram of the outside open loop. The gain of the internal loop must be greater than unity at frequencies up to and somewhat beyond the desired crossover frequency of the outside loop and the gain,  $K_1$ , may be set to give the proper crossover.

Much of Fig. 17 can be constructed by using approximate straight line diagrams, but portions of the diagram in the frequency region near crossover of the internal loop should be corrected by using accurate magnitude and phase values from the log magnitude-angle diagram (Nichols chart, see Chap. 21).

**Rate and Lead Network Feedback.** In some systems, it is necessary to obtain higher gain at the low frequencies. This can be obtained in a system using internal feedback by adding a lead network to the rate feedback. Figure 18 shows the log magnitude and phase diagram of such an internal feedback element along with the system diagram leading to the open loop diagram of  $C/E(s)$ .

For the higher order lead networks, the inner loop may be unstable by itself. In such a situation it is necessary to determine the number of

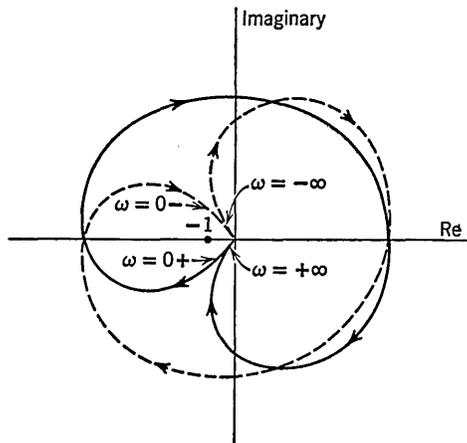


FIG. 19. Nyquist diagram of unstable inner loop as indicated by two clockwise rotations.

positive real poles in the closed inner loop in order to apply a stability criterion to the outer loop. An *example* of such a system is shown in Figs. 19 and 20. The closed inner loop contains two positive real poles as indicated by the two encirclements of the  $-1$  point by the Nyquist sketch of the open inner loop transfer function. For the outer loop, and therefore the whole system, to be stable, the Nyquist plot must encircle the  $-1$  point twice counterclockwise. It does this as indicated in Fig. 20.

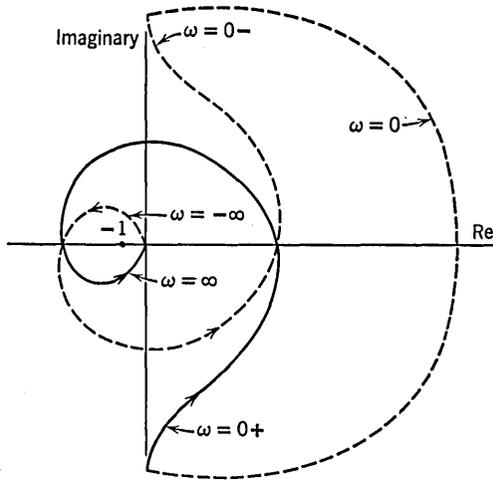


FIG. 20. Nyquist diagram of outside loop. Stability is indicated by two counter-clockwise rotations about  $-1$ .

**Lead Network Feedback.** A lead network is a rate measuring device that is somewhat inferior in performance to the components mentioned in the paragraph above, but because of simplicity and low cost is often used in place of these more expensive components. Such a network is equivalent to a tachometer, or other rate device, at low frequencies, but it does not have rate characteristics above a frequency which is equal to  $1/T$  where  $T$

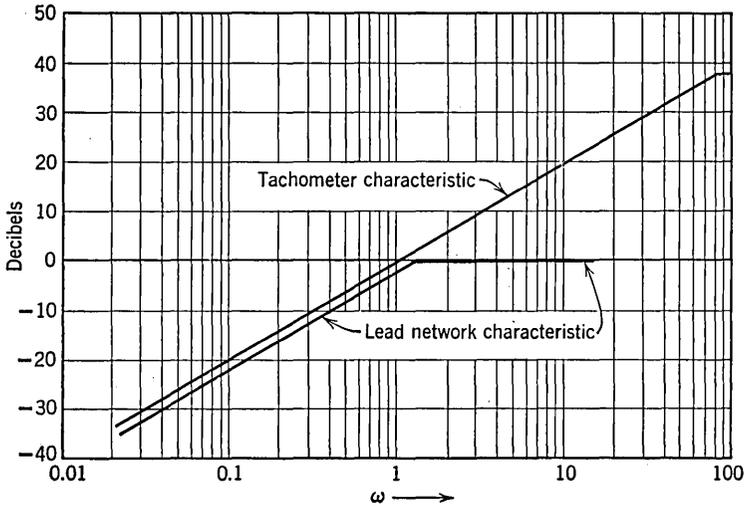


FIG. 21. Comparison of tachometer and lead network characteristics.

is the time constant of the network. Figure 21 shows a comparison of tachometer and network characteristics. The tachometer also has a high-frequency droop in its log magnitude diagram, but this is usually well above any frequencies of interest in the feedback control system. The lead network can also have a rate characteristic out to high frequencies by reducing the network time constant, but this lowers the gain of the circuit at the frequencies of interest. Such a lead network feedback is particularly useful in systems in which a d-c voltage is one of the intermediate outputs. An *example* is the voltage rate feedback around the amplidyne in a voltage regulator.

**Multiloop Systems.** In the preceding discussion, internal feedback loops have been closed to form a portion of the open outside loop of a feedback control system. In the same way this second, or outside, loop can be closed by using the log magnitude-angle diagram, and becomes a portion of a third feedback control loop. This procedure can be extended to any number of concentric feedback loops, such as may be present in a complex feedback control system. However the block diagram of complex control systems often is not in the form of concentric loops. Chapter 20 shows how intertwined block diagrams can usually be transformed into concentric loops by making use of superposition rules.

### Alternate Methods of Representation

The preceding paragraphs have shown how the log magnitude diagram of a power element can be modified by the addition of series or feedback components to obtain the log magnitude diagrams synthesized in Sec. 1. There are several other forms in which these same data may be presented and handled to obtain the same desired results. Some of the more commonly used forms are the *Nyquist diagram*, the *inverse complex plane diagram*, and the *root locus plot*.

**Nyquist Diagram and Inverse Complex Plane Diagram.** The use of these diagrams is described in detail in Chap. 9 of Ref. 1. Since these diagrams contain exactly the same information as the log magnitude diagram, essentially the same principles as described above may be used. The steps may be summarized:

1. Select the starting axis (type of system) and gain factor from the static error coefficient requirements.
2. From stability and transient response requirements, determine the maximum allowable  $M$  and draw in this  $M$  circle.
3. By using the gains established in step 1 and the chosen power element, draw a Nyquist diagram.
4. Add frequency sensitive networks (or proper internal feedback loops) as needed to reshape the diagram to avoid the required  $M$  contour. See

Fig. 22. (This is a trial and error process which will become more efficient with the user's experience.)

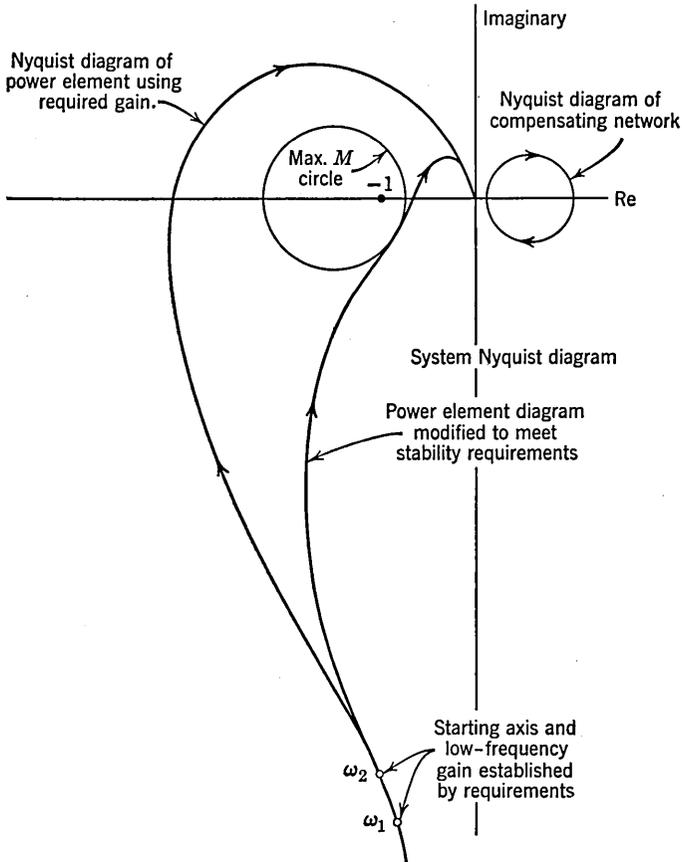


FIG. 22. Nyquist diagram showing synthesis procedure.

The Nyquist diagram is used in this discussion because of its historical position, although it is somewhat easier to use the inverse complex plane plot in this type presentation.

**Root Locus Plots.** (See Ref. 6.) This is essentially a complex plane graphical representation of the pole-zero configuration synthesis presented in Sect. 1. In this plot, the locations of the closed loop poles are traced on the complex plane as the open loop gain is varied. Use of this diagram may be broadly outlined in steps analogous to those of the preceding paragraph:

1. Select the closed loop poles from system specifications of performance and stability. These may be located on the complex plane.

2. Start with the poles and zeros of the power element and draw the root locus plot.

3. Add loop pole and zero combinations to modify the root locus plot to pass through the required closed loop poles. Reference 6 indicates optimum selections of added pole and zero configurations to achieve the desired changes in locus shape.

Figure 23 indicates the above steps.

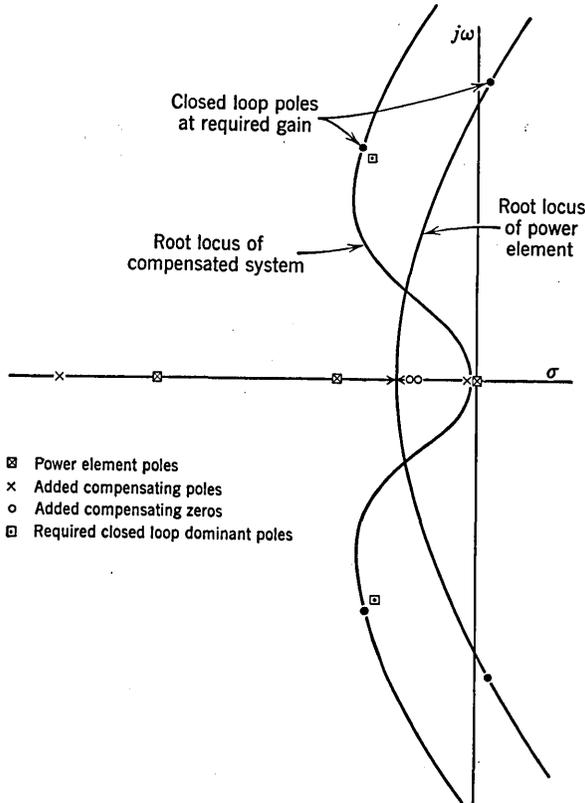


FIG. 23. Root locus plots showing synthesis procedure.

**Comparison of Alternate Methods of Representation with the Log Magnitude Diagram.** The log magnitude and phase diagrams contain the same information as the Nyquist or inverse complex plane diagram. Because the log magnitude diagram is much easier to construct than the other two diagrams, and required modifications to meet specifications are

more easily visualized and constructed on the log magnitude diagram, there is normally no reason for using the Nyquist or inverse diagram in control system design work. In very complex systems, containing several positive real roots, it may be desirable to make rough order of magnitude Nyquist sketches to check rotations about the  $-1$  point in order to check stability but the actual numerical work should be done using log magnitude diagrams.

The log magnitude diagram is also easier to construct than the root locus plot and would normally be used for problems concerned with stability, bandwidth, static and low-frequency errors. However, the root locus plot has more specific information regarding actual transient response characteristics and damping factor, and would be used in problems in which this type of information is of primary importance.

### Design Aids

**Charts of Electric Networks.** Tables 3 through 7 show many of the electric networks useful in compensating d-c systems. All these networks are of the resistor and capacitor type since any practical type of frequency characteristic can be obtained with these components. Inductance is also a useful circuit component, but large time constants cannot be obtained in sizes competitive with resistance-capacitance components.

TABLE 3. STABILIZING NETWORKS: LEAD NETWORKS WITH 20 DB/DECADE SLOPE (Ref. 2)

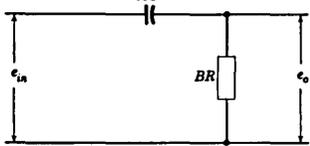
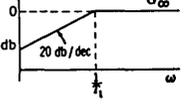
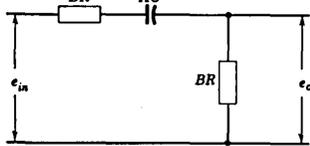
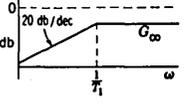
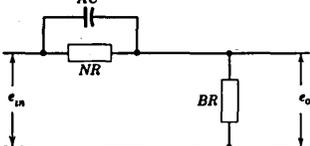
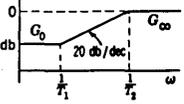
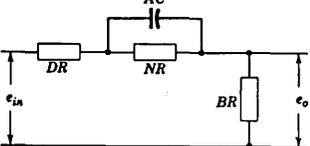
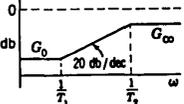
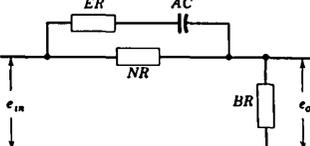
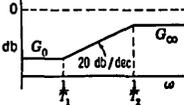
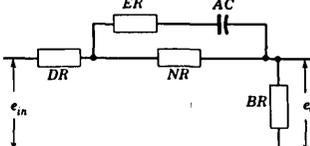
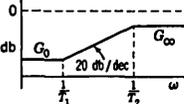
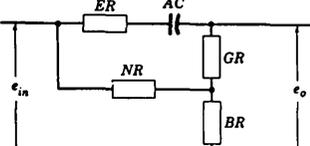
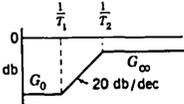
Network	Attenuation Characteristic
<p>(a)</p> 	 <p><math>G_0 = 0</math>      <math>G_\infty = 1</math></p>
<p>(b)</p> 	 <p><math>G_0 = 0</math>      <math>G_\infty = \frac{1}{1 + \frac{D}{B}}</math></p>
<p>(c)</p> 	 <p><math>G_0 = \frac{1}{1 + \frac{N}{B}}</math>      <math>G_\infty = 1</math></p>
<p>(d)</p> 	 <p><math>G_0 = \frac{1}{1 + \frac{D+N}{B}}</math>      <math>G_\infty = \frac{1}{1 + \frac{D}{B}}</math></p>
<p>(e)</p> 	 <p><math>G_0 = \frac{1}{1 + \frac{N}{B}}</math>      <math>G_\infty = \frac{1}{1 + \frac{EN}{B(E+N)}}</math></p>
<p>(f)</p> 	 <p><math>G_0 = \frac{1}{1 + \frac{D+N}{B}}</math>      <math>G_\infty = \frac{1}{1 + \frac{D}{B} + \frac{EN}{B(E+N)}}</math></p>
<p>(g)</p> 	 <p><math>G_0 = \frac{B}{B+N}</math>      <math>G_\infty = \frac{B(E+G+N) + GN}{(B+N)(E+G) + BN}</math></p>

TABLE 3. STABILIZING NETWORKS (Continued)

Transfer Function	$T_1$	$T_1$
$\frac{T_{1s}}{T_{1s} + 1}$	$ABRC$	...
$\frac{T_{2s}}{T_{2s} + 1}$	$A(B + D)RC$	$ABRC$
$G_0 \frac{(T_{1s} + 1)}{(T_{2s} + 1)}$	$ANRC$	$\frac{B}{B + N} T_1$
$G_0 \frac{(T_{1s} + 1)}{(T_{2s} + 1)}$	$ANRC$	$\left[ \frac{B + D}{B + D + N} \right] T_1$
$G_0 \frac{(T_{1s} + 1)}{(T_{2s} + 1)}$	$A(E + N)RC$	$\left[ \frac{B + \frac{EN}{E + N}}{B + N} \right] T_1$
$G_0 \frac{(T_{1s} + 1)}{(T_{2s} + 1)}$	$A(E + N)RC$	$\left[ \frac{B + D + \frac{EN}{E + N}}{B + D + N} \right] T_1$
$G_0 \frac{(T_{1s} + 1)}{(T_{2s} + 1)}$	$A \left[ (E + G + N) + \frac{GN}{B} \right] RC$	$A \left[ (E + G) + \frac{BN}{B + N} \right] RC$

TABLE 4. STABILIZING NETWORKS: LEAD NETWORKS WITH 40 DB/DECADE SLOPE (Ref. 2)

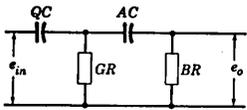
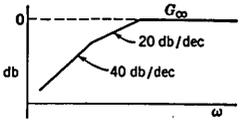
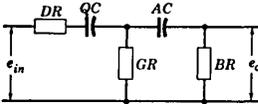
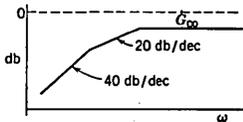
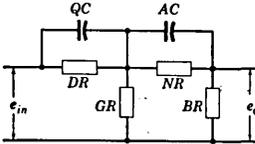
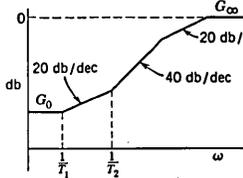
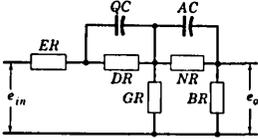
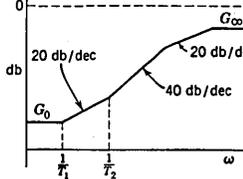
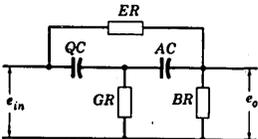
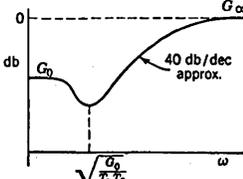
Network	Attenuation Characteristic
 <p>(a)</p>	 <p><math>G_0 = 0</math>                      <math>G_\infty = 1</math></p>
 <p>(b)</p>	 <p><math>G_0 = 0</math>                      <math>G_\infty = \frac{1}{1 + \frac{(B+G)D}{BG}}</math></p>
 <p>(c)</p>	 <p><math>G_0 = \frac{1}{1 + \frac{N}{B} + \frac{(B+G+N)D}{BG}}</math>                      <math>G_\infty = 1</math></p>
 <p>(d)</p>	 <p><math>G_0 = \frac{1}{1 + \frac{N}{B} + \frac{(B+G+N)(D+E)}{BG}}</math>                      <math>G_\infty = \frac{1}{1 + \frac{(B+G)E}{BG}}</math></p>
 <p>(e)</p>	 <p><math>G_0 = \frac{1}{1 + \frac{E}{B}}</math>                      <math>G_\infty = 1</math></p>

TABLE 4. STABILIZING NETWORKS (Continued)

Transfer Function	$T_1$	$T_2$
$\frac{T_1 T_2 s^2}{T_1 T_2 s^2 + \left[ T_1 \left( 1 + \frac{G}{B} \right) + T_2 \right] s + 1}$	ABRC	GQRC
$\frac{T_1 T_2 s^2}{T_1 T_2 s^2 \left[ 1 + \frac{(B+G)D}{BG} \right] + \left[ T_1 \left( 1 + \frac{G}{E} \right) + T_2 \left( 1 + \frac{D}{G} \right) \right] s + 1}$	ABRC	GQRC
$\frac{(T_1 s + 1)(T_2 s + 1)}{T_1 T_2 s^2 + \left\{ T_1 \left( 1 + \frac{N}{B} \right) + T_2 \left[ 1 + \frac{(B+G)D}{BG} \right] \right\} s + \frac{1}{G_0}}$	DQRC	$ANRC = \frac{AN}{DQ} T_1$
$\frac{(T_1 s + 1)(T_2 s + 1)}{T_1 T_2 \left[ 1 + \frac{(B+G)E}{BG} \right] s^2 + \left\{ T_1 \left[ 1 + \frac{N}{B} + \frac{(B+G+N)E}{BG} \right] + T_2 \left[ 1 + \frac{(B+G)(D+E)}{BG} \right] \right\} s + \frac{1}{G_0}}$	DQRC	$ANRC = \frac{AN}{DQ} T_1$
$\frac{T_1 T_2 s^2 + \left[ T_1 \frac{G}{E} + T_2 \left( \frac{B}{B+E} \right) \right] s + G_0}{T_1 T_2 s^2 + \left[ T_1 + T_2 \left( 1 + \frac{A}{Q} \right) \right] s + 1}$	$A \frac{BE}{B+E} RC$	GQRC

TABLE 5. STABILIZING NETWORKS: LAG NETWORKS WITH 20 DB/DECADE SLOPE (Ref. 2)

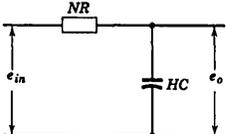
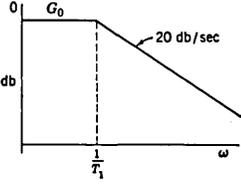
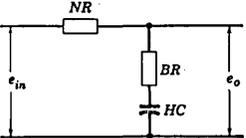
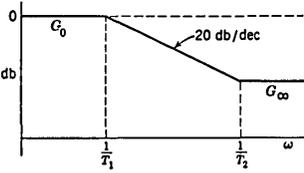
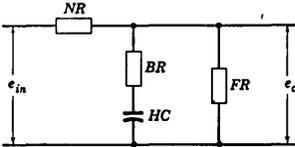
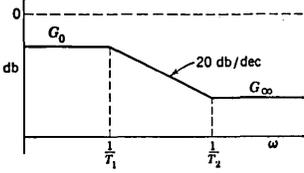
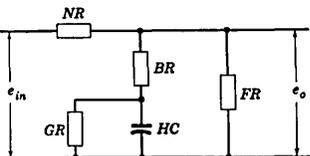
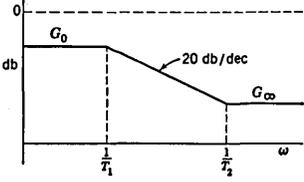
Network	Attenuation Characteristic
 <p>(a)</p>	 <p><math>G_0 = 1</math>                      <math>G_\infty = 0</math></p>
 <p>(b)</p>	 <p><math>G_0 = 1</math>                      <math>G_\infty = \frac{1}{1 + \frac{N}{B}}</math></p>
 <p>(c)</p>	 <p><math>G_0 = \frac{1}{1 + \frac{N}{F}}</math>                      <math>G_\infty = \frac{1}{1 + \frac{N}{B} + \frac{N}{F}}</math></p>
 <p>(d)</p>	 <p><math>G_0 = \frac{1}{1 + \frac{N}{B} + \frac{N}{G} + \frac{N}{F}}</math>                      <math>G_\infty = \frac{1}{1 + \frac{N}{B} + \frac{N}{F}}</math></p>

TABLE 5. STABILIZING NETWORKS (Continued)

Transfer Function	$T_1$	$T_2$
$\frac{1}{T_1 s + 1}$	$NHRC$	$0$
$\frac{T_2 s + 1}{T_1 s + 1}$	$\frac{(B + N)}{B} T_2$	$BHRC$
$G_o \frac{(T_2 s + 1)}{(T_1 s + 1)}$	$\left[ 1 + \frac{FN}{B(F + N)} \right] T_2$	$BHRC$
$G_o \frac{(T_2 s + 1)}{(T_1 s + 1)}$	$\left[ \frac{1 + \frac{N}{B} + \frac{N}{F}}{1 + \frac{N}{B + G} + \frac{N}{F}} \right] T_2$	$\left( \frac{BG}{B + G} \right) HRC$

TABLE 6. STABILIZING NETWORKS: LAG NETWORKS WITH 40 DB/DECADE SLOPE (Ref. 2)

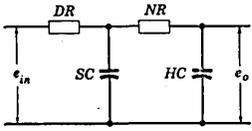
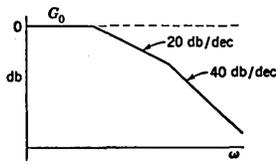
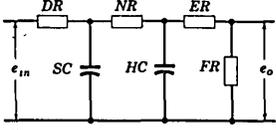
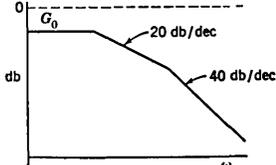
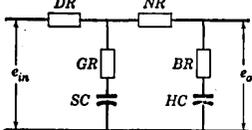
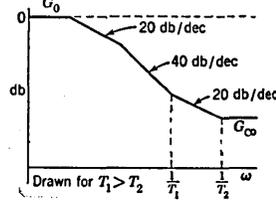
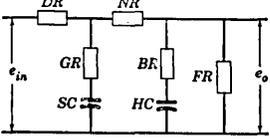
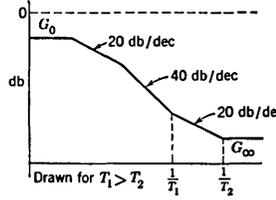
Network	Attenuation Characteristic
 <p>(a)</p>	 <p><math>G_0 = 1</math>      <math>G_\infty = 0</math></p>
 <p>(b)</p>	 <p><math>G_0 = \frac{F}{D + E + F + N}</math>      <math>G_\infty = 0</math></p>
 <p>(c)</p>	 <p><math>G_0 = 1</math>      <math>G_\infty = \frac{1}{\left(1 + \frac{D}{G}\right) \left(1 + \frac{N}{B}\right) + \frac{D}{B}}</math></p>
 <p>(d)</p>	 <p><math>G_0 = \frac{F}{D + F + N}</math>      <math>G_\infty = \frac{\frac{F}{F + N}}{\left[1 + \frac{FN}{B(F + N)}\right] \left[1 + \frac{D}{G}\right] + \frac{D(B + F)}{B(F + N)}}</math></p>

TABLE 6. STABILIZING NETWORKS (Continued)

Transfer Function	$T_1$	$T_2$
$\frac{1}{T_1 T_2 s^2 + \left[ T_1 \left( 1 + \frac{D}{N} \right) + T_2 \right] s + 1}$	HNRC	DSRC
$\frac{1}{T_1 T_2 s^2 + \left[ T_1 \left( 1 + \frac{D}{N} \right) + T_2 \left( \frac{E+F+N}{F} \right) \right] s + \frac{1}{G_0}}$	$\frac{E+F}{F}$ HNRC	DSRC
$\frac{(T_1 s + 1)(T_2 s + 1)}{T_1 T_2 \left[ \left( 1 + \frac{D}{G} \right) \left( 1 + \frac{N}{B} \right) + \frac{D}{B} \right] s^2 + \left[ T_1 \left( 1 + \frac{D}{B} + \frac{N}{B} \right) + T_2 \left( 1 + \frac{D}{G} \right) \right] s + 1}$	BHRC	GSRC
$\frac{\frac{F}{F+N} (T_1 s + 1)(T_2 s + 1)}{T_1 T_2 \left\{ \left[ 1 + \frac{FN}{B(F+N)} \right] \left[ 1 + \frac{D}{G} \right] + \frac{D(B+F)}{B(F+N)} \right\} s^2 + \left\{ T_1 \left[ 1 + \frac{FN}{B(F+N)} + \frac{D(B+F)}{B(F+N)} \right] + T_2 \left( 1 + \frac{D}{G} + \frac{D}{F+N} \right) \right\} s + \frac{F}{G_0(F+N)}}$	BHRC	GSRC

TABLE 7. STABILIZING NETWORKS: LEAD-LAG NETWORKS (Ref. 2)

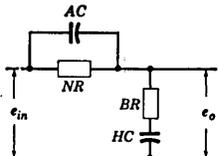
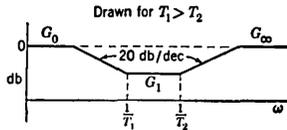
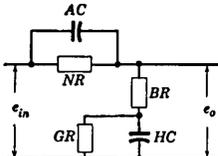
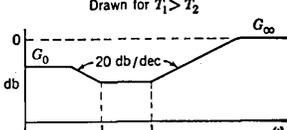
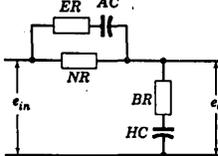
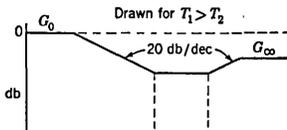
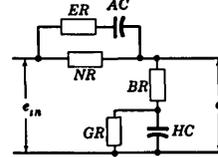
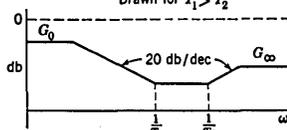
Network	Attenuation Characteristic
 <p>(a)</p>	<p>Drawn for <math>T_1 &gt; T_2</math></p>  $G_1 = \frac{T_1 + T_2}{T_1(1 + \frac{N}{B}) + T_2}$ <p><math>G_0 = 1</math>                      <math>G_\infty = 1</math></p>
 <p>(b)</p>	<p>Drawn for <math>T_1 &gt; T_2</math></p>  $G_0 = \frac{1}{1 + \frac{N}{B + G}}$ <p><math>G_\infty = 1</math></p>
 <p>(c)</p>	<p>Drawn for <math>T_1 &gt; T_2</math></p>  <p><math>G_0 = 1</math>                      <math>G_\infty = \frac{1}{1 + \frac{EN}{B(E + N)}}</math></p>
 <p>(d)</p>	<p>Drawn for <math>T_1 &gt; T_2</math></p>  $G_0 = \frac{1}{1 + \frac{N}{B + G}}$ <p><math>G_\infty = \frac{1}{1 + \frac{EN}{B(E + N)}}</math></p>

TABLE 7. STABILIZING NETWORKS (Continued)

Transfer Function	$T_1$	$T_2$
$\frac{(T_1s + 1)(T_2s + 1)}{T_1T_2s^2 + \left[ T_1 \left( 1 + \frac{N}{B} \right) + T_2 \right] s + 1}$	BHRC	ANRC
$\frac{(T_1s + 1)(T_2s + 1)}{T_1T_2s^2 + \left[ T_1 \left( 1 + \frac{N}{B} \right) + T_2 \right] s + \frac{1}{G_0}}$	$\frac{BG}{B+G}$ HRC	ANRC
$\frac{(T_1s + 1)(T_2s + 1)}{T_1T_2 \left[ 1 + \frac{EN}{B(E+N)} \right] s^2 + \left[ T_1 \left( 1 + \frac{N}{B} \right) + T_2 \right] s + 1}$	BHRC	$A(E+N)RC$
$\frac{(T_1s + 1)(T_2s + 1)}{T_1T_2 \left[ 1 + \frac{EN}{B(E+N)} \right] s^2 + \left\{ T_1 \left( 1 + \frac{N}{B} \right) + T_2 \left[ 1 + \frac{EN}{(B+G)(E+N)} \right] \right\} s + \frac{1}{G_0}}$	$\frac{BG}{B+G}$ HRC	$A(E+N)RC$

TABLE 7. STABILIZING NETWORKS (Continued)

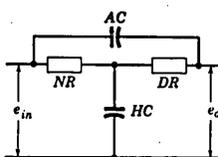
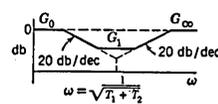
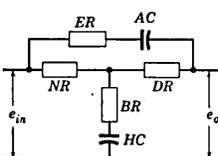
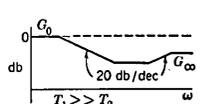
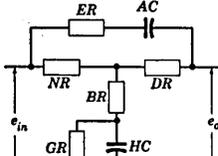
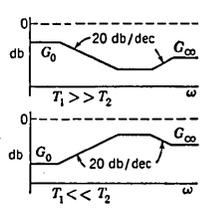
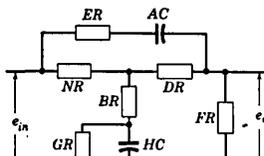
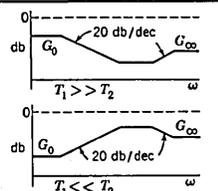
Network	Attenuation Characteristic
 <p>(e)</p>	 $G_1 = \left(1 + \frac{D}{N}\right) \frac{T_2}{T_1} + 1$ $\omega = \sqrt{\frac{1}{T_1} + \frac{1}{T_2}}$ <p><math>G_0 = 1</math>                      <math>G_\infty = -1</math></p>
 <p>(f)</p>	 $G_0 = 1$ $G_\infty = \frac{1 + \frac{DN}{B(D+E+N)}}{1 + \frac{N(D+E)}{B(D+E+N)}}$ <p><math>T_1 \gg T_2</math></p>
 <p>(g)</p>	 $G_0 = \frac{1}{1 + \frac{N}{B+G}}$ $G_\infty = \frac{1 + \frac{DN}{B(D+E+N)}}{1 + \frac{N(D+E)}{B(D+E+N)}}$ <p><math>T_1 \gg T_2</math> (top curve) <math>T_1 \ll T_2</math> (bottom curve)</p>
 <p>(h)</p>	 <p>Attenuation curves are for fairly large values of <math>F</math> only</p> $G_0 = \frac{1}{1 + \frac{N}{B+G} + \frac{1}{F} \left( D + N + \frac{DN}{B+G} \right)}$ $G_\infty = \frac{1 + \frac{DN}{B(D+E+N)}}{1 + \frac{N(D+E)}{B(D+E+N)} + \frac{1}{F} \left( \frac{E}{D+E+N} \right) \left( D + N + \frac{DN}{B} \right)}$

TABLE 7. STABILIZING NETWORKS (Continued)

Transfer Function	$T_1$	$T_2$
$\frac{T_1 T_2 s^2 + T_2 s + 1}{T_1 T_2 s^2 + \left[ T_1 \left( 1 + \frac{N}{D} \right) + T_2 \right] s + 1}$	$\frac{DN}{D+N} \text{HRC}$	$A(D+N)RC$
$\frac{T_1 T_2 \left[ 1 + \frac{DN}{B(D+E+N)} \right] s^2 + (T_1 + T_2) s + 1}{T_1 T_2 \left[ 1 + \frac{N(D+E)}{B(D+E+N)} \right] s^2 + \left[ T_1 \left( 1 + \frac{N}{B} \right) + T_2 \right] s + 1}$	$\text{BHRC}$	$A(D+E+N)RC$
$\frac{T_1 T_2 \left[ 1 + \frac{DN}{B(D+E+N)} \right] s^2 + \left\{ T_1 + T_2 \left[ 1 + \frac{DN}{(B+G)(D+E+N)} \right] \right\} s + 1}{T_1 T_2 \left[ 1 + \frac{N(D+E)}{B(D+E+N)} \right] s^2 + \left\{ T_1 \left( 1 + \frac{N}{B} \right) + T_2 \left[ 1 + \frac{N(D+E)}{(B+G)(D+E+N)} \right] \right\} s + \frac{1}{G_0}}$	$\frac{BG}{B+G} \text{HRC}$	$A(D+E+N)RC$
$\frac{T_1 T_2 \left[ 1 + \frac{DN}{B(D+E+N)} \right] s^2 + \left\{ T_1 + T_2 \left[ 1 + \frac{DN}{(B+G)(D+E+N)} \right] \right\} s + 1}{T_1 T_2 \left[ 1 + \frac{N(D+E)}{B(D+E+N)} + \frac{1}{F} \left( \frac{E}{D+E+N} \right) \left( D+N + \frac{DN}{B} \right) \right] s^2 + \left\{ T_1 \left( 1 + \frac{N}{B} \right) + \frac{T_1}{F} \left( D+N + \frac{DN}{B} \right) + T_2 \left[ 1 + \frac{N(D+E)}{(B+G)(D+E+N)} \right] \right\} s + \frac{T_2}{F} \left( \frac{E}{D+E+N} \right) \left( D+N + \frac{DN}{B+G} \right) \right\} s + \frac{1}{G_0}}$	$\frac{BG}{B+G} \text{HRC}$	$A(D+E+N)RC$

TABLE 8. MECHANICAL COMPONENTS, LEAD

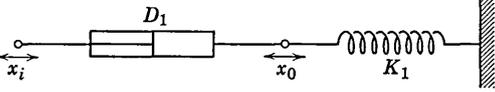
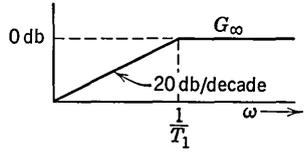
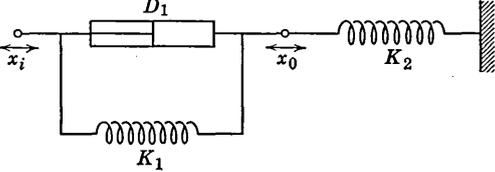
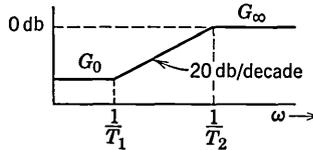
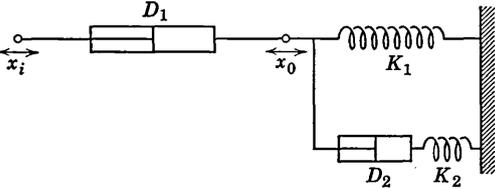
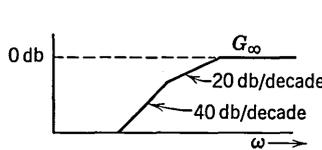
Mechanical Lead Network	Log Magnitude Characteristic	Transfer Function	$T_1$	$T_2$
<p>(a) </p>	<p></p> <p><math>G_0 = 0</math>                      <math>G_\infty = 1</math></p>	$\frac{T_1 s}{T_1 s + 1}$	$\frac{D_1}{K_1}$	<p>...</p>
<p>(b) </p>	<p></p> <p><math>G_0 = \frac{1}{1 + \frac{K_2}{K_1}}</math>                      <math>G_\infty = 1</math></p>	$G_0 \frac{T_1 s + 1}{T_2 s + 1}$	$\frac{D_1}{K_1}$	$G_0 T_1$
<p>(c) </p>	<p></p> <p><math>G_0 = 0</math>                      <math>G_\infty = 1</math></p>	$\frac{T_1 T_2 s^2}{T_1 T_2 s^2 + [T_1 + (1 + \frac{K_2}{K_1}) T_2] s + 1}$	$\frac{D_1}{K_1}$	$\frac{D_2}{K_2}$

TABLE 9. MECHANICAL COMPONENTS, LAG

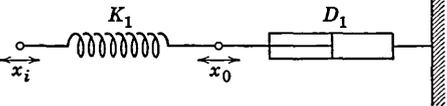
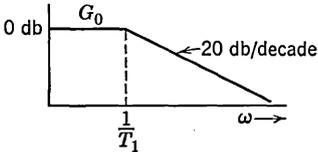
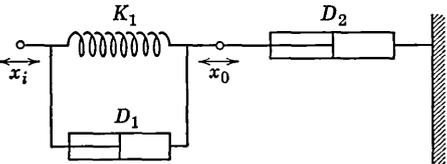
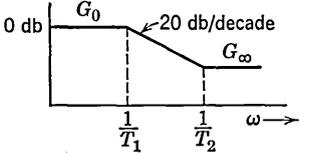
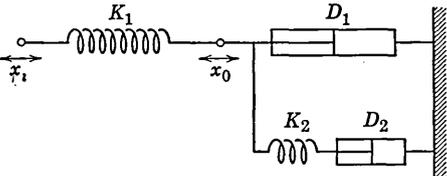
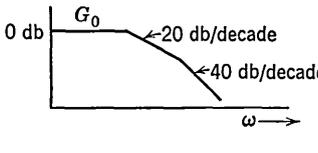
Mechanical Lag Network	Log Magnitude Characteristic	Transfer Function	$T_1$	$T_2$
<p>(a) </p>	<p></p> <p><math>G_0 = 1</math>                      <math>G_\infty = 0</math></p>	$\frac{1}{T_1 s + 1}$	$\frac{D_1}{K_1}$	...
<p>(b) </p>	<p></p> <p><math>G_0 = 1</math>                      <math>G_\infty = \frac{1}{1 + D_2/D_1}</math></p>	$\frac{T_2 s + 1}{T_1 s + 1}$	$\frac{T_2}{G_\infty}$	$\frac{D_1}{K_1}$
<p>(c) </p>	<p></p> <p><math>G_0 = 1</math>                      <math>G_\infty = 0</math></p>	$\frac{1}{T_1 T_2 s^2 + [T_1 + (1 + \frac{K_2}{K_1}) T_2] s + 1}$	$\frac{D_1}{K_1}$	$\frac{D_2}{K_2}$

TABLE 10. MECHANICAL COMPONENTS, LAG-LEAD

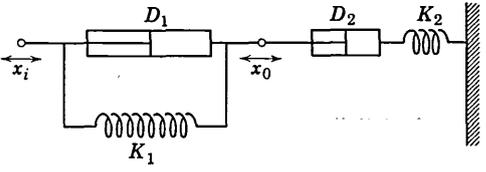
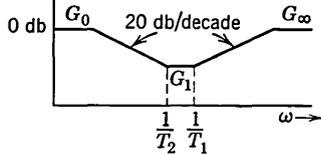
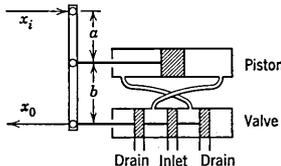
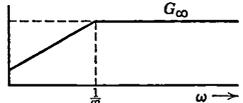
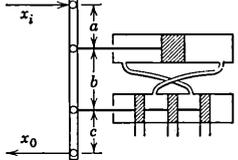
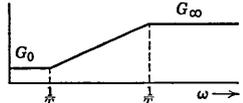
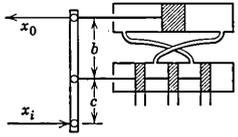
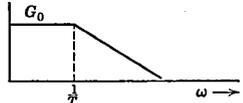
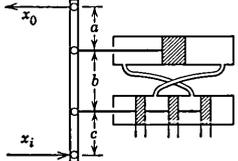
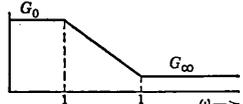
Mechanical Lag-Lead Network	Log Magnitude Characteristic	Transfer Function	$T_1$	$T_2$
	 <p> <math>G_0 = G_\infty = 1</math>    <math>G_1 = \frac{T_1 + T_2}{T_1 + (1 + \frac{K_2}{K_1})T_2}</math> </p>	$\frac{(T_1 s + 1)(T_2 s + 1)}{T_1 T_2 s^2 + [T_1 + (1 + \frac{K_2}{K_1})T_2]s + 1}$	$\frac{D_1}{K_1}$	$\frac{D_2}{K_2}$

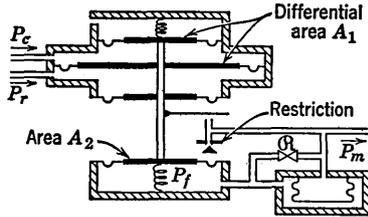
TABLE 11. MECHANICAL-HYDRAULIC COMPONENTS

Mechanical-Hydraulic Network	Log Magnitude Characteristic	Transfer Function	$T_1$	$T_2$
	 <p><math>G_0 = 0</math>      <math>G_\infty = \frac{b}{a}</math></p>	$\frac{G_\infty T_1 s}{T_1 s + 1}$	$\frac{a}{(a+b)K}$	...
	 <p><math>G_0 = \frac{c}{a+b}</math>      <math>G_\infty = \frac{b+c}{a}</math></p>	$\frac{G_0(T_1 s + 1)}{T_2 s + 1}$	$\frac{b+c}{cK}$	$\frac{a}{(a+b)K}$
	 <p><math>G_0 = \frac{b}{c}</math>      <math>G_\infty = 0</math></p>	$\frac{G_0}{T_1 s + 1}$	$\frac{b+c}{cK}$	...
	 <p><math>G_0 = \frac{a+b}{c}</math>      <math>G_\infty = \frac{a}{b+c}</math></p>	$\frac{G_0(T_1 s + 1)}{T_2 s + 1}$	$\frac{b+c}{cK}$	$\frac{a}{(a+b)K}$

$K$  = velocity of piston per unit valve displacement

TABLE 12. PNEUMATIC COMPENSATING COMPONENTS. APPROXIMATE RELATIONSHIPS FOR HIGH LOOP GAIN CONTROLLERS,  $\epsilon \ll 1$

LEAD

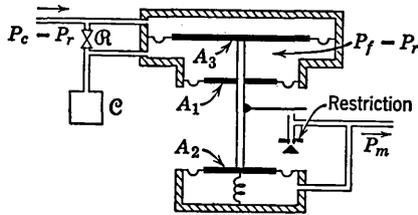
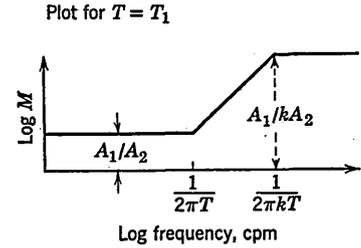


▲ = Pressure source

$$\frac{P_m - P_0}{P_c - P_r} = \frac{A_1}{A_2} \left[ \frac{1 + T_1 s}{1 + k T_1 s} \right]$$

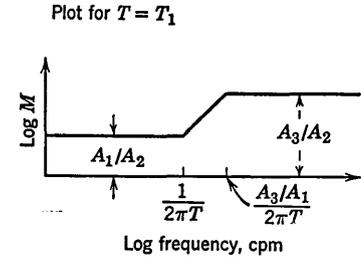
$$T_1 = RC$$

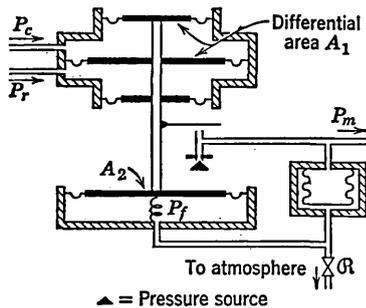
$k$  = change in  $P_f$  for a unit change in  $P_m$  when  $R$  is completely closed.



$$\frac{P_m - P_0}{P_c - P_r} = \frac{A_1}{A_2} \left[ \frac{1 + (A_3/A_2) T_1 s}{1 + T_1 s} \right]$$

$$T_1 = RC$$



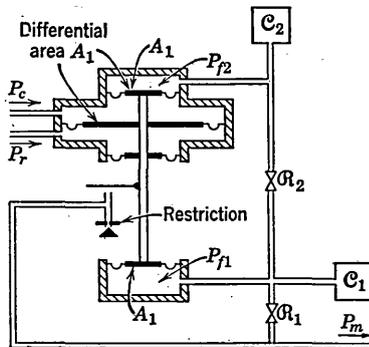
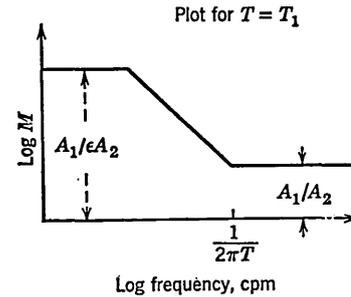


### LAG

$$\frac{P_m - P_0}{P_c - P_r} = \frac{A_1}{A_2 k} \left[ \frac{1 + 1/T_1 s}{1 + \epsilon/k T_1 s} \right]$$

$$T_1 = RC$$

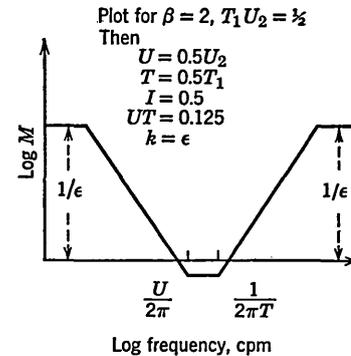
$\epsilon$  = a system constant related to the loop gain.



### LAG-LEAD

$$\frac{P_m - P_0}{P_c - P_r} = (1 + \beta T_1 U_2) \cdot \left[ \frac{U_2/s + 1 + \frac{T_1 s}{1 + \beta T_1 U_2}}{\epsilon U_2/s + \epsilon T_1 s + 1} \right],$$

where  $\epsilon \beta T_1 U_2 \ll 1$ ,  $T_1 = R_1 C_1$ ,  $U_2 = 1/R_2 C_2$ ,  $\beta = 1 + C_2/C_1$ ,  $I = \text{interaction factor}$ .



Reproduced by permission from *Process Instruments and Controls Handbook*, edited by D. M. Considine. Copyright 1957. McGraw-Hill Book Company.

*Symbols in Table 12*

- $P_c$  = controlled variable-pressure signal  
 $P_r$  = referenced input  
 $P_f$  = feedback pressure  
 $P_m$  = output pressure  
 $P_0$  = equilibrium value of  $P_m$  and  $P_f$  when  $P_c = P_r$   
 $A_1, A_3$  = differential areas acted on by pressures as shown in diagram  
 $A_2$  = area of feedback diaphragm  
 $R$  = variable resistance in line  
 $C$  = capacitance associated with  $P_f$

**Charts of Mechanical Networks.** The great majority of feedback control systems use electrical signals in the error channel, but in some cases it is necessary or desirable to use mechanical devices to handle the signals. The signals are most commonly in the form of linear or angular positions. Tables 8 through 10 show a few mechanical networks for obtaining phase lead and phase lag characteristics, for linear motion. Components for rotational motion can be readily derived from these. Oil-filled dashpot elements indicated in these networks have a disadvantage that their characteristics change with extreme variations of temperatures so that these networks may not be applicable in some systems. Several more mechanical networks are given in Ref. 7.

**Charts of Mechanical-Hydraulic Networks.** Effective stabilized networks can also be formed from hydraulic system components. Many hydraulic systems contain one or more stages of hydraulic amplification consisting of control valve and output piston components. When mechanical feedback around these stages is used, lead or lag characteristics can be obtained as indicated schematically in Table 11.

**Chart of Mechanical-Pneumatic Networks.** Pneumatic systems allow considerable flexibility in generating control functions. Table 12 shows several typical circuits. Additional circuits are given in Ref. 8.

### 3. COMPENSATING NETWORKS: A-C SYSTEMS

Many of the components that are commonly used are alternating-current devices. Selsyns are accurate and reliable error measuring devices; a-c amplifiers are relatively simple and drift-free devices compared with their d-c counterparts; a-c tachometers have reached a high state of development for small, low-power servos; and the two-phase a-c motor is a popular drive for low-power, fast servos. Much analysis using a modulated a-c carrier is the same as that used in d-c systems, but there are additional design problems as described in the following sections.

**Requirements for Ideal Series Networks.** Frequently a-c servo-mechanisms are required to operate at frequencies above that indicated

by the motor time constant. In such systems a rate producing component is required to stabilize the control system. For a series stabilizing component, this usually takes the form of an electric lead network.

The special problem of an a-c rate network is that the network must take the derivative of the envelope of the carrier without shifting the phase of the carrier. Consider a carrier of time function  $\cos \omega_c t$  modulated by the signal  $m \cos a\omega_c t$ , where  $a$  is a fraction much less than 1. (The useful control frequencies in an a-c feedback control system must be quite low relative to the carrier frequency.) If this signal is operated on by a lead network of transfer function  $1 + Ts$ , where  $T$  is the lead time constant, the required output signal is

$$m[\cos a\omega_c t \cos \omega_c t - T a \omega_c \sin a\omega_c t \cos \omega_c t]$$

$$= m \cos \omega_c t [\cos a\omega_c t + a\omega_c T \cos (a\omega_c t + 90^\circ)].$$

Figure 24 is a vector diagram of these signals, showing the phase lead produced in the modulating signal. Note that the phase of the carrier is

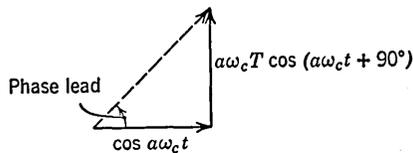


FIG. 24. Phase lead produced by ideal lead network on modulating signal in all a-c servo system.

not changed although the phase of the modulating signal is. The equation can be changed to the equivalent expression

$$\frac{m}{2} \{ [\cos (1 + a)\omega_c t + \cos (1 - a)\omega_c t]$$

$$+ a\omega_c T [\cos ((1 + a)\omega_c t + 90^\circ) \cos ((1 - a)\omega_c t - 90^\circ)] \}.$$

This equation indicates that the modulated carrier signal is actually composed of two frequencies somewhat displaced about the carrier. Furthermore, it indicates the necessary frequency response of a network to obtain an effective lead characteristic at the modulation frequency,  $a\omega_c$ . Figure 25 shows the magnitude and phase characteristic indicated by this equation for several values of  $T$ .

**Practical Networks.** The ideal frequency characteristic shown in Fig. 25 cannot be attained with physical networks. However, the frequency response can be approximated over part of the frequency rang by several practical networks. The most common of these are the bridged T and parallel-T networks. (See Table 7(a).) The bridged-T networks

have characteristics similar to the ideal network of Fig. 25 whereas the parallel-T has the rate portion only. Addition of a parallel channel to the parallel-T network makes it possible to approximate the lead characteristics of Fig. 25.

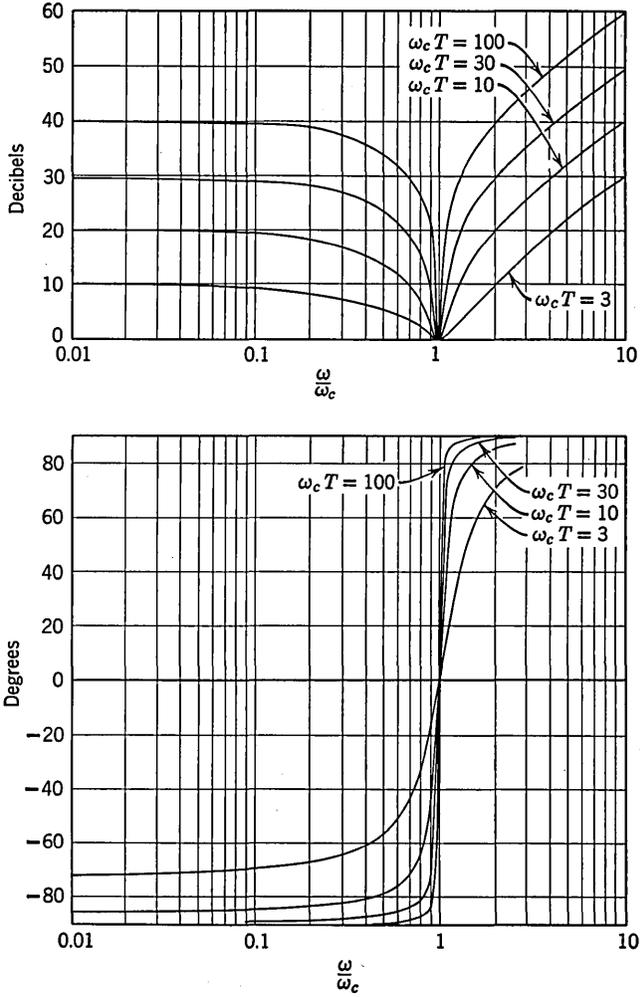


Fig. 25. Characteristics of ideal lead networks for use in a-c feedback control systems.

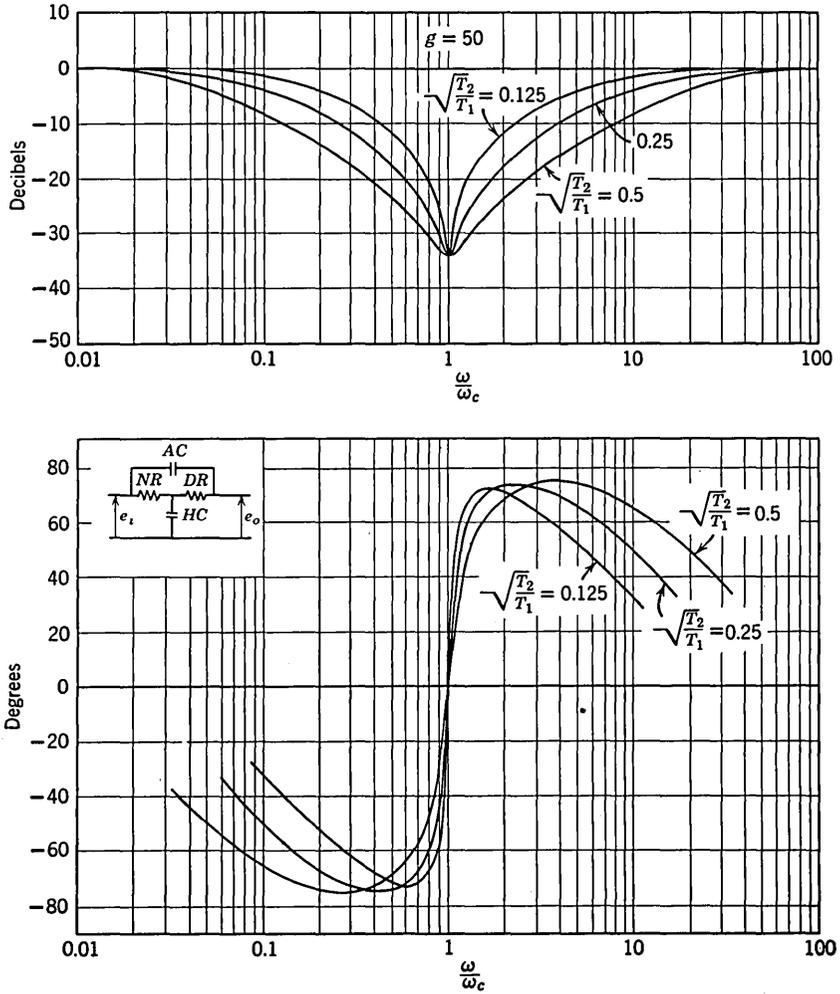


FIG. 26. Characteristics of the bridged-T network.

$$\frac{e_0}{e_i} = \frac{(s^2/\omega_c^2) + \sqrt{T_2/T_1}(s/\omega_c) + 1}{(s^2/\omega_c^2) + g\sqrt{T_2/T_1}(s/\omega_c) + 1}$$

$$\omega_c^2 = \frac{1}{T_1 T_2}, \quad T_1 = \frac{DN}{D+N} HRC, \quad T_2 = A(D+N)RC,$$

$$g = \left(1 + \frac{N}{D}\right) \frac{T_1}{T_2} + 1.$$

As indicated in Table 7(a) the transfer function of the bridged-T network is

$$\frac{e_0}{e_i} = \frac{T_1 T_2 s^2 + T_2 s + 1}{T_1 T_2 s^2 + \{[1 + (N/D)]T_1 + T_2\}s + 1}$$

where

$$T_1 = \left( \frac{DN}{D + N} \right) HRC,$$

$$T_2 = A(D + N)RC.$$

The circuit is adjusted until  $T_1 T_2 = 1/\omega_c^2$ , where  $\omega_c$  is the carrier frequency. The minimum value of the transfer function occurs at  $\omega = \omega_c$ , where the value is

$$\frac{1}{g} = \frac{1}{1 + [1 + (N/D)](T_1/T_2)}.$$

The above equation can be rewritten

$$\frac{e_0}{e_i} = \frac{(s^2/\omega_c^2) + \sqrt{T_2/T_1}(s/\omega_c) + 1}{(s^2/\omega_c^2) + \sqrt{T_2/T_1}g(s/\omega_c) + 1}.$$

Characteristics of this network for several values of  $T_2/T_1$  are shown in Fig. 26. It is seen that this can approximate the ideal characteristic of Fig. 25. The factor  $T_2/T_1$  largely influences the rate of change of angle near  $\omega_c$  whereas  $g$  determines the magnitude of phase change that can be obtained.

Characteristics of the parallel-T network can be found in Ref. 5.

**Sensitivity to Carrier Frequency Shift.** The most serious weakness of a-c feedback control systems using series networks is that the system performance is impaired by normal shifts in the frequency of the carrier. For *example*, in an aircraft power system, the 400-cps power may be frequency regulated to only 5 or 10 per cent. In many a-c feedback control systems using series stabilizing networks, this amount of frequency shift will render the system completely useless.

Figure 27 shows the effects of a carrier frequency shift on the operation of an a-c stabilizing network. These are:

1. The gain of the control systems at low frequencies is increased. This may result in saturation in subsequent elements in the control system.

2. The phase of the carrier is shifted. This means that any phase sensitive devices such as discriminators or a-c motors will not operate at best efficiency.

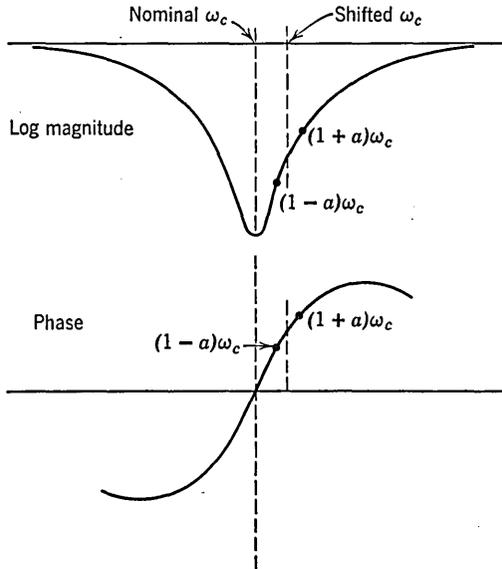


FIG. 27. Effect of carrier frequency shift on the operation of an a-c series stabilizing network.

3. The phase lead at the control system frequencies ( $a\omega_c$ ) is decreased. Thus the network does not perform the function for which it was intended and instability may result.

It is noticed that "fast" control systems, which are designed to have a high crossover frequency (large values of  $a$ ) are less susceptible to carrier frequency shifts than systems which have a large effective lead network time constant.

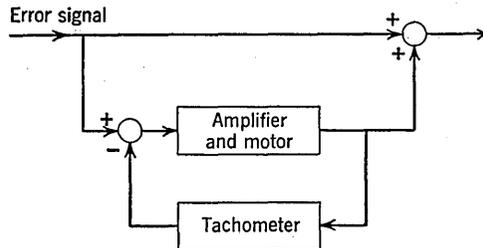
**Tachometer Stabilization.** Alternating-current tachometers are excited by the carrier frequency along with other components in an a-c control system and generate an amplitude-modulated signal proportional to velocity. This operation is not hindered by reasonable changes of the carrier frequency so that this method of stabilization is not subject to the limitations of a-c stabilizing networks. The analysis or synthesis of a-c systems using tachometers can proceed just as in the case of the d-c system described in Sect. 2 under Rate Feedback.

**Other Techniques.** The previous sections have discussed means of stabilizing all a-c feedback control systems by using rate-producing components. Often it is necessary to add gain at low frequencies and reduce this gain below the desired crossover frequency by means of an integrating or reset component. This is the type of system (using d-c signals) described in Sect. 2, Phase Lag Compensation.

Theoretically, an a-c carrier lag network can be constructed by using a bridged-T circuit in the feedback channel of a feedback amplifier. However, since the time constant of a lag network has to be considerably larger than used in a lead network, the effect of carrier frequency shifts usually makes this method impractical.

Commonly lag networks are obtained by rectifying the carrier to direct current and then using d-c networks. This procedure then is the same as that of Sect. 2 and the system is no longer an all a-c system.

Another method to obtain an effective lag network is to use a small "reset" servo in parallel with the signal channel, as shown in Fig. 28.



$$\text{Reset servo transfer function} = \frac{\frac{K}{s(1 + T_m s)}}{1 + \frac{K}{(1 + T_m s)}} = \frac{K}{s[(1 + K) + T_m s]}$$

Fig. 28. Reset servo channel in parallel with signal channel.

The servo channel has high gain at low frequencies but, because of the tachometer feedback, this gain falls below that of the regular signal channel below the crossover frequency.

#### 4. OPEN-CLOSED LOOP CONTROL

*Open-closed loop control*, sometimes called *schedule and trim*, is not so much a different kind of control as it is a different way of visualizing or synthesizing a control system. The principle is that an open loop control system, although not accurate enough for the complete control, responds predictably and stably to an input signal, and it can be used as an approximate, or first order correction, control system. Then the required accuracy can be obtained as a correction or "trim" to the open loop and is accomplished by the use of a relatively slow but high gain feedback loop. This is illustrated in the block diagram of Fig. 29, which is modified for analysis purposes in Fig. 30. The only closed loop to consider is the easily stabilized, low crossover frequency loop and that the required high-frequency response is obtained by the open-ended forcing function. In

the actual system this forcing action is attained by the parallel signal channel to the power element.

An allied situation exists when there is difficulty in measuring the control system output accurately and immediately. The trouble may be in an

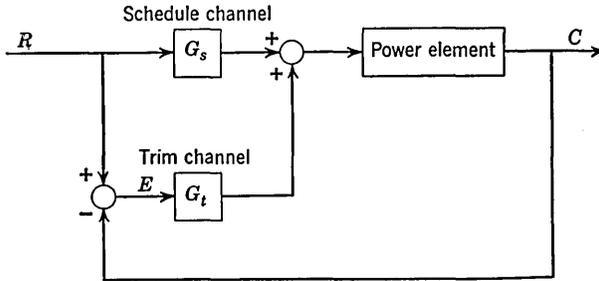


FIG. 29. Block diagram of open-closed loop control.

inherent delay in the measuring device, such as the time lag of a thermocouple measuring temperature, or it may be caused by the need for a smoothing and averaging process to attain accuracy from noisy data. In such cases, the delayed but accurate measurement can be used in a trim feedback control loop and then an internal, fast response, feedback loop is formed by using an alternate measurement. This alternate quantity is

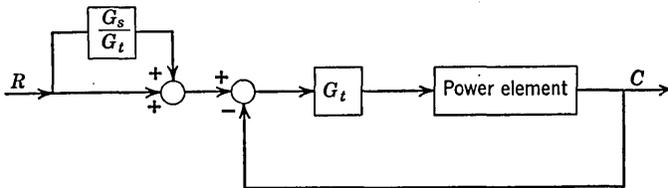


FIG. 30. Modification of Fig. 29 for analysis.

related to the desired output in a known way, such as a pressure change which accompanies a change in temperature, but accuracy of such a relationship is not high enough to use the alternate quantity as the ultimate measurement of the desired output.

Open-closed loop control is covered in detail in Ref. 9.

ACKNOWLEDGMENT

Tables 3 to 7 are reproduced with permission from H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. II, Wiley, New York, 1955.

REFERENCES

1. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. I, Wiley, New York, 1951.
2. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. II, Wiley, New York, 1955.
3. J. G. Truxal, *Automatic Feedback Control System Synthesis*, McGraw-Hill, New York, 1955.
4. Paul E. Smith, Jr., Design regulating systems by error coefficients, *Control Eng.*, **2**, 69-75 (1955).
5. Leonard Stanton, Theory and application of parallel-T resistance-capacitance frequency-selective networks, *Proc. I.R.E.*, **34**, 447-456 (1946).
6. W. R. Evans, *Control System Dynamics*, McGraw-Hill, New York, 1954.
7. J. E. Gibson, 14 ways to generate control functions mechanically, *Control Eng.*, **2**, 65-69 (1955).
8. D. M. Considine, Editor, *Process Instruments and Controls Handbook*, McGraw-Hill, New York, 1957.
9. John R. Moore, Combination open-cycle, closed-cycle control systems, *Proc. I.R.E.*, **39**, 1421-1432 (1951).

## Noise, Random Inputs, and Extraneous Signals

*D. L. Lippitt*

1. Introduction	24-01
2. Mathematical Description of Noise	24-02
3. Measurement of Noise	24-06
4. System Response to Noise	24-11
5. System Design in the Presence of Noise	24-15
References	24-19

### 1. INTRODUCTION

Linear systems can be designed to obtain a desired response to commands and disturbances which may be exactly defined either by an equation or by a graphical plot (Chaps. 19 through 23). In many cases inputs can be described adequately only in a statistical manner. *Examples* are the jitter observed in automatic radar tracking systems and gust disturbances to an aircraft. This chapter covers methods for:

- (a) Measuring and describing statistical inputs.
- (b) Computing the system response to such inputs.
- (c) Specifying optimum designs.

## 2. MATHEMATICAL DESCRIPTION OF NOISE

Random processes are described in Chap. 12, Sect. 16, and Chap. 13, Sect. 2. It is sufficient to note here that a *random process* has a complete set of probability distribution functions. If these distributions are independent of time, the process is *stationary* and its characteristics can be defined by time averages (Ref. 1).

**Autocorrelation.** The most useful description of a random process for control system analysis is the *autocorrelation function*  $\phi$  defined by eq. (1) for a stationary function of time  $x(t)$ :

$$(1) \quad \phi_{xx}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^{+T} x(t)x(t + \tau) dt.$$

Figure 1 graphically illustrates eq. (1). For nonstationary processes the autocorrelation may be described by an ensemble average that is a function

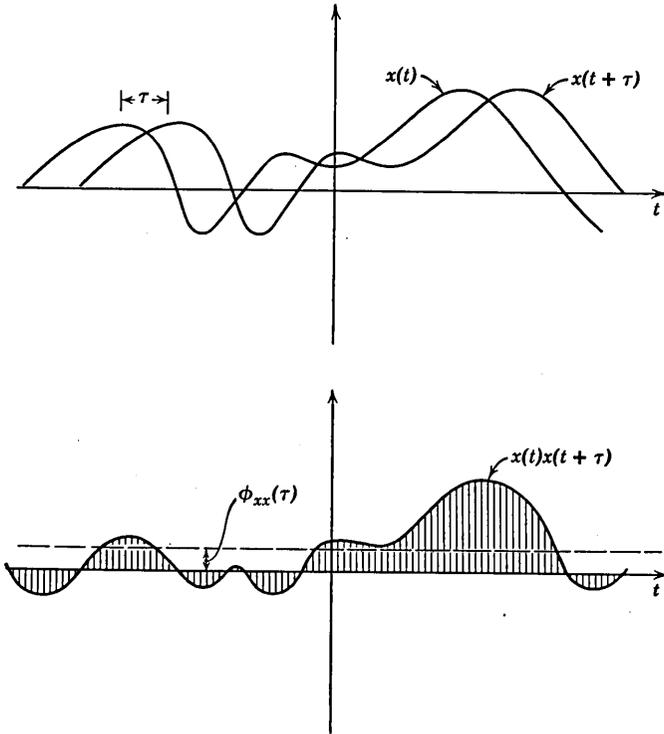


FIG. 1. Illustration of the computation of the autocorrelation function.

of time as well as  $\tau$ . This definition is given in eq. (2):

$$(2) \quad \phi_{xx}(t, \tau) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x_1 x_2 P(t, x_1, t + \tau, x_2) dx_1, dx_2,$$

where  $P(t, x_1, t + \tau, x_2)$  is the joint probability density of  $x_1$  at time  $t$  and  $x_2$  at time  $(t + \tau)$ . *Important properties of the autocorrelation function for stationary series are:*

$$(3) \quad \phi_{xx}(\tau) \leq \phi_{xx}(0),$$

$$(4) \quad \phi_{xx}(\tau) = \phi_{xx}(-\tau),$$

$$(5) \quad \phi_{xx}(0) = \overline{x^2}.$$

In eq. (5), the bar indicates a time average.

An interesting *example* is the autocorrelation of the function  $\sin \omega t$ .

$$x(t) = A \sin \omega t.$$

$$(6) \quad \begin{aligned} \phi_{xx}(\tau) &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^{+T} A^2 \sin \omega t \sin \omega(t + \tau) dt \\ &= A^2/2 \cos \omega \tau. \end{aligned}$$

Although  $\sin \omega t$  is not strictly stationary, this *example* illustrates the effect of a pronounced periodicity in noise data. If it exists, it will show up in the autocorrelation as cosine function.

**Cross-Correlation.** In some cases a control system will have two inputs,  $x$  and  $y$ , which are not completely independent. The relationship is expressed by the *cross-correlation function* defined by eq. (7) for stationary series:

$$(7) \quad \phi_{xy}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^{+T} x(t)y(t + \tau) dt.$$

For nonstationary series  $\phi_{xy}$  must be expressed as an ensemble average as given by eq. (8):

$$(8) \quad \phi_{xy}(t, \tau) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyP(x, t, y, t + \tau) dx dy.$$

*Important properties of the cross-correlation function for stationary series are:*

$$(9) \quad [\phi_{xy}(\tau)]_{\max} < \overline{x^2} \quad \text{or} \quad \overline{y^2} \quad (\text{whichever is larger}),$$

$$(10) \quad \phi_{xy}(\tau) = \phi_{yx}(-\tau).$$

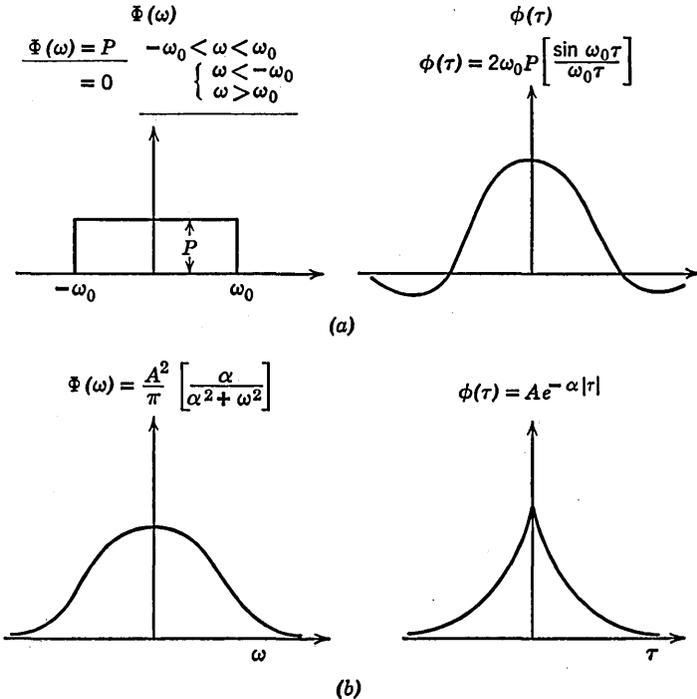


FIG. 2. (a) and (b) Examples of autocorrelation and spectral density pairs.

The autocorrelation of the sum of two correlated functions is given by eq. (11):

$$(11) \quad \phi_{x+y}(\tau) = \phi_{xx}(\tau) + \phi_{yy}(\tau) + \phi_{xy}(\tau) + \phi_{yx}(\tau).$$

If  $x(t)$  and  $y(t)$  are independent:

(a) The cross-correlations become constants equal to the products of their means or  $\bar{x}\bar{y}$ .

(b) The autocorrelation of the sum becomes the sum of the individual autocorrelations plus twice the product of the means, or  $\phi_{xx}(\tau) + \phi_{yy}(\tau) + 2\bar{x}\bar{y}$ .

**Spectral Density.** An alternate description of a *stationary random process* is the spectral density,  $\Phi(\omega)$ . It is a measure of the distribution of energy in the frequency spectrum. For a voltage wave the units would be volts<sup>2</sup> per radian per second.

The following discussion is not rigorous but will show the *physical significance* of  $\Phi(\omega)$ . Assume that several samples of noise of duration  $T$

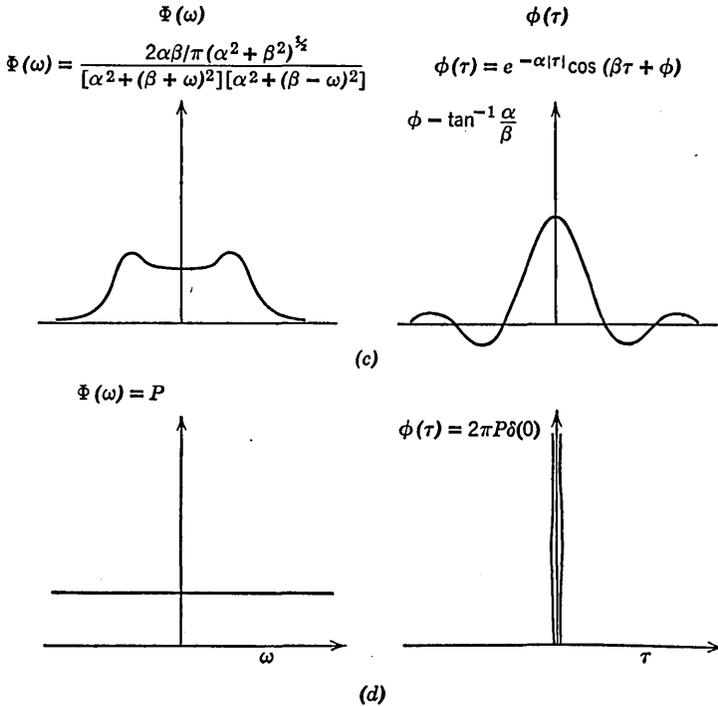


FIG. 2. (c) and (d) Examples of autocorrelation and spectral density pairs.

seconds have been expanded in a Fourier series of the form shown in eq. (12):

$$(12) \quad x(t) = \frac{c_0}{2} + \sum_{n=1}^{\infty} c_n \cos(\omega_n t + \psi_n),$$

where  $\omega_n = 2n\pi/T$ ,

$$a_n = \frac{2}{T} \int_0^T x(t) \cos \omega_n t dt,$$

$$b_n = \frac{2}{T} \int_0^T x(t) \sin \omega_n t dt,$$

$$c_n^2 = a_n^2 + b_n^2; \psi_n = \tan^{-1}(b_n/a_n).$$

Assume that  $T$  is very long compared to the longest periodicity present in the function. If the  $c$ 's and  $\psi$ 's for a given value of  $n$  are considered over a large number of samples, it will be found that the  $\psi$ 's are uniformly distributed between  $+\pi$  and  $-\pi$  and that the  $c_n^2$ 's have an average value  $\overline{c_n^2}$ . A knowledge of  $\overline{c_n^2}$ 's for a given process is sufficient to predict the

output of a linear control system with a transfer function  $G(j\omega)$  between the noise input and the system output. The mean square of the output is

$$(13) \quad \sigma_c^2 = |G(0)|^2 \frac{c_0^2}{2} + \sum_{n=1}^{\infty} \overline{c_n^2} |G(\omega_n)|^2.$$

The experimental determination of the  $\overline{c_n^2}$ 's would be relatively inefficient. However, the  $\overline{c_n^2}$ 's are related to the spectral density by eq. (14) for large  $T$ .

$$(14) \quad \overline{c_n^2} = 2\Phi(\omega_n)\Delta\omega,$$

where

$$\omega_n = n\Delta\omega = 2\pi n/T.$$

Hence if the average value of the input is zero,  $c_0$  is zero and eq. (13) becomes eq. (15):

$$(15) \quad \sigma_c^2 = 2 \int_0^{\infty} \Phi(\omega) |G(\omega)|^2 d\omega.$$

The spectral density is related to the autocorrelation function by the Fourier cosine transform as shown in eqs. (16) and (17):

$$(16) \quad \Phi(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \phi(\tau) \cos \omega\tau d\tau,$$

$$(17) \quad \phi(\tau) = \int_{-\infty}^{+\infty} \Phi(\omega) \cos \omega\tau d\omega.$$

The cross-spectral density  $\Phi_{xy}(\omega)$  bears the same transform relationship to the cross-correlation function  $\phi_{xy}(\tau)$  as the spectral density does to the autocorrelation function.

Figure 2 shows several pairs of spectral density functions and autocorrelation functions.

### 3. MEASUREMENT OF NOISE

The greatest problem involved in the analysis of the response of a control system to a random input is obtaining the required characteristics of the input. If the input is stationary, long samples are necessary and numerous calculations must be made. In most cases the use of high-speed digital computers is necessary if the job is to be completed in reasonable time. If the input is nonstationary, the magnitude is multiplied many times since the results must be calculated separately for each value of time.

**Calculation of  $\phi(\tau)$  for Stationary Inputs.** The most straightforward methods of analysis if the input is stationary are to compute the auto-

correlation function defined by eq. (1). The approximate form for calculation is given by eq. (18):

$$(18) \quad \phi(m \Delta\tau) = \frac{1}{N - m + 1} \sum_{n=0}^{N-m} x_n x_{n+m},$$

where  $\Delta\tau$  is the time interval at which values of the function are read and  $x_n$  is the value of the function  $n \Delta\tau$  seconds from the beginning of the sample.

**Sampling Rate.** The value of  $\Delta\tau$  is set by *Shannon's sampling theorem*

$$(19) \quad \Delta\tau = \frac{1}{2f_s}.$$

The value of  $f_s$  is the highest frequency present in the data. In general, this will not be known. In most control systems there are considerations other than noise which set an upper bound on the system band pass so that a filter may be inserted in the device which records the sample to eliminate frequencies not of interest. This is desirable since by increasing  $\Delta\tau$  the number of calculations is reduced.

**Required Range of  $\tau$ .** The maximum value of  $\tau$  is determined by the use to which  $\phi(\tau)$  is to be put. If it is desired to compute the variance of a system output, reference to eq. (29) will show that  $m_{\max} \Delta\tau$  should equal the longest anticipated settling time of the system output to an impulse applied at the noise input.

An insight to the effect of using a finite value,  $m_{\max}$ , can be obtained by performing the integration of eq. (16) over a finite range or by multiplying the true autocorrelation function by a function  $u(\tau)$  which equals unity for  $-T < \tau < +T$  and zero elsewhere. Then the approximate spectrum is given by eq. (20):

$$(20) \quad \Phi_{\text{approx}}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} u(\tau)\phi(\tau) \cos \omega\tau \, d\tau.$$

Since the true spectrum is the Fourier transform of  $\phi(\tau)$ , and  $(T/\pi) (\sin \omega T/\omega T)$  is the transform of  $u(\tau)$ , the approximate spectrum is given by eq. (21):

$$(21) \quad \Phi_{\text{approx}}(\omega) = \frac{T}{\pi} \int_{-\infty}^{+\infty} \Phi(\omega - \alpha) \frac{\sin \alpha T}{\alpha T} \, d\alpha.$$

For *example*, if the time function were a pure sine wave of frequency  $\omega_0$  of unity power, the true spectrum would be given by eq. (22):

$$(22) \quad \Phi(\omega) = \frac{1}{2}[\delta(\omega_0) + \delta(-\omega_0)],$$

where  $\delta(\omega) =$  impulse at  $\omega$ .

Then if the autocorrelation were computed only for  $(-T < \tau < +T)$ , the approximate spectrum computed from the result would be given by eq. (23):

$$(23) \quad \Phi_{\text{approx}}(\omega) = \frac{T}{2\pi} \left[ \frac{\sin(\omega_0 - \omega)T}{(\omega_0 - \omega)T} + \frac{\sin(\omega_0 + \omega)T}{(\omega_0 + \omega)T} \right].$$

Figure 3 shows a plot of these results.

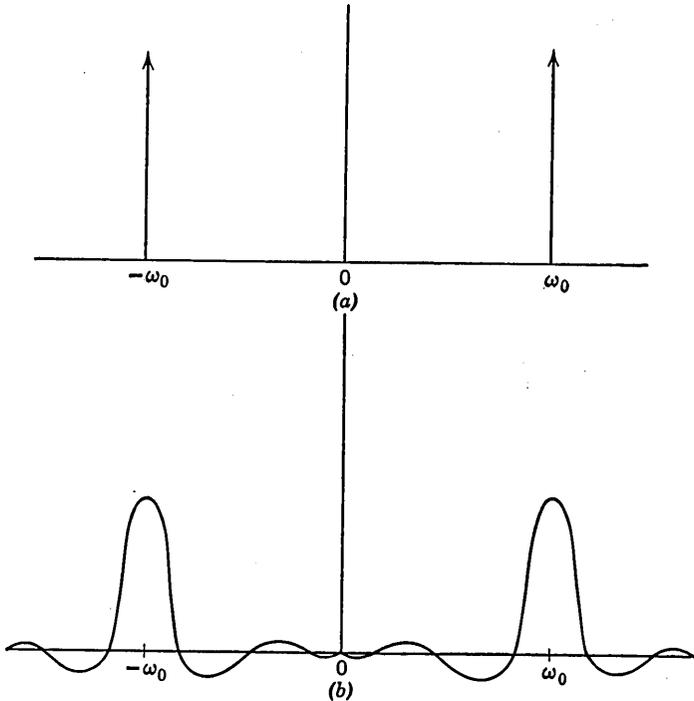


FIG. 3. Effect of limiting the range of  $\tau$  on computation of the spectral density: (a) exact spectrum; (b) approximate spectrum.

From this *example*, it can be seen that the usable range of  $(\tau)$  limits the resolutions of the measurement. Also, a limited range of  $(\tau)$  may lead to negative values of  $\Phi(\omega)$  which are physically impossible.

**Required Sample Length.** The required sample length for computing  $\phi(\tau)$  depends on the use to which the result will be put and to a certain extent upon the frequencies contained in the input function. *A useful rule of thumb is that  $N$  in eq. (18) should be at least 10 times the maximum value of  $m$ .* As a check, the autocorrelation can be computed for two samples of equal length. If then the results are nearly equal, the samples are

probably long enough. If not, the average of the two should be compared with results from a sample twice as long and so forth until agreement is reached.

Once an approximate knowledge of the frequency components of the noise has been obtained, a better estimate can be made of the required sample length. Reference to eq. (28) shows that computing the system output from the autocorrelation function of the input is equivalent to computing the output directly by convolution and averaging the square. Hence, if a sample of the system output  $T$  seconds long is sufficient to give an accurate measure of the output mean square, a sample  $T + T_s$  seconds long, where  $T_s$  is the system settling time, is sufficiently long to compute the autocorrelation of the input. If  $\phi_0(\tau)$  is the estimated output correlation, the ratio of standard deviation of the computed output mean squares  $\sigma^2_{[\sigma_0^2]}$  for samples  $T$  seconds long to the true mean square is given by eq. (24):

$$(24) \quad \frac{\sigma^2_{[\sigma_0^2]}}{\sigma_0^2} = \left[ \frac{4}{T^2 \phi^2(0)} \int_0^T (T - \tau) \phi_0^2(\tau) d\tau \right]^{1/2} .$$

Figure 4 shows a plot of this ratio where  $\phi_0(\tau)$  is  $(e^{-\alpha|\tau|})$  as a function of  $(\alpha T)$ . References 2, 3, 4 give a more detailed consideration to the problem.

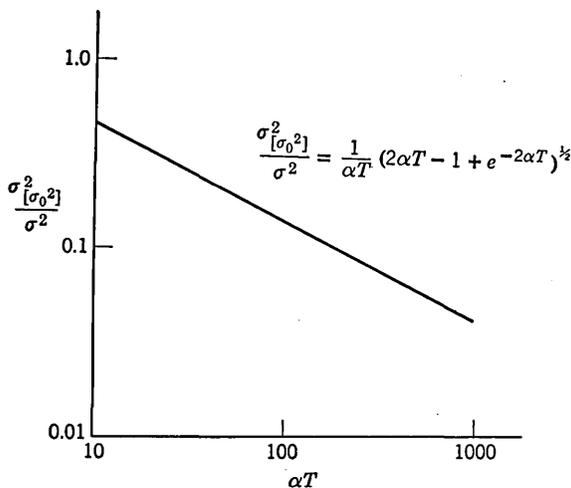


FIG. 4. Standard deviation of errors in computing the output mean square error from a finite length sample.

**Nonstationary Inputs.** Any process observed in nature is nonstationary in the strict sense of the term. However, in many cases the input characteristics will vary so slowly that samples long enough to compute

$\phi(\tau)$  can be considered stationary. In this case the techniques previously discussed are applicable.

A slightly more difficult problem exists when the input *changes too rapidly* to obtain a sufficiently long sample, but it still does not change appreciably during the settling time of the control system. In this case, several recordings of the input must be obtained over the range of characteristics of interest. Then short samples can be drawn from common points and the autocorrelation averaged. The amount of computation required is vastly increased.

One type of *slowly changing* nonstationary function can be treated in a more simple manner. If the frequency components retain the same amplitudes relative to each other but where absolute magnitude increases or decreases with time, a single recording can be used. The input is divided into several short samples. The autocorrelations of each sample are normalized, so that the function is divided by  $\phi(0)$ , and the normalized functions are averaged. The autocorrelation for any specific time is then the averaged normalized autocorrelation function times the mean square value of the input corresponding to that time. Note that this technique is helpful only if it is known that the spectrums have the same form. Otherwise, more data would have to be taken to establish the point.

If the input characteristics *vary appreciably* during a settling time of system, the problem becomes immensely complicated. To compute  $\phi(\tau, t_1)$  it is necessary to average the products from many recordings of the input. *At least one hundred products would be necessary to obtain 10 per cent accuracy when the output of a control system is calculated.*

**Correlation Computers.** Special computers for the computation of correlation functions can be built where many correlations must be done and high-speed digital computers are not available. The basic principle is illustrated in Fig. 5. The noise is recorded on a media such as magnetic

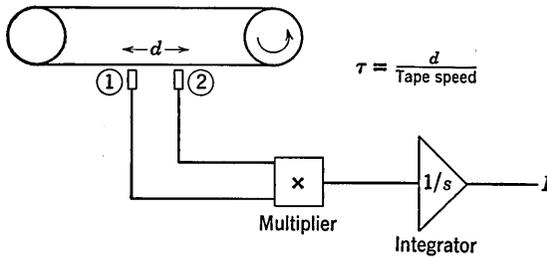


FIG. 5. Correlation computer.

tape and played back through two reading heads spaced a distance  $d$  apart. The second reading head receives the same signal as the first except

that it is delayed by a time equal to  $d$  divided by the tape speed. The two outputs are then multiplied and the result integrated. The output of the integrator is then given by eq. (25):

$$(25) \quad I = \int f(t)f(t - \tau) dt.$$

$I$  may then be divided by the time interval over which the run is made to get the autocorrelation. Other possible recording media are photographic film and ink recordings tracked by hand (Ref. 5). Still other methods use pulse sampling and storage.

Computers of this type depend on the availability of the noise in the right form. For instance, radar tracking data are usually taken with a moving picture camera so that the data must be read frame by frame before they are useful. Also, several records may have to be combined to arrive at the noise. In such cases, it would be simpler to use a general purpose digital computer.

The *accuracy* of such systems is limited by the recording mechanism, multiplier, and integrator. To obtain reasonable accuracy these units become bulky and expensive.

#### 4. SYSTEM RESPONSE TO NOISE

**Time Domain Methods.** The response of a linear dynamic system to an input  $x(t)$  is given by the integral in eq. (26):

$$(26) \quad c(t) = \int_0^{\infty} x(t - \tau)g(\tau) d\tau.$$

Substituting this eq. (26) in eq. (1) gives the autocorrelation function of the output if the input is stationary:

$$(27) \quad \begin{aligned} \phi_{cc}(\tau) &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^{+T} \int_0^{\infty} g(s) \int_0^{\infty} g(r)x(t + \tau - r)x(t - s) dt ds dr \\ &= \int_0^{\infty} g(s) \int_0^{\infty} g(r)\phi_{xx}(\tau + s - r) ds dr. \end{aligned}$$

The mean square value of the output is obtained by setting  $\tau = 0$  in eq. (27):

$$(28) \quad \overline{c^2(t)} = \int_0^{\infty} g(s) \int_0^{\infty} g(r)\phi_{xx}(s - r) ds dr.$$

By an appropriate change of variables the alternate form of eq. (28) is given by eqs. (29) and (30):

$$(29) \quad \overline{c^2(t)} = \int_{-\infty}^{+\infty} \phi_{xx}(\tau) \phi_{gg}(\tau) d\tau,$$

$$(30) \quad \phi_{gg}(\tau) = \int_{-\infty}^{\infty} g(t)g(t + \tau) dt.$$

Strictly speaking the transition from eq. (28) to eq. (29) is possible only if the input  $x(t)$  is stationary. However, if the characteristics of  $x(t)$  vary only slightly during one settling time of the control system and if  $\phi(\tau)$  has been computed from a sample which is long compared to one settling time, eq. (28) is approximately true for nonstationary inputs.

For the more general case of a linear time-varying system with a nonstationary input, the mean square output at time ( $t$ ) is given by eq. (31):

$$(31) \quad \widetilde{\overline{c^2(t)}} = \int_0^{\infty} g(t, s) \int_0^{\infty} g(t, r) \phi(t - r, r - s) ds dr.$$

The autocorrelation function is defined in this case by eq. (2). The function  $g(t, s)$  is defined as the effect on the system output at time  $t$  of an impulse applied at time ( $t - s$ ). The wavy line over  $c^2(t)$  in the equation indicates an ensemble average rather than a time average.

**Frequency Domain Methods.** For cases where the spectral density is known, the mean square of the output is given by eq. (32):

$$(32) \quad \overline{c^2(t)} = 2 \int_0^{\infty} \Phi(\omega) |G(j\omega)|^2 d\omega.$$

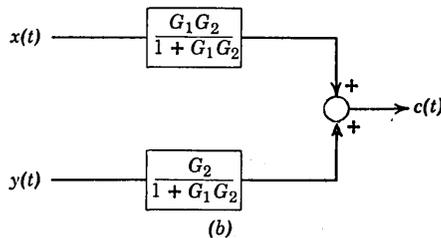
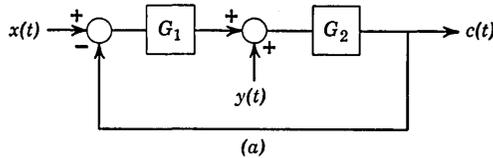


FIG. 6. Control system with two inputs: (a) actual circuit; (b) equivalent circuit.

A more general case is shown in Fig. 6 where there are two inputs to a control system and the inputs may be correlated. In terms of the equivalent circuit, the mean square output is given by eq. (33):

$$(33) \quad \overline{c^2(t)} = 2 \int_0^\infty [ |G_x(j\omega)|^2 \Phi_{xx}(\omega) + |G_y(j\omega)|^2 \Phi_{yy}(\omega) + G_x^*(j\omega) G_y(j\omega) \Phi_{xy}(\omega) + G_x(j\omega) G_y^*(j\omega) \Phi_{yx}(\omega) ] d\omega.$$

The starred transfer functions are complex conjugates.

**Computer Methods.** A modification of eq. (28) leads to an analog computer method for computing noise output. If the input noise has a constant spectral density  $K$ , the autocorrelation function becomes a delta function at  $\tau = 0$  with strength  $2\pi K$ . Then the system output is given by eq. (34):

$$(34) \quad \overline{c^2(t)} = 2\pi K \int_0^\infty g^2(\tau) d\tau.$$

Generating the impulse response by analog techniques, squaring it, and integrating the result give the mean square output. If the input noise

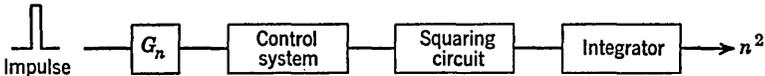


FIG. 7. Computer simulation to compute the mean square output for a correlated noise input.

does not have a constant spectral density, the output can be computed from the system shown in Fig. 7. The filter transfer function is specified by eq. (35):

$$(35) \quad G_n(j\omega) G_n^*(j\omega) = \Phi_n(\omega).$$

The same technique can be used for linear time-varying systems, and for nonstationary random inputs if the inputs are equivalent to stationary noise passing through a linear time-varying filter. For this case eq. (34) becomes eq. (36):

$$(36) \quad \widetilde{\overline{c^2(t)}} = 2\pi K \int_0^\infty g^2(t, \tau) d\tau,$$

where  $g(t, \tau)$  has the definition given following eq. (31). The function  $g(t, \tau)$  with  $\tau$  as the variable can be generated from the adjoint of the control system and a shaping filter (Ref. 6). The following is quoted from Ref. 7. "The adjoint is found from the analog of the original by:

1. Turning each element in the loop around and reversing the direction of signal flow.

2. Letting the variation of time-varying element start from some time  $t_1$  and run backward relative to the action in the original system.

3. Interchange the input and output of the system. The new input is  $\delta(t - t_1)$ ."

The output is then  $g(t, \tau)$  as a function of  $\tau$ . This output can then be squared and integrated to give a machine solution of eq. (36). Figure 8 shows an example of a system (a) and its adjoint (b).

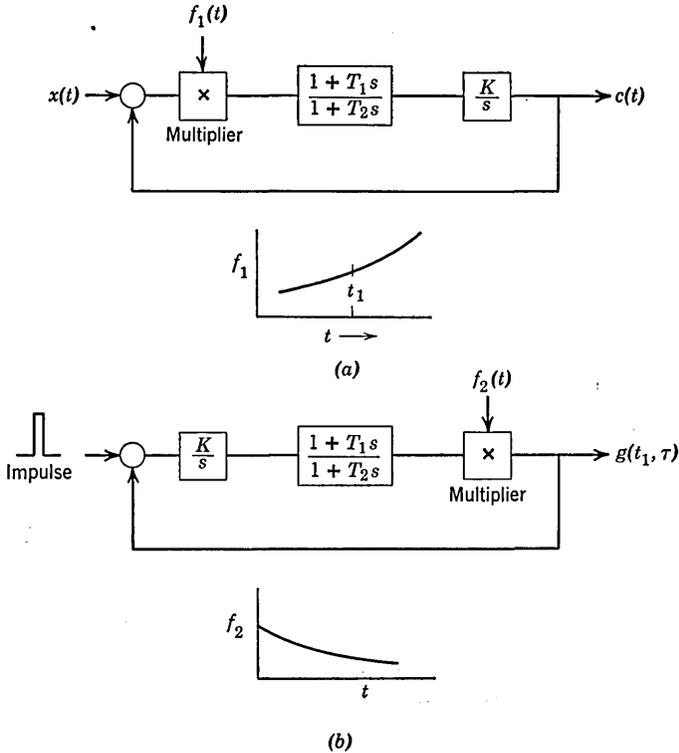


FIG. 8. Analog of a system (a) and its adjoint (b).

**Noise Generators.** Many nonlinear control systems and systems involving a human operator will not yield readily to analytical techniques. In these cases simulation using a random noise generator is required. Several such generators (Refs. 8, 9, 10) are available. In general, the output is a flat spectrum with various amplitude distributions possible. Where the spectrum of the true input noise is known, a shaping filter, as specified in eq. (35), can be used to modify the output of the noise generator.

Although the use of a noise generator and a simulated system provides a simple solution to many complex problems, it is also a time-consuming

one. The methods of the previous section of this chapter provide an indication of the sample lengths required where non-time-varying systems are tested with stationary inputs. For time-varying systems, the answers will be of interest at one or more times during the run. The number of runs required is determined by standard statistical methods.

## 5. SYSTEM DESIGN IN THE PRESENCE OF NOISE

Previous sections of this chapter have shown methods for describing a random input or disturbance and methods for computing the response of a system to these inputs. It remains then to establish procedures which can be used to apply these methods to the design of control systems. Unfortunately each problem is a little or greatly different from any general case so that the designer must examine his problem and determine what methods are adequate.

**Mean Square Error Criteria.** Practically all the work covered in this section aims at minimizing the mean square error of the system. In some cases restraints are placed upon the solution in an attempt to conform more nearly to the practical situation. The *limitations* of this approach are listed below.

1. The mean square error may not be the proper criteria. For instance, in a gun fire control system the object is to maximize the probability of destroying the target.
2. The data concerning the system inputs will seldom be exact enough to warrant an extended analysis or to justify the system complexity required to realize the desired response.
3. The optimum design may be very sensitive to practical limitations of the system, such as gain variations.

As a result, it is suggested that formal methods of optimum design are good guides for a design but that more useful results are obtained by starting with a conventional design and varying the parameters to minimize the noise error as computed by formulas in Sect. 3 of this chapter or by analog computers.

**Optimum Design for Stationary Random Inputs.** For the case where the signal and the noise enter the system at the same point and are both random and stationary and their cross-correlation is zero, the linear filter giving the least mean square error is given by eq. (37) (Ref. 11):

$$(37) \quad G_{\text{opt}}(j\omega) = \frac{\Phi_s(\omega)}{\Phi_s(\omega) + \Phi_n(\omega)} G_d(j\omega),$$

where  $\Phi_s(\omega)$  = signal spectral density,

$\Phi_n(\omega)$  = noise spectral density,

$G_d(j\omega)$  = desired transfer function if noise was not present.

The mean square error is given by eq. (38):

$$(38) \quad \sigma_{\epsilon \min}^2 = \int_{-\infty}^{+\infty} \frac{\Phi_s(\omega)\Phi_n(\omega)}{\Phi_s(\omega) + \Phi_n(\omega)} |G_d(j\omega)|^2 d\omega.$$

Unfortunately,  $G_{\text{opt}}(j\omega)$  will usually be physically unrealizable since its impulse response will have values for  $t < 0$ . *The best physically realizable filter is derived from eq. (37) as follows.* Consider a filter made up of two filters in cascade as in Fig. 9. The input spectral density



FIG. 9. Optimum filter broken into two series filters.

$[\Phi_I(\omega) = \Phi_s(\omega) + \Phi_n(\omega)]$  is split into two functions  $\Phi_I^+(\omega)$  and  $\Phi_I^-(\omega)$  where  $\Phi^+$  contains all the poles and zeros in the upper half-plane and  $\Phi_I^-$  contains all the poles and zeros of the lower half-plane. Then  $G_1(\omega)$  is given by eq. (39):

$$(39) \quad G_1(j\omega) = 1/\Phi_I^+(\omega);$$

for example, if

$$(40) \quad \Phi_s(\omega) = \frac{K_s^2}{\omega^2 + a^2},$$

$$(41) \quad \Phi_n(\omega) = K_n^2.$$

Then the input spectral density is

$$(42) \quad \Phi_I(\omega) = \frac{K_n^2(\omega^2 + b^2)}{\omega^2 + a^2}; \quad b^2 = \frac{K_s^2}{K_n^2} + a^2,$$

and

$$(43) \quad G_1(j\omega) = \left( \frac{a + j\omega}{b + j\omega} \right) \left( \frac{1}{K_n} \right).$$

The transfer function of  $G_2(j\omega)$  is obtained by taking the physically realizable part of the remaining part of the optimum transfer function,  $[\Phi_s(\omega)G_d(j\omega)/\Phi_I^-]$ . This is obtained by finding the impulse response of this remaining part and constructing a filter that will give the same result for positive values of time. For instance, if the desired response to the signal is unity, the term in brackets becomes, for our example:

$$(44) \quad \frac{\Phi_s(\omega)G_d(j\omega)}{\Phi_I^-(\omega)} = \frac{K_s^2}{K_n(b - j\omega)(a + j\omega)}.$$

For positive values of time, the impulse response of this filter is

$$(45) \quad g^+(t) = \frac{K_s^2 e^{-at}}{K_n(a+b)}.$$

The transfer function of a realizable filter which has this impulse response is obtained by taking the Fourier transform of eq. (45).

$$(46) \quad G_2(j\omega) = \frac{K_s^2}{K_n} \left( \frac{1}{a+b} \right) \frac{1}{a+j\omega}.$$

The complete filter is then the cascade of  $G_1(j\omega)$  and  $G_2(j\omega)$ , or simply

$$(47) \quad G_{12}(j\omega) = \frac{K_s^2/K_n^2}{(a+b)(b+j\omega)}.$$

By referring to eq. (42) it is seen that as the noise amplitude becomes very small,  $b$  becomes very large and the gain of the filter becomes unity, and that as the noise becomes very large, the gain goes to zero and gives an intuitive check of our results. If  $G_{12}(j\omega)$  is known, conventional methods may be used to synthesize a control system approximating the optimum response.

In some applications a lag can be tolerated in the output or conversely a lead is desired. In these cases  $G_d(j\omega)$  is no longer unity, but it is given by  $e^{j\omega T}$  where  $T$ , the time displacement, is plus for a lead and negative for a lag. Figure 10 shows a plot of the impulse response of  $G(j\omega)$  without

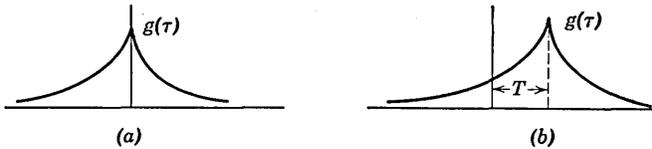


FIG. 10. Effect on the optimum response of allowing a lag  $T$ : (a) zero lag; (b) lag equals  $T$ .

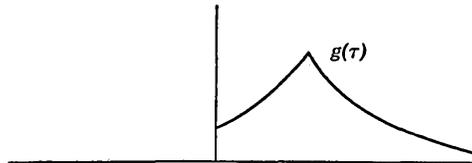


FIG. 11. Physically realizable part of 10(b).

regard for physical realizability for both zero and negative values of  $T$ . For the case of the lag the realizable transfer function for  $G(j\omega)$  would have the impulse response shown in Fig. 11. This response is physically

realizable but would require a very complex network. The advantage of accepting the signal delayed by  $T$  seconds is that it would be more accurate when it arrived.

The same results can be obtained by writing an equation for the optimum impulse response of the control system. This is known as the *Weiner-Hoft equation*:

$$(48) \quad \phi_{ss}(\tau) = \int_0^{\infty} \phi_{II}(\tau)g(\tau) d\tau.$$

In eq. (48)  $\phi_{ss}(\tau)$  is the autocorrelation of the signal,  $\phi_{II}(\tau)$  is the autocorrelation of the input, and  $g(\tau)$  is the optimum impulse response. The Fourier transform of  $g(t)$  is the same as  $G_{12}(j\omega)$  derived previously. Equation (48) can be solved numerically if the autocorrelations are the result of experimental measurement and no analytical expressions for them are available.

**Nonstationary Inputs.** Equation (48) can be extended to the case of nonstationary inputs by rewriting it as follows (Refs. 12 through 15):

$$(49) \quad \phi_{ss}(t, t - \tau) = \int_0^{\infty} \phi_{II}(t - \tau, t - r)g(t - r, t) dr,$$

where  $g(t - r, t)$  is the response of the control system at time  $t$  to an impulse applied at time  $(t - r)$  and the autocorrelation functions are defined by eq. (2). Equation (49) will usually require a numerical solution.

The response  $g(t - r, t)$  is the optimum only at time  $t$ . A different time will in general require a different response so that the system must be time-varying. The values of  $g(t - r, t)$  are most easily obtained by using adjoint techniques with an analog computer (Ref. 7). A cut-and-try variation of parameters can be used to approximate the calculated optimum response.

**System Optimization Under Constraints.** The response called for by the minimum squared error criteria may call for unrealistic demands on components by requiring extended ranges of operation to prevent saturation or by requiring highly rated components to prevent overheating. Hence, the optimization should more practically be carried out under the restraints of rms power dissipated in the output or signal level at various points in the system (Refs. 16 and 17).

**EXAMPLE.** Suppose that the input consists solely of a signal with spectral density  $\Phi_I(\omega)$  and it is desired to limit the rms velocity of the output. The mean square error and velocity are given by eqs. (50) and (51) where  $G(j\omega)$  is the system transfer function.

$$(50) \quad \sigma_\epsilon^2 = 2 \int_0^\infty \Phi_I(\omega) |1 - G(j\omega)|^2 d\omega,$$

$$(51) \quad \left| \frac{dc(t)}{dt} \right|^2 = 2 \int_0^\infty \omega^2 \Phi_I(\omega) |G(j\omega)|^2 d\omega.$$

The quantity  $I$  is formed:

$$(52) \quad I = \sigma_\epsilon^2 + \lambda \left| \frac{dc(t)}{dt} \right|^2$$

$$(53) \quad = \int_0^\infty (\Phi_I(\omega) |1 - G(j\omega)| + \lambda \omega^2 \Phi_I(\omega) |G(j\omega)|^2) d\omega.$$

The object is to minimize  $I$  by the proper choice of  $G(j\omega)$ . The result is a function of  $\lambda$  the Lagrangian multiplier. With this result, the value of  $(dc(t)/dt)^2$  is plotted versus  $\lambda$  and a value of  $\lambda$  is picked that gives the minimum error for the velocity below the specified limit. The solution of eq. (53) is quickly obtained in this case by noting that  $I$  would be the mean square error if  $\lambda \omega^2 \Phi_I(\omega)$  were a noise input. Hence, the optimum transfer function without regard to realizability is

$$(54) \quad G_{\text{opt}}(j\omega) = \frac{\Phi_I(\omega)}{\Phi_I(\omega) + \lambda \omega^2 \Phi_I(\omega)},$$

which has the same form as eq. (37). Hence, the methods of treating eq. (37) can be used in this case.

**Summary.** The designer should be careful not to expect too much of formal techniques of optimum design. On account of shortcomings in the mean square error criteria and the difficulty of obtaining accurate statistical description of the inputs, they will seldom produce practical designs. They do, however, form useful guide posts and establish limits on performance.

## REFERENCES

1. H. F. James, N. B. Nichols, and R. S. Philips, *Theory of Servomechanisms*, Mass. Inst. Technol. Radiation Laboratory Series, Vol. 25, McGraw-Hill, New York, 1947.
2. W. B. Davenport, R. A. Johnson, and D. Middleton, Statistical errors in measurements on random time function, *J. Appl. Phys.*, **23**, 377-388 (1952).
3. J. W. Tukey, Sampling Theory of Power Spectrum Estimates, *Symposium on Application of Auto-Correlation Analysis to Physical Problems*, Woods Hole, Mass., Publication NAVEXOS-P-735, June 1949.
4. S. O. Rice, Mathematical analysis of random noise, *Bell System Tech. J.*, **23**, 282-332 (1944); **24**, 46-56 (1945).

5. N. J. Zabusky, The Mechanical Correlation Computer, Mass. Inst. Technol. Servomechanisms Lab. Rept. No. 6506-ER-32 (ASTIA AD-45).
6. J. H. Laning and R. H. Battin, An application of analog computers to the statistical analysis of time-variable networks, *I.R.E. Trans. Circuit Theory*, **CT-2**, 44-49, March 1955.
7. J. A. Aseltine and R. R. Favreau, Weighting functions for time varying feedback systems, *Proc. I.R.E.*, **42**, 1559-1564 (1954).
8. Goodyear Aircraft Corporation, Random Noise Generator for Simulation Studies, Rept. GER-6436, Dec. 1954.
9. Electronic Associates, Inc., Low Frequency Gaussian Noise Generator Specification for Model 201, Long Branch, N. J.
10. R. R. Bennett and A. S. Fulton, The generation and measurement of ultra low frequency random noise, *J. Appl. Phys.*, **22**, 1187-1192 (1951).
11. H. W. Bode and C. E. Shannon, A simplified derivation of linear least square smoothing and prediction theory, *Proc. I.R.E.*, **38**, 417-425 (1950).
12. A. H. Koschmann and J. G. Truxal, Optimum linear filtering of non-stationary time series, *Proc. Natl. Electronics Conference*, pp. 119-127, National Electronics Conference, Inc., Menasha, Wis., 1954.
13. R. C. Davis, One theory of prediction of non-stationary processes, *J. Appl. Phys.*, **23**, 1047-1053 (1952).
14. L. A. Zadeh and J. R. Ragazzini, An extension of Wiener's theory of prediction, *J. Appl. Phys.*, **21**, 645-655 (1950).
15. R. C. Booton, An optimization theory for time-varying linear systems with non-stationary statistical inputs, *Proc. I.R.E.*, **40**, 977-981 (1952).
16. G. C. Newton, Compensation of feedback control systems subject to saturation, *J. Franklin Inst.*, **254**, 281-296, 391-413 (1952).
17. J. H. Westcott, Synthesis of optimum feedback systems satisfying a power limitation, Paper 53-A-17, *Frequency Response Symposium*, Am. Soc. Mech. Engrs., New York, Dec. 1953.

## Nonlinear Systems

W. M. Gaines

1. Definitions	25-01
2. General Nonlinear System Problem	25-03
3. Methods of Analysis: Linearization	25-07
4. Methods of Analysis: Describing Function	25-13
5. Methods of Analysis: Phase Plane, Graphical Solution of System Equations	25-36
6. Other Methods of Analysis	25-43
7. Nonlinear System Compensation	25-48
References	25-66

### 1. DEFINITIONS

**Definition of Nonlinear System.** A nonlinear system or element is a system or element described by a nonlinear differential equation. The significant feature of a nonlinear differential equation is that the principle of superposition is not applicable. Therefore, in a nonlinear system the simultaneous signal level and operating range of *all* inputs must be specified in order to define the system performance.

**Classes of Nonlinearity: Defined by Static Characteristic** (Refs. 1 and 2). An *essential nonlinearity* is a nonlinearity which is necessary to the basic operation of the system. A relay servo is a good example. By its definition, this type of nonlinearity must be included in any system analysis. It is also referred to as an *intentional nonlinearity*.

*Parasitic nonlinearities* are those nonlinearities which are small or incidental defects in an otherwise linear system. This type of nonlinearity is also referred to as an *incidental nonlinearity*. Neglecting parasitic nonlinearities often does not compromise the early analysis and the preliminary designs can be based on this assumption. However, in more complex, high performance systems and/or systems which are required to follow large step inputs, these nonlinearities may result in unsatisfactory or even unstable operation. These nonlinear effects must, therefore, be considered in a complete analysis.

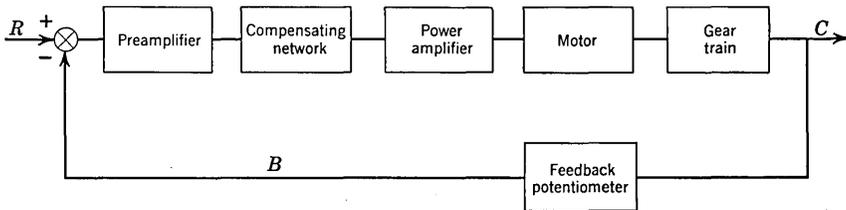


FIG. 1. Components of a simple system.

**EXAMPLES.** Typical sources of parasitic or incidental nonlinearities in a simple system, Fig. 1, are:

- (a) Preamplifier and/or power amplifier saturation.
- (b) Preamplifier, power amplifier, or motor sensitivity limit (due to Coulomb friction or stiction).
- (c) Hysteresis in motor characteristics.
- (d) Motor velocity and acceleration limits (due to magnetic saturation).
- (e) Inherent nonlinear motor characteristics which are linearized for analysis, e.g., series or 2-phase motor characteristics.
- (f) Granularity in the followup potentiometer.
- (g) Backlash and stiction in the gearing.

**Classes of Nonlinearity: Defined by Dynamic Effect.** A *slowly varying nonlinearity* is a nonlinearity which changes system characteristics slowly relative to the response time of the system. Such nonlinearities are often related to the variation of an independent parameter outside the control system. The variation of the effectiveness of the control surfaces of a missile with altitude is typical of this type of nonlinearity.

A *rapidly varying nonlinearity* is a nonlinearity which changes system characteristics at an appreciable rate with respect to the response time of the system. Such nonlinearities are usually functions of the dependent variable, e.g., amplifier saturation.

## 2. GENERAL NONLINEAR SYSTEM PROBLEM

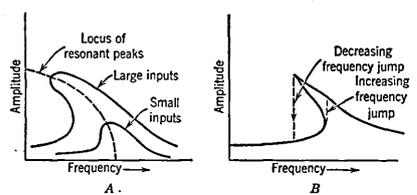
**Importance of Considering Nonlinearities.** Conceivably a system could be designed to operate in an essentially linear manner for all expected inputs. Such a design would probably not be economically feasible. The ratings of power elements, amplifiers, and precision of gear trains, etc., required for such a design would be prohibitive from space, weight, and cost considerations. Fortunately such a design is usually not required and may not be optimum even if achieved. Normally smooth operation is required only for small deviations about the operating point. For large deviations, saturation of amplifiers and/or power elements is often acceptable. The problem for large deviations is to assure that the change in system operating characteristics does not result in unsatisfactory performance when synchronizing on the new operating point.

For small deviations, there exists the problem of minimizing the nonlinearities which cause deadband or signal distortion. Again it is not usually practical to eliminate the cause. The designer attempts to reduce the effect to acceptable limits. The threshold at which an error signal will cause control action is initially a static problem. However, those nonlinearities which are important at low signal levels can cause a self-sustained oscillation of low amplitude. It is necessary to assure that the amplitude and frequency of these oscillations are not sufficient to affect adversely system performance.

Of great importance also is the fact that linear operation represents only a small cross-section of the possible methods of achieving the desired operation. It certainly is to be anticipated that in many cases the best performance can be achieved by nonlinear operation. In fact in a number of cases techniques using nonlinear elements have been employed to improve the performance of servos without other improvements in existing controllers and motors. (See Sect. 7.)

**State of Art in Nonlinear System Study.** There are two types of tasks in nonlinear studies, *analysis* and *synthesis*. Practical analytical techniques of treating nonlinear systems are not available except for the most rudimentary and simple systems. Fortunately, the task of *analysis* lends itself to solution by computers. The analog and digital computer no longer makes it necessary to rely upon tedious hand methods of calculation. Although the ennui of calculating performance of nonlinear systems has been removed, the problem of organizing a general theory still remains. The continuing advancement of any scientific field is dependent upon the ability of the persons involved to generalize, classify, and catalog their results. This provides the foundation for future develop-

TABLE 1. COMMON NONLINEAR PHENOMENA

Type	Condition for Occurrence	Characteristics and/or Effects	General Remarks
1. Jump resonance	Occurs in systems with lightly damped resonances, and nonlinear restoring forces that are excited sinusoidally. Describing function analysis shows that $ G  \cos \gamma > 1$ must be true for phenomena to occur in single loop system with simple saturation. ( $ G $ = open loop gain, $\gamma$ = phase margin). (See Ref. 3.)	When excited by a sinusoidal driving signal, the resonance is normal for small amplitude signals. Theoretically, for larger amplitude signals the resonance bends as in <i>A</i> below. In practice the three-valued function cannot be measured but the response will appear to <i>jump</i> as in <i>B</i> below. The jump will occur at different values of frequency depending upon whether the frequency is increasing or decreasing. The phase characteristic exhibits a corresponding jump.	For jump resonance to occur the system must be second order or higher. To have significant bending the damping must be 0.1 or less in a second order system. The phenomena can occur in systems with saturation or increasing gain characteristics. The bending is to the right for increasing gain characteristics and to the left for saturation. The normalized second order equation of the type $dx^2/dt^2 + dx/dt + f(x) = F \cos \omega t$ (Duffing's equation) has been solved for various forms of the function $f(x)$ and each case has exhibited the jump resonance when the viscous damping was small. (See Ref. 4.) The existence of the jump resonance can be confirmed by the use of describing functions. (See Refs. 3 and 5.)
			
2. Limit cycle or bounded	Occurs in unstable or conditionally stable nonlinear systems. Describing func-	An unstable linear system can exhibit oscillations that grow without bound. A nonlinear system that is unstable	Limit cycle oscillations can arise from a wide variety of system conditions. Conditionally stable systems with sat-

oscillations

tion analysis shows that conditions of  $|G| = 1$  and  $\gamma = 0$  must be met for phenomena to occur in simple systems. For a stable system with an unstable limit cycle it is necessary to excite the system beyond the level of the limit cycle to obtain self-sustained oscillations.

can oscillate at fixed amplitudes. Such oscillations are referred to as *limit cycle oscillations*. Limit cycles can be either *stable* or *unstable* depending upon whether the oscillation converges or diverges from the conditions represented. Depending upon the system characteristics, the limit cycle oscillation can vary from nearly simple harmonic oscillation to a highly nonlinear, relaxation type oscillation. Self-excited oscillations arising in a stable system with unstable limit cycle are referred to as *soft oscillations*. Self-sustained oscillations which occur after the system has been excited to a given level (unstable limit cycle) are referred to as *hard oscillations*.

uration will contain both a stable and unstable limit cycle and an unstable system with saturation will have one stable limit cycle. System imperfections that appear at low signal levels (backlash, friction, etc.) can, under the proper conditions, cause limit cycle oscillations. Existence of this type of limit cycle makes it necessary to define instability in terms of the acceptable magnitude of an oscillation since a low level nonlinear oscillation may or may not be detrimental to performance of the system. Because soft and hard types of oscillations can exist, the designer must specify the input range completely in evaluation or synthesis of a nonlinear system. Limit cycles can be most correctly explained by use of the phase plane; however, the magnitude and fundamental frequency of the limit cycle can be estimated to a first order of magnitude by means of describing functions.

### 3. Subharmonic generation

Appears in nonlinear systems excited sinusoidally. No general rules are available defining the necessary conditions for occurrence. The phenomena have been observed in lightly damped systems with nonlinear restoring force and in systems with nonlinear energy delays.

When the output contains subharmonics of the input exciting frequency, the phenomena is referred to as *subharmonic generation*.

Systems with elements having hysteresis, i.e., backlash, magnetic hysteresis, friction, have been known to exhibit this type of performance when excited with a sinusoidal input. The transition from harmonic to subharmonic operation can be quite sudden, but once the subharmonic is established, it is often quite stable. (See Ref. 8 and its bibliography.)

TABLE 1. COMMON NONLINEAR PHENOMENA (*Continued*)

Type	Condition for Occurrence	Characteristics and/or Effects	General Remarks
4. Inter-modulation effect on gain	Occurs in amplitude-sensitive nonlinear systems excited by two or more frequencies. The frequencies can be separate inputs or one input with a complex waveform. The amplitude of the complex wave must be sufficient to enter the nonlinear region.	Because of the amplitude-sensitive nonlinearity the frequencies will be intermodulated. This causes the original frequency components to have different amplitudes and phase shift than obtained from the nonlinear system with only one frequency present. This can be interpreted as a different phase shift and/or attenuation through an element. The effect is also apparent when noise is present with the signal.	In a simple saturating system the effect can be explained quite easily. Two frequencies are considered. After the saturation the amplitude of both will be reduced beyond that expected if only one frequency had been present. If we are considering the effective gain with respect to one of the two frequencies, the gain will have been reduced. This gain reduction in the open loop can be interpreted as reducing the gain crossover and therefore increasing the phase shift of the closed loop for the frequency being considered. (See Ref. 7.) By considering one of the frequencies as an extraneous signal, the effect of noise on the performance of a saturating system can be envisioned. The effect is particularly significant if the amplitude or phase shift of the closed loop is important to system performance.

ment. A great deal of work of this type remains to be done in nonlinear system analysis.

The major problem for the systems engineer lies in *synthesis* of a control. In synthesis one needs in addition to a full appreciation of the characteristics and a complete understanding of the nature of the task to be performed: (a) methods of rapidly, approximately estimating the effect of different types of compensation in order to allow selection of a potentially good approach and (b) having selected a general approach, a logical design procedure which converges on the "best" design. Although no such generally satisfactory method exists, the methods of linearization, describing functions, and phase plane analysis are powerful analytical tools for attacking nonlinear problems. (See Sects. 3, 4, and 5.) By and large these methods are not exact but often suffice for preliminary design calculations. *The majority of these methods attempt to linearize the problem sufficiently to allow the use of the well-known techniques used in the study of linear systems.* Because of the difficulty in providing generalized design criteria or design charts for any but the simplest nonlinear system, the synthesis of a system using nonlinear elements is primarily a cut-and-try process tempered with common sense.

**Unusual Phenomena Peculiar to Nonlinear Systems.** Many unusual phenomena occur in nonlinear systems. In a linear system the response to a given input defines the response to be expected from any input. This is not true for a nonlinear system. Cases arise where performance may completely deteriorate between a step response and a sinusoidal response. Table 1 summarizes some of the more common types of nonlinear phenomena which have been catalogued. The list demonstrates that the designer must be aware of the peculiar characteristics exhibited by nonlinear systems and completely specify the operating conditions in order to proceed with an intelligent, efficient design. The types of nonlinear phenomena described in Table 1 are those most commonly encountered in nonlinear feedback control systems. Other types of nonlinear phenomena have been catalogued; frequency entrainment, asynchronous excitation and asynchronous quenching, parametric excitation, etc. See Ref. 6 and its bibliography for more details on nonlinear phenomena.

### 3. METHODS OF ANALYSIS: LINEARIZATION

Frequency response analysis can be used only when the system is described by linear constant coefficient equations. Certain nonlinear systems can be linearized by use of the perturbation theory. The method assumes that for very small deviations about the operating point the system is linear. The perturbation method determines the coefficients of the new linear equation describing the performance of the system. Once

reduced to a linear form the usual frequency response techniques can be applied.

**Method of Evaluating Linearized Coefficients.** A nonlinear function of one or more variables can be linearized if the function is analytic. (Refs. 6 and 9.) Expand the function in a Taylor series about the operating point and neglect all the second order and higher derivatives. One thereby considers only the incremental change about a nominal value.

If the function is  $f(x_1, x_2, x_3, \dots, x_n) = f(x_i)$ , then the Taylor expansion about point  $a_i$  ( $x_1 = a_1, x_2 = a_2, x_3 = a_3$ , etc.) is

$$y = f(x_i) = f(a_i) + \sum_{i=1}^n \left. \frac{\partial f(x_i)}{\partial x_i} \right|_{x_i=a_i} (x_i - a_i) + \frac{1}{2} \sum_{i=1}^n \left. \frac{\partial^2 f(x_i)}{\partial x_i^2} \right|_{x_i=a_i} (x_i - a_i)^2 \\ + \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^n \left. \frac{\partial^2 f(x_i, x_k)}{\partial x_i \partial x_k} \right|_{\substack{x_i=a_i \\ x_k=a_k}} (x_i - a_i)(x_k - a_k) + \dots, \quad k \neq i,$$

or

$$y = f(x_i) \approx f(a_i) + \sum_{i=1}^n \left. \frac{\partial f(x_i)}{\partial x_i} \right|_{x_i=a_i} (x_i - a_i) = y_0 + \Delta y,$$

and

$$(1) \quad \Delta y = \sum_{i=1}^n \left. \frac{\partial f(x_i)}{\partial x_i} \right|_{x_i=a_i} (x_i - a_i)$$

for small values of  $x_i - a_i$ . Equation (1) is a linear relation between  $\Delta y$  and  $x_i - a_i$  or  $\Delta x_i$ , the incremental changes.

The accuracy of the approximation can be estimated by evaluating the next terms in the Taylor series, e.g., the second order terms are

$$\frac{1}{2} \sum_{i=1}^n \left. \frac{\partial^2 f(x_i)}{\partial x_i^2} \right|_{x_i=a_i} (x_i - a_i)^2 \\ + \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^n \left. \frac{\partial^2 f(x_i, x_k)}{\partial x_i \partial x_k} \right|_{\substack{x_k=a_k \\ x_i=a_i}} (x_i - a_i)(x_k - a_k), \quad k \neq i.$$

**Alternate Method of Evaluation of Linearized Coefficients.** The same linearization can be accomplished by substituting  $a_1 + \Delta x_1, a_2 + \Delta x_2$ , etc., for  $x_1, x_2$ , etc., in the original function, and neglecting all second order and higher terms, i.e., terms containing the product of the type  $\Delta x_1 \Delta x_2$ . Here  $a_1$  and  $a_2$  are the values at the operating point and  $\Delta x_1$  and  $\Delta x_2$  are small deviations.

EXAMPLE. Consider the nonlinear function  $f(x, y) = w = xy$ . Substituting  $x = x_0 + \Delta x$ ,  $y = y_0 + \Delta y$ ,  $w = w_0 + \Delta w$

$$w_0 + \Delta w = x_0 y_0 + \Delta y x_0 + x_0 \Delta y + \Delta x \Delta y.$$

If  $\Delta x$  and  $\Delta y$  are small, the term  $\Delta x \Delta y$  is small and can usually be neglected.

$$w_0 + \Delta w \approx x_0 y_0 + \Delta x y_0 + \Delta y x_0.$$

Since  $w_0 = x_0 y_0$ , the remaining two terms must be deviations, therefore,

$$(2) \quad \Delta w \approx \Delta x y_0 + \Delta y x_0.$$

Equation (2) is a linear expression for  $\Delta w$ . For small variations about the operating point  $x_0, y_0$ , eq. (2) describes the performance. The coefficient  $y_0$  is the *gain* between  $\Delta x$  and  $\Delta w$ , and  $x_0$  is the gain between  $\Delta y$  and  $\Delta w$ .

The same relationship would have been obtained by using the Taylor series expansion.

**Table of Useful Algebraic Approximations for Linearization.**

Table 2 is useful when making the above substitution into nonrational equations. The terms in Table 2 were determined by considering the

TABLE 2. USEFUL ALGEBRAIC APPROXIMATIONS

$m \ll 1$		
Algebraic Expression	Approximation	Next Term in Series
1. $\frac{1}{1+m}$	$1 - m$	$+m^2$
2. $(1+m)^n$	$1 + mn$	$+\frac{n(n-1)m^2}{2}$
3. $e^m$	$1 + m$	$+\frac{m^2}{2}$
4. $\log_e(1+m)$	$m$	$-\frac{m^2}{2}$
5. $\sin(m)$	$m$	$-\frac{m^3}{6}$
6. $\cos(m)$	$1$	$-\frac{m^2}{2}$
7. $(1+m_1)(1+m_2)$	$1 + m_1 + m_2$	$+m_1 m_2$

series expansion of the closed form. The last column of the table can be used to estimate the accuracy of the approximation.

To use the table, it is necessary to work the expression into a nondimensional form.

EXAMPLE. Consider the flow through a variable orifice. Flow,  $q$ , is given by

$$q = C_d A \sqrt{P},$$

where  $A$  = orifice area, a variable,  
 $P$  = pressure drop, a variable,  
 $C_d$  = flow coefficient, a constant.

By substituting the incremental change form

$$q_0 + \Delta q = C_d(A_0 + \Delta A_0)\sqrt{P_0 + \Delta P}.$$

Dividing the quantity under the radical sign by  $P_0$  yields

$$q_0 + \Delta q = C_d(A_0 + \Delta A_0)\sqrt{P_0} \sqrt{1 + \Delta P/P_0}.$$

Using approximation 2 in Table 2 for  $n = \frac{1}{2}$  yields the expression

$$q_0 + \Delta q \approx C_d(A_0 + \Delta A_0)\sqrt{P_0} \left(1 + \frac{1}{2} \frac{\Delta P}{P_0}\right).$$

By expanding and neglecting higher order and constant terms this reduces to

$$\Delta q = \left(\frac{C_d A_0}{2\sqrt{P_0}}\right) \Delta P + (C_d \sqrt{P_0}) \Delta A,$$

where the terms in parenthesis are the equivalent gain between  $\Delta P$  and  $\Delta A$  and flow,  $\Delta q$ .

**Graphically Linearizing System Characteristic.** The analytical expression for the function need not be known. If the graphical relation-

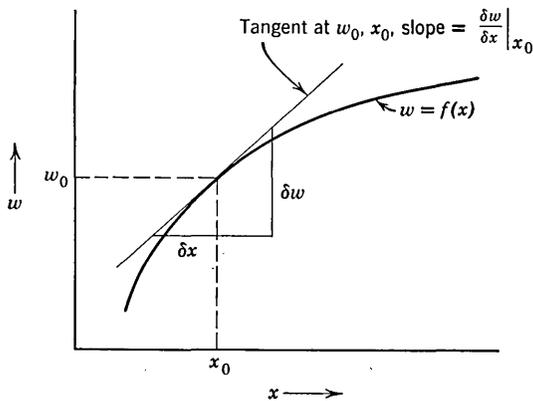


FIG. 2. Determination of the linearized coefficient from a plot of the function.

ship between the variables is known, a linear expression can be obtained by considering incremental departures from the nominal values at the particular operating condition being considered. For Fig. 2 then

$$\begin{aligned}
 w_0 + \Delta w &= f(x_0) + (\text{slope of function at } x_0, w_0) \Delta x \\
 &= w_0 + \left. \frac{\delta w}{\delta x} \right|_{x_0} \Delta x
 \end{aligned}$$

and

$$\Delta w = \left. \frac{\delta w}{\delta x} \right|_{x_0} \Delta x.$$

The method can be extended to functions of more than one variable by obtaining the slopes from the appropriate curves. *Note that in dealing with functions of more than one variable the slope must be taken so as to be independent of all variables but the ones being considered.* In Fig. 3 a function of

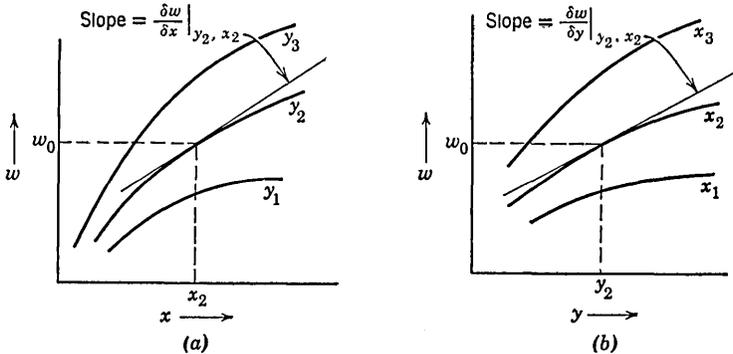


FIG. 3. Determination of the linearized coefficients from cross-plots of a function of two variables: (a)  $xw$  plot and (b)  $yw$  plot.

two variables is cross-plotted to obtain the independent slopes from separate plots. From these the linear relation for small deviations at values of the variable  $x_2, y_2$  is

$$\Delta w = \left. \frac{\delta w}{\delta x} \right|_{x_2, y_2} \Delta x + \left. \frac{\delta w}{\delta y} \right|_{x_2, y_2} \Delta y.$$

For functions of two variables that are reasonably regular it is usually possible to pick the values for calculating the slopes from a single plot of the function and avoid the labor of cross-plotting the functions (see Ref. 9).

**Use of Linearized Coefficients.** The characteristics of the function or system are approximated by the linearized coefficients for small varia-

tions from the operating point. Therefore, for small disturbances, only the deviation terms need to be considered in determining the stability. The

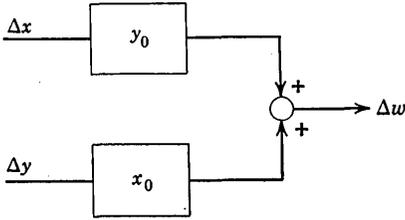


FIG. 4. Equivalent block diagram for the linearized small deviation expression for the function  $w = xy$ .

“constant” terms defining the operating point will remain the same to a first order approximation. Under these conditions the block diagram for the multiplication  $w = xy$  is shown in Fig. 4.

For small disturbances, the approximation can be substituted into the system diagram and the usual methods of linear analysis and compensation used. Note that one is now dealing only with the deviations

and not the total variable. For further applications of this method see Sect. 7.

**Limitations.** The above approximations are limited to small deviations from the operating point. The errors get progressively worse as the signal level is increased, and considerable care must be exercised when dealing with large excursions. The validity of the approximation can be checked by evaluating the next terms in the Taylor series or the last column of Table 2.

*The perturbation theory is valid only if the derivative of the function exists.* The method would be of doubtful value in dealing with a relay characteristic.

The method is sometimes limited when the operating point is at 0. One or more of the variables will then have a steady-state value of 0, and terms involving that variable and the deviation can be lost. This can lead to an indeterminate or inaccurate solution. For instance, in the example of linearizing  $w = xy$  if either  $x_0$  or  $y_0$  were 0, the variation of  $\Delta w$  with the corresponding value of  $\Delta x$  or  $\Delta y$  is zero. See eq. (2). Although the error of the approximation is still  $\Delta x \Delta y$  this term becomes significant with respect to the other terms as  $x_0$  and  $y_0$  become small. Consider the example of the flow through an orifice. The linearized deviation expression is repeated here:

$$\Delta q = \left( \frac{C_d A_0}{2\sqrt{P_0}} \right) \Delta P + (C_d \sqrt{P_0}) \Delta A.$$

As  $P_0 \rightarrow 0$ , this expression becomes indeterminate and the analysis based on the approximation under these conditions loses significance. A linear expression no longer adequately describes these situations.

#### 4. METHODS OF ANALYSIS: DESCRIBING FUNCTION

##### General

**Definition of Describing Function.** When an element is excited by a sinusoidal signal, the describing function is the ratio of the fundamental of the output signal to the sinusoidal input signal. A describing function of an element may be a complex quantity, characterizing both amplitude and phase relations between the input and output. The describing function may be a function of both signal amplitude and frequency.

**Use of the Describing Function.** Within the validity of the basic assumptions, the describing function, representing the nonlinear element, can be substituted directly into the system equations for the nonlinear characteristics. The use of the describing function quasi-linearizes the frequency response equations. Since the describing function will be a function of amplitude, the system frequency response will be a function of both frequency and amplitude. The quasi-linearization of the system equations in the frequency domain allows the use of the Nyquist criterion to determine stability.

Although the dependency of the frequency response upon both frequency and amplitude complicates the calculations slightly, practical methods are available which require little effort beyond that normally required in the plotting of a Nyquist diagram.

**Usefulness of Describing Function.** The describing function provides an approach which allows a solution in the frequency domain. The ability to manipulate the system equation in the frequency domain is valuable because: (a) frequency response techniques developed for linear system are available for synthesis, (b) synthesis and analysis can be handled with relative ease, and (c) the technique is not limited to systems with few energy storage elements.

**Basic Assumptions of Describing Function Method.** If the input to the nonlinear element is sinusoidal, then it is assumed that:

1. The output is periodic and of the same fundamental period as the input signal.
2. Only the fundamental of the output wave need be considered in a frequency response analysis.
3. The nonlinear element is not time-varying.
4. Only one nonlinear element is considered to exist in the system.

Assumption 1 implies that no subharmonics are generated.

If the element used in a system is driven by sinusoidal signal, the output of the device by assumption 2 is considered to be sufficiently filtered by

the system characteristics so that the signal fed back into the input of the nonlinear element is essentially sinusoidal. The degree to which the input to the nonlinear element must be "essentially sinusoidal" is determined by how critical the nonlinearity is to the wave shape of the driving signal.

While the describing function cannot be obtained for a system in which the coefficients are time-varying because the output would not reach a steady-state periodic solution, the describing function can be obtained for an element with characteristics which are dependent on frequency. In such a case the describing function will be a function of both amplitude and frequency of the signal.

If a system contains *two nonlinearities* of major importance, it is still possible to get a describing function for the system. Often the easiest way to obtain the describing function in this case is to lump the characteristics of the two nonlinearities and obtain an over-all describing function. In general, it is not practical to consider each nonlinearity separately.

**Theory of Describing Functions.** Consider the system of Fig. 5. It is convenient in describing function analysis to have the nonlinear and

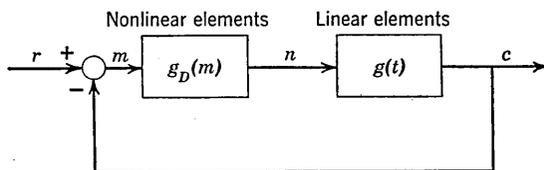


Fig. 5. Block diagram of simplified nonlinear system.

linear elements separated as in the figure. In the following it will be assumed that this has been done.

The output,  $n$ , of the nonlinear element is related to the input,  $m$ , by

$$n = [g_D(m)]m.$$

If the input is sinusoidal, then, by the assumptions made, the error  $M(j\omega)$  must be sinusoidal and of the fundamental frequency.

$$N(j\omega) = G_D[M(j\omega)]M(j\omega).$$

The output,  $N(j\omega)$ , can be represented by a Fourier series:

$$N(j\omega) = N_1(j\omega) + N_2(j2\omega) + N_3(j3\omega) + \dots,$$

where  $N_1(j\omega)$  is the first harmonic of the output  $N(j\omega)$ .

By definition, the describing function is

$$G_{D_1}(|M|, \omega) = \frac{N_1(j\omega)}{M(j\omega)} = \left| \frac{N_1(j\omega)}{M(j\omega)} \right| \text{angle} \left( \frac{N_1(j\omega)}{M(j\omega)} \right),$$

where  $G_{D_1}(|M|, \omega)$  is the describing function as a function of amplitude,  $|M|$ , and frequency,  $\omega$ .

Usually for convenience ( $|M|, \omega$ ) and the subscript 1 are dropped; this leaves  $G_D$  as the symbol for describing functions.

Within the validity of the assumption that the harmonics are sufficiently filtered by the linear system elements so that the feedback signal contains essentially the fundamental, the harmonics of the output of the nonlinear element can be neglected, and the describing function  $G_D$  can be used as a series element in the frequency response analysis.

$G_D$  can be determined by conventional Fourier series analysis. (See Chap. 14 and Ref. 10.)

**Stability Criterion.** For the system of Fig. 5 the frequency response is approximated by

$$\frac{C}{R}(j\omega) = \frac{G_D G(j\omega)}{1 + G_D G(j\omega)},$$

where  $G_D$  = the describing function,

$G(j\omega)$  = the frequency sensitive portion of the system.

For a minimum phase shift system (see Chap. 23) the system will be critically stable when the denominator is zero or

$$1 + G_D G(j\omega) = 0,$$

or

$$(3) \quad G(j\omega) = -1/G_D,$$

or

$$(4) \quad -G_D = 1/G(j\omega).$$

When eq. (3) or (4) is satisfied, the system will have a sustained oscillation of the amplitude and frequency which satisfies the equation (Ref. 11).

**Limitations and Accuracy of Describing Function Method.** There are two major disadvantages with the describing function analysis:

1. There is no convenient method to determine the accuracy. A method proposed by Johnson (Ref. 12) becomes laborious for more than the most simple systems.

2. Frequency response analysis allows prediction of the transient response. A describing function analysis allows at best a qualitative interpretation. These designs can therefore predict stability and frequencies of oscillation, but are limited to crude rules of thumb in prescribing

a given stable response. It must be pointed out that such approximations are often all that is justified by the accuracy of other system data available.

In a wide variety of applications, use of describing functions has allowed prediction of the frequency and amplitude of oscillations within 20 per cent. It is difficult to generalize, but usually the method will be most accurate when  $G_D$  and  $G(j\omega)$  are varying rapidly in the region of intersection (Ref. 12). Erroneous results may be obtained in some cases when the intersection is approximately tangential (Ref. 13).

Because the assumption of sinusoidal input to the nonlinear element is not exact, the method works best when  $G_D$  is not sensitive to wave shape of the driving function.

NOTE. *The describing function method of analysis is the only practical analytical technique for treating nonlinear systems which are higher than second order.*

**Describing Function: Methods of Presentation**

**Inverted Nyquist Diagram.** The equation for sustained oscillation,

$$-G_D = 1/G(j\omega),$$

indicates that an intersection of the  $-G_D$  and  $1/G(j\omega)$  loci is a point having a given frequency and amplitude at which sustained oscillation can occur.

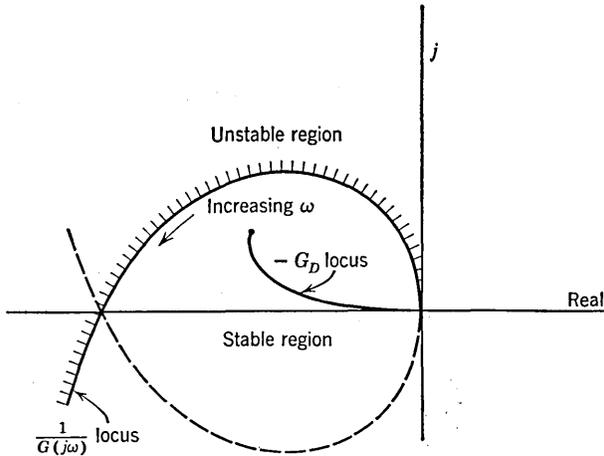


FIG. 6. Inverted Nyquist diagram for a stable nonlinear system showing the stable and unstable regions for the describing function locus.

In a normal inverted Nyquist plot the  $-1$  point should be on the left as the  $1/G(j\omega)$  locus is traversed in the direction of increasing frequency for a minimum phase shift system. In this case, there is no longer a fixed

$-1$  point, but the system is stable if the  $-G_D$  locus is entirely on the left of the  $1/G(j\omega)$  locus.

The stability criteria is then as follows: *If a locus of all possible values of  $G_D$  is plotted, then the system will be stable if the  $-G_D$  locus does not intersect the  $1/G(j\omega)$  locus and  $G_D$  locus lies completely on the left-hand side of the  $1/G(j\omega)$  locus when the  $1/G(j\omega)$  locus is traversed in the direction of increasing frequency. (Valid for minimum phase functions.)*

A stable system is shown in Fig. 6.

NOTE. *By plotting in this manner the frequency sensitive,  $G(j\omega)$ , and amplitude sensitive,  $G_D$ , portions of the system have been separated and can be considered independently.*

**Nyquist Diagram.** Equation 3 leads to a Nyquist diagram as indicated in Fig. 7.

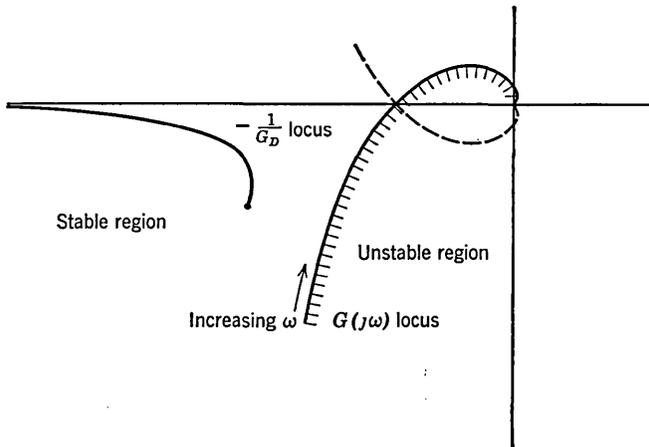


FIG. 7. Nyquist diagram for a stable nonlinear system showing the stable and unstable regions for the describing function locus.

The stability criteria is as follows: *If a locus of all values of  $-1/G_D$  is plotted, then the system will be stable if the  $-1/G_D$  locus does not intersect the  $G(j\omega)$  locus and the  $1/G_D$  locus lies entirely on the left-hand side of the  $G(j\omega)$  locus when the  $G(j\omega)$  locus is traversed in the direction of increasing frequency.*

A stable system is shown in Fig. 7.

**Log-Angle Plane Representation.** It is sometimes more convenient to work with magnitude and phase angle semi-independently. This can be done in the case of describing functions by use of the log magnitude-angle plot. These are the familiar coordinates used on the Nichols charts. (See Chap. 21, Sect. 7).

For the case of a nonlinear system, the critical point is

$$(5) \quad \begin{aligned} 20 \log_{10} |G(j\omega)| &= 20 \log_{10} |1/G_D| \\ \angle G(j\omega) &= -180^\circ - \angle G_D. \end{aligned}$$

If the conditions of eqs. (5) are met, the system is unstable. A typical plot of a stable system servo is given in Fig. 8. As long as the two loci do not intersect, the system will be stable.

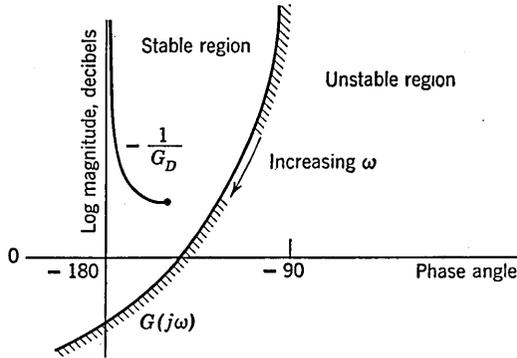


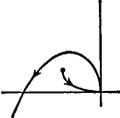
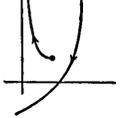
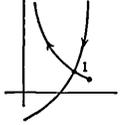
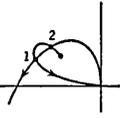
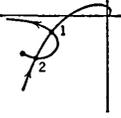
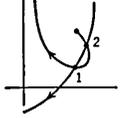
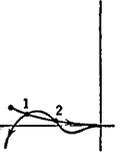
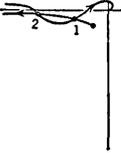
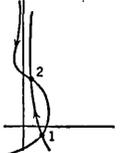
FIG. 8. Log magnitude-angle diagram for a stable nonlinear system showing the stable and unstable regions for the describing function locus.

**Typical Loci for Nonlinear Systems.** Table 3 shows a number of different types of loci which can be expected. In each of these diagrams, the frequency and amplitude loci have been plotted. The arrows indicate increasing frequency and amplitude of signal input to the nonlinear element.

System *B* has an intersection at point 1. This indicates that the system will be unstable at amplitudes less than those at point 1 and will be stable for larger amplitudes. This system, therefore, will be unstable and oscillate at the amplitude and frequency of point 1. If a small disturbance is introduced into the system *B* described in Table 3, the system will appear unstable and the amplitude of oscillation will increase until point 1 is reached. If the amplitude of oscillation becomes larger than this, the system appears stable and any oscillation would tend to die down to that corresponding to point 1. The amplitude and frequency corresponding to point 1 are the amplitude and frequency of the sustained oscillation.

Point 1 of system *B* is called a *convergent point* because disturbances at either side tend to converge at these conditions. This is contrasted with point 2 of system *C*, Table 3, which is a *divergent point* since disturbances which are not large enough to give this value of  $G_D$  will decay and disturb-

TABLE 3. TYPICAL LOCI FOR AMPLITUDE SENSITIVE NONLINEAR SYSTEMS

Diagram Type	Inverted Nyquist Diagram	Nyquist Diagram	Log-Angle Diagram
Stability criteria	$-G_D = \frac{1}{G(j\omega)}$	$-\frac{1}{G_D} = G(j\omega)$	$20 \log_{10}  G(j\omega) $ $= 20 \log_{10}  1/G_D ,$ $\angle G(j\omega)$ $= -180^\circ - \angle G_D$
A. Stable system			
B. System with a convergent point			
C. System with a convergent point 1 and a divergent point 2; Case I, stable for small signal			
D. System with a convergent point 1 and a divergent point 2; Case II, unstable for small signals and very large signals			

ances which are larger will result in oscillations which tend to increase in amplitude.

In system  $D$ , point 1 is convergent and 2 is divergent.

**Comparison of the Methods of Presentation.** All the methods are equally valid. The designer may thus choose the one with which he is most familiar and/or the one which best fits the design problem. In the inverted Nyquist diagram many of the simpler describing functions are bounded, whereas in the Nyquist diagram the describing functions will be infinite for some conditions. The other factors that influence the selection of one form of the Nyquist diagram are still applicable. (See Chap. 21.)

The log-angle method of depicting the stability of the system has advantages in synthesizing a system containing nonlinearities. It is somewhat quicker to plot since  $G(j\omega)$  can be obtained directly from a Bode diagram. This method of display also lends itself to use with Nichols charts and templates. (See Sect. 7.)

**Frequency Variant Describing Functions.** (See Ref. 12.) Describing functions which vary with frequency as well as signal amplitude will appear graphically like the typical plot of Fig. 9. The describing function becomes a surface in three dimensions (magnitude, phase, frequency), and if this surface is pierced by the frequency locus (also plotted

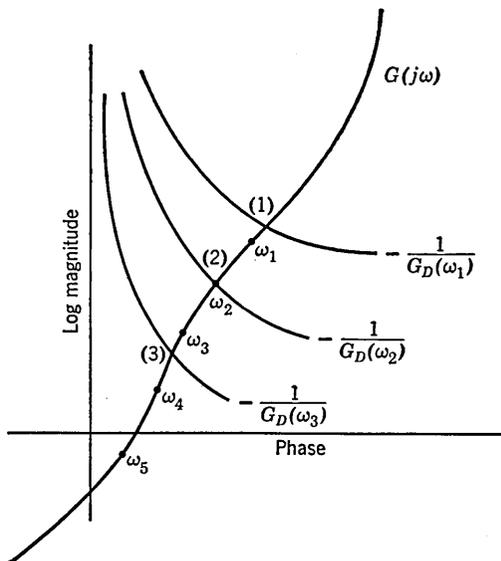


FIG. 9. Log magnitude-angle diagram for a nonlinear system with a frequency and amplitude sensitive nonlinearity. Intersections (1) and (3) are not significant because at these intersections  $G(j\omega)$  and  $G_D(\omega)$  do not have the same frequency.

in three dimensions), the system will be unstable at the frequency and amplitude of the intersection. In other words, to have a significant intersection, it is necessary to have the intersection of the  $G_D(j\omega, |M|)$  and  $G(j\omega)$  loci occur at the same frequency. This is indicated in Fig. 9 at  $\omega_2$ .

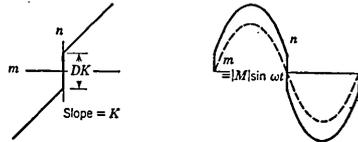
### Tables of Useful Describing Functions

**Amplitude Sensitive Nonlinearities.** Table 4 gives some of the more common describing functions for simple amplitude sensitive nonlinearities, with corresponding graphs in Figs. 10-16.

TABLE 4. USEFUL AMPLITUDE SENSITIVE DESCRIBING FUNCTIONS

Type of Non-linearity	Nonlinear Characteristic	Output Wave Shape, $n$ (Input = $m =  M  \sin \omega t$ )	Equations of Describing Function	Graph of Describing Function
Saturation			$\frac{G_D}{K} = \frac{2}{\pi} \left( \sin^{-1} \frac{S}{ M } + \frac{S}{ M } \cos \sin^{-1} \frac{S}{ M } \right)$	Fig. 10
Deadband			$\frac{G_D}{K} = \frac{2}{\pi} \left( \frac{\pi}{2} - \sin^{-1} \frac{B}{2 M } - \frac{B}{2 M } \cos \sin^{-1} \frac{B}{2 M } \right)$	Fig. 11
Hysteresis			$\frac{G_D}{K} = \sqrt{a^2 + b^2} \angle \tan^{-1} \frac{b}{a}$ $b = \frac{H}{\pi  M } \left( \frac{H}{ M } - 2 \right)$ $a = \frac{1}{\pi} \left[ \frac{\pi}{2} + \sin^{-1} \left( \frac{ M  - H}{ M } \right) + \left( \frac{ M  - H}{ M } \right) \cos \sin^{-1} \left( \frac{ M  - H}{ M } \right) \right]$	Fig. 12
Negative deficiency (type 1)			$\frac{G_D}{K} = \sqrt{a^2 + b^2} \angle \tan^{-1} \frac{b}{a}$ $b = -\frac{1}{\pi} \left( \frac{D}{ M } \right)^2$ $a = \frac{1}{\pi} \left( \pi + \frac{2D}{ M } \cos \sin^{-1} \frac{D}{2 M } \right)$	Fig. 13a

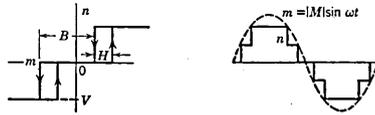
Negative deficiency (type 2)



$$\frac{G_D}{K} = 1 + \frac{2}{\pi} \frac{D}{|M|}$$

Fig. 13

Relay



$$G_D = \frac{4}{\pi} \frac{V}{|M|} \sin \left( \frac{\beta + \psi}{2} \right) \angle -\frac{\psi - \beta}{2}$$

Fig. 14

$$\beta = \cos^{-1} \left( \frac{B + H}{2|M|} \right)$$

$$= \cos^{-1} \left( \frac{B - H}{2|M|} \right)$$

Granularity

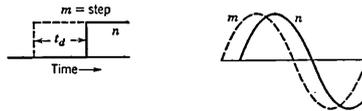


$$G_D = \frac{2}{\pi} \frac{V}{|M|} \left( \sum_{a=1}^{A_i} \sqrt{4 - (2a - 1)^2 \frac{B^2}{|M|^2}} \right)$$

Fig. 15

$$A_i = \text{largest integer value of } \frac{|M|}{B} + \frac{1}{2}, \frac{|M|}{B} > \frac{1}{2}$$

Delay time



$$G_D = e^{-i\omega t_d}$$

(There is no harmonic distortion in delay time. This expression is exact.)

Variable gain

$$n = m^K$$

$$G_D = \frac{2}{\sqrt{\pi}} |M|^{K-1} \frac{\Gamma[(K+1)/2]}{\Gamma[(K+2)/2]}$$

Fig. 16

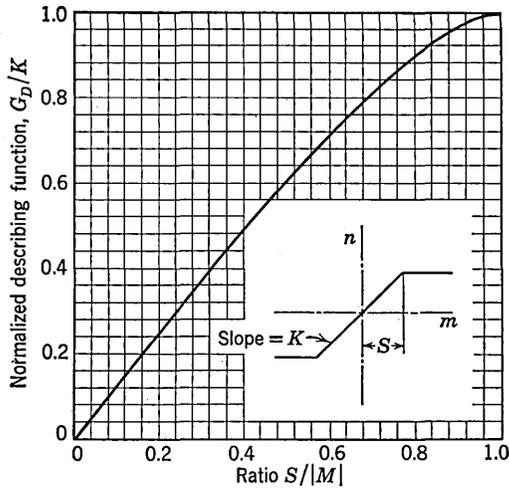


FIG. 10. Describing function for saturation.

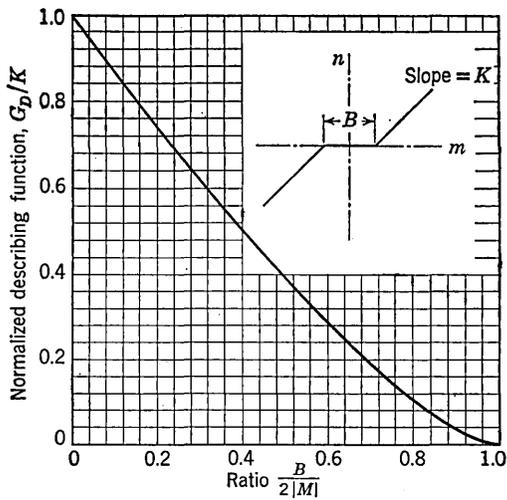


FIG. 11. Describing function for deadband.

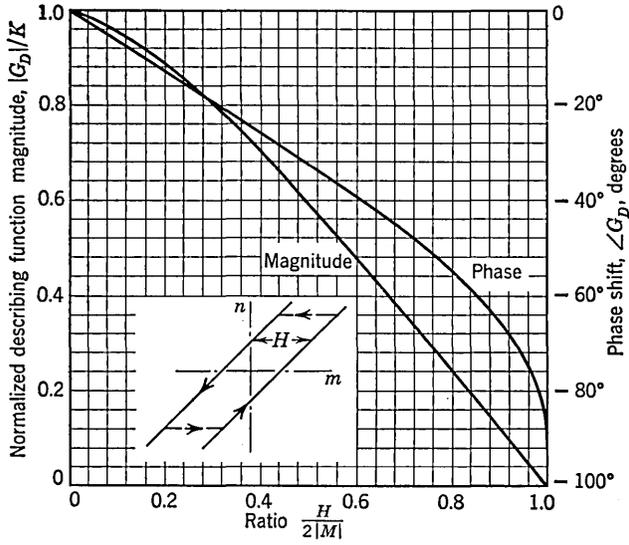
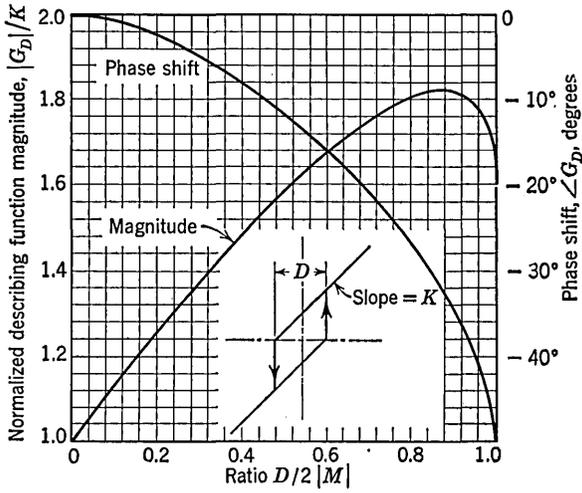
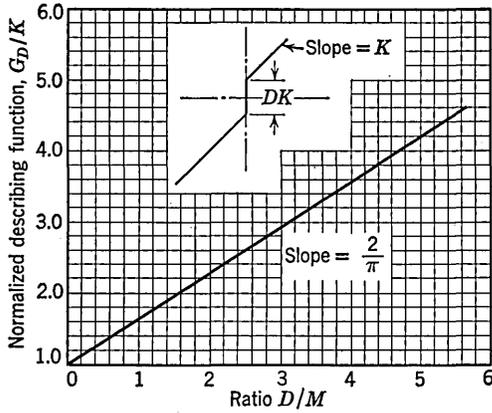


FIG. 12. Describing function for hysteresis.



(a)



(b)

FIG. 13. Describing function for negative deficiency: (a) type 1, (b) type 2.

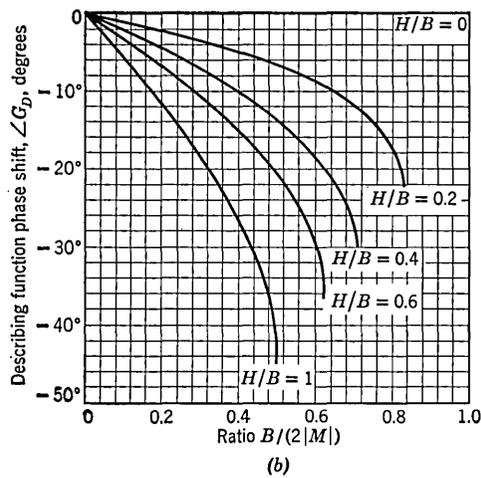
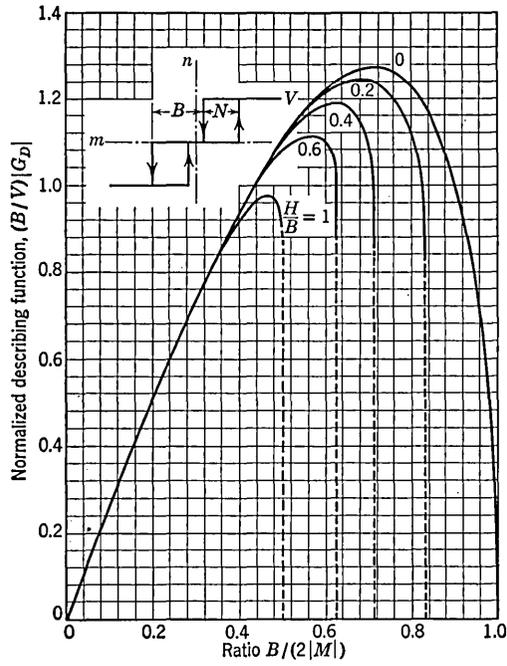


FIG. 14. Describing function for a relay contactor with hysteresis: (a) magnitude, (b) phase shift.

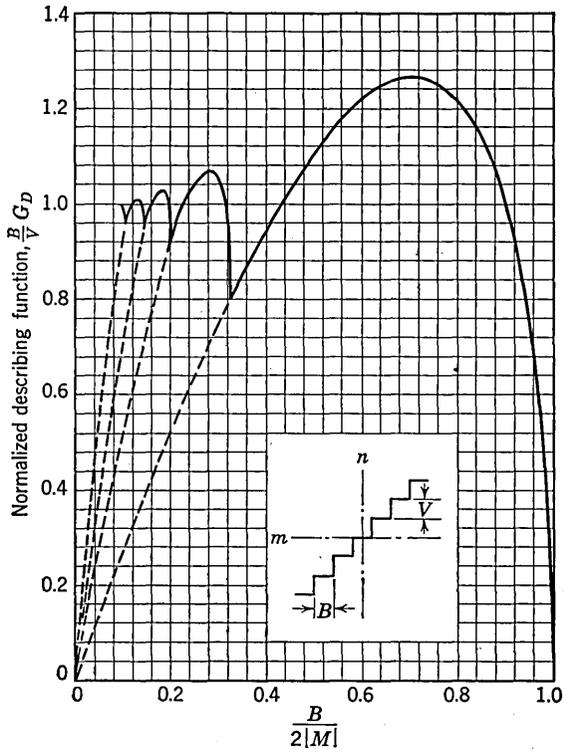


FIG. 15. Describing function for granularity. Dashed lines are used when granularity has a finite number of steps.

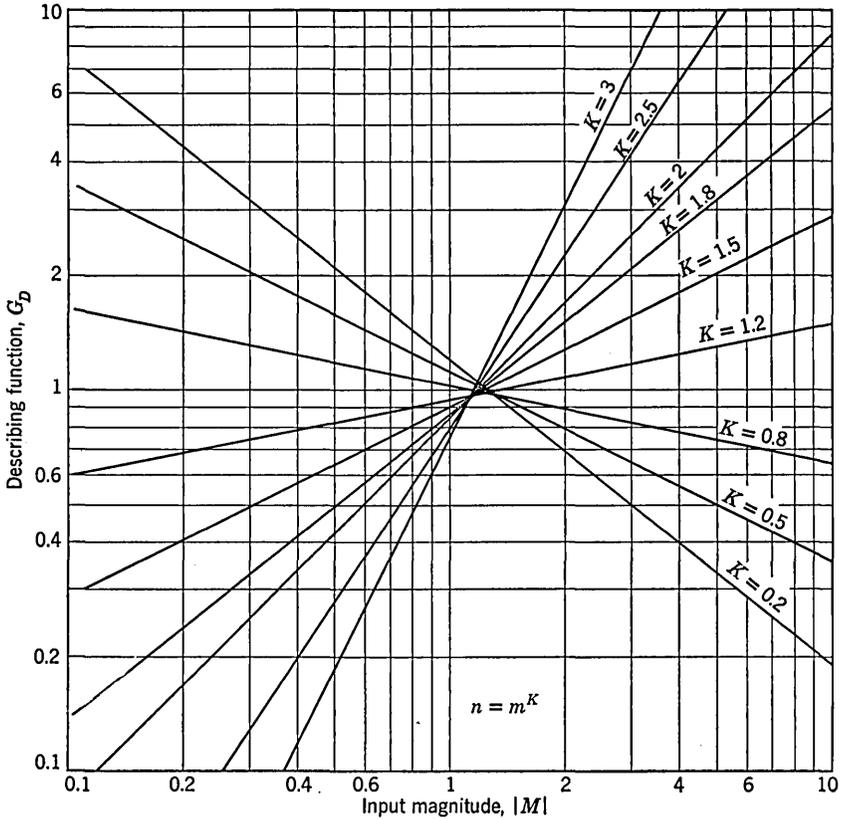


FIG. 16. Describing functions for variable gain elements,  $n = m^K$ .

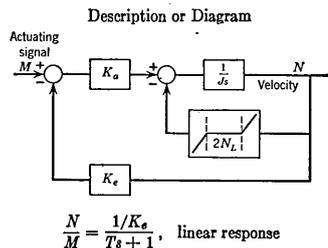
**Complex Nonlinearities.** Table 5 gives the block diagrams for nonlinearities which are frequency sensitive or which cannot be conveniently separated from frequency sensitive elements. Included in this table are the describing functions for several of the nonlinear elements, and graphs are given in Figs. 17-19.

The block diagrams can be mechanized on an analog computer and, in general, if extensive investigation is needed, a computer solution is recommended. See the instruction manuals of the particular computer to be used for details on computer circuits.

TABLE 5. DESCRIPTION OF TYPICAL COMPLEX NONLINEARITIES

- Type of Nonlinearity
- Motor velocity limiting
    - $K_a$  = acceleration constant
    - $K_e$  = viscous damping constant
    - $J$  = motor-load inertia
    - $T$  = motor time constant
    - $T = J/K_e K_a$
    - $M' = M/K_e K_a$ , normalized input
    - $N_L$  = limiting velocity
  - Motor acceleration limiting
    - $K_a, K_e, J, T, M'$  same as above
    - $\dot{N}_L$  = limiting acceleration
    - $T\dot{N}_L$  = torque limit referred to normalized input  $M'$
  - Backlash (d-c shunt motor driving load)
    - $J_M, J_L$  = motor and load inertia respectively referred to same speed; gearing inertia divided between motor and load, depending upon location of backlash

2-3-30

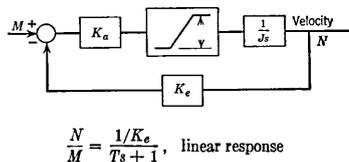


Describing Functions  
Fig. 17

- Assumptions
- Linear machine performance except for limiting.
  - Limiting of motor speed is abrupt, i.e., limiting is absolute and does not occur gradually.

General Remarks

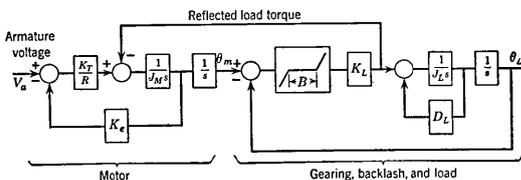
Limiting velocity appears as increased damping; i.e., an effective increase in  $K_e$  and therefore an apparent decrease in motor gain. In general the decreasing gain and phase shift of this type of limiting tend to make the system more stable but sluggish when limiting occurs. The block diagram can be extended to higher order systems. Note that in modifying the block diagram to suit a different situation when a variable is limited, the derivatives of that variable must go to zero in a physical system. The describing function for the simple one energy storage motor shown in the block diagram has been determined experimentally in Ref. 16 and is given in Fig. 17. The configuration for velocity limiting is applicable to hydraulic, d-c and a-c servo motors and can be easily mechanized on an analog computer.



- Linear machine performance except for limiting.
- Limiting of motor acceleration is abrupt.

Limiting acceleration appears as an increase in motor inertia to torque ratio; i.e., a decrease in  $K_a/J$  and therefore an apparent increase in motor time constant. However, because the damping is not affected, the motor low-frequency gain ( $1/K_e$ ) is unchanged. The resulting increase in phase shift with this type of limiting can lead to serious performance deterioration. Torque limiting slows the initial response and also the rate of correction for overshoots. The latter can result in large overshoots. The describing function of the simple one energy storage motor has been determined experimentally in Ref. 16 and is given in Fig. 18. The effect of acceleration limiting can be estimated by simply modifying  $K_a$  to account for the limiting. The gain change can be obtained by using the describing function for saturation, Table 4.

Fig. 18



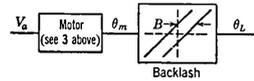
- No gearing bounce.
- Only viscous damping.
- Lumped parameter representation of gearing, motor and load (see column 1).

The large number of variables in the complete problem precludes the derivation of simple describing functions. Describing functions have been derived for a simplified configuration assuming no viscous damping and either (a) a unidirectional Coulomb friction force (see Ref. 14) or (b) a  $J_m \gg J_L$  so that the motor motion is not affected by reflected load torques (see Ref. 15). Relationships have also been determined experimentally (see Ref. 53). The difficulty of handling the complex describing functions, the restrictions of the basic assumptions, and the desirability of using multiple feedbacks in

...

$K_L$  = effective spring constant between motor and load  
 $D_L$  = effective load damping with respect to fixed reference  
 $B$  = backlash angle  
 $K_T, R, K_e$  = motor torque, impedance and velocity, constants

4. Simplified backlash (high load damping)



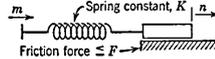
25-31

5. Simplified Coulomb friction

$$\frac{N_1}{M} = G_D$$

$$= \frac{1}{1 + j \frac{4F}{\pi K |N_1|}}$$

$N_1$  = fundamental of output



1.  $J_m \gg J_L$ , i.e., reflected load torques have a negligible effect.
2.  $\sqrt{J_L K} / D_L \ll 1$ , i.e., high load damping so that the load does not coast.
3. Instantaneous acceleration of the load up to speed when the gears engage.

1. Zero load mass,  $J_L = 0$ .
2. Zero viscous damping.
3. Friction force equal and opposite to applied force up to a maximum of  $F$ .
4. No effect on driving motion of reflect load forces.

practical systems limit the usefulness of this approach. The approximations mentioned in the text or below in 4 and 5 usually suffice for preliminary studies, and a computer study is usually necessary for a more thorough investigation. See Sect. 7 for a more detailed discussion of backlash effects.

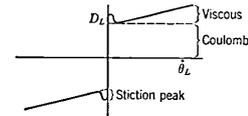
The block diagram can be modified to include (a) Coulomb friction by making  $D_L$  a nonlinear function, (b) speed and acceleration limiting by modifying the motor blocks as shown in 1 and 2 above, (c) multiple backlash by properly dividing the inertia, springiness, etc.

This presentation is useful for many types of instrument servos where there is little coasting by the load. Use the hysteresis describing function, Table 4, for backlash where  $H$  = total backlash angle.

Table 4  
Fig. 12

This describing function is useful when Coulomb friction is a major influence on performance (see Ref. 17). Magnitude and phase are the same as for a simple time constant where frequency  $T\omega = 4F / (\pi K |N_1|)$ . A more exact analysis can be made by using analog computers and introducing friction as a nonlinear damping term,  $D_L$ , i.e.;

Fig. 19



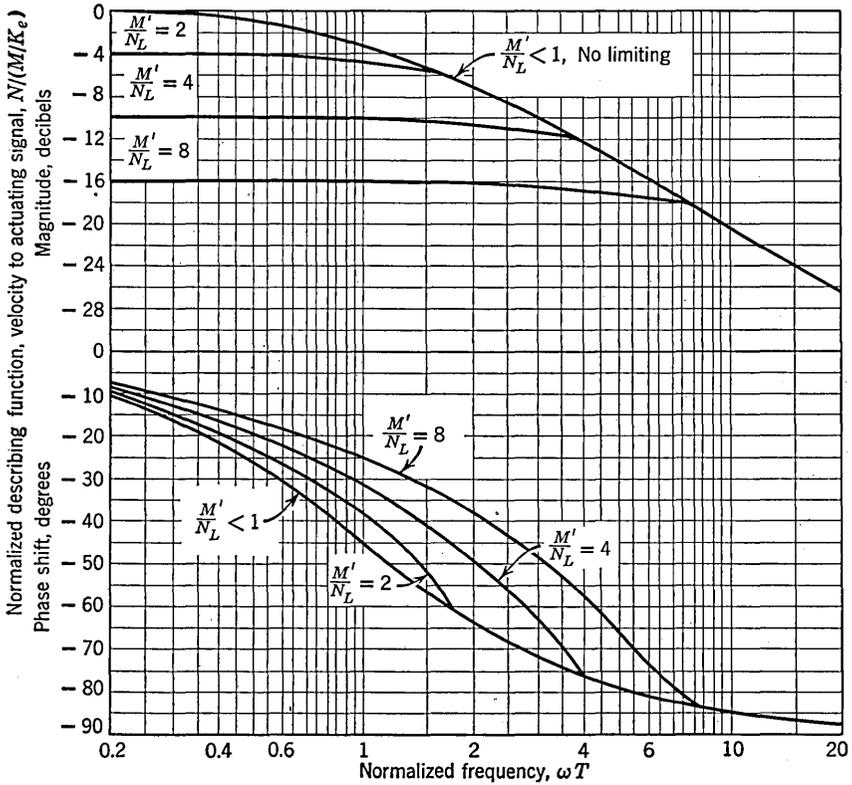


FIG. 17. Describing function of a single servomotor subject to velocity limiting;  $M'$  = normalized magnitude of input, radians per second;  $N_L$  = saturated speed, radians per second (maximum speed);  $T$  = unsaturated motor time constant, seconds (Ref. 16).

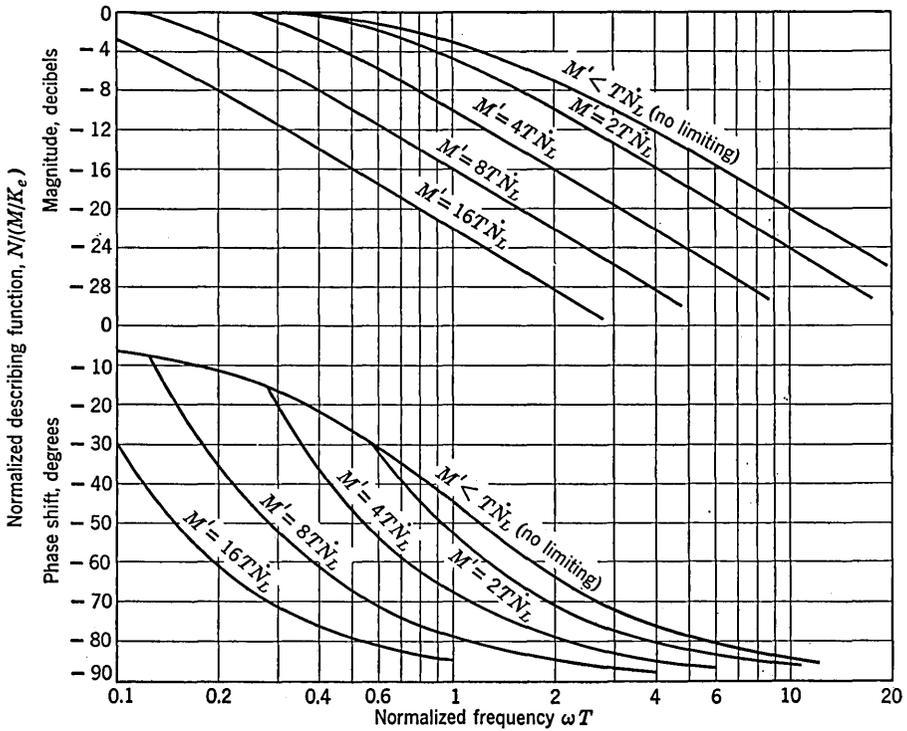


FIG. 18. Describing function of motor subject to acceleration limiting (Ref. 16).

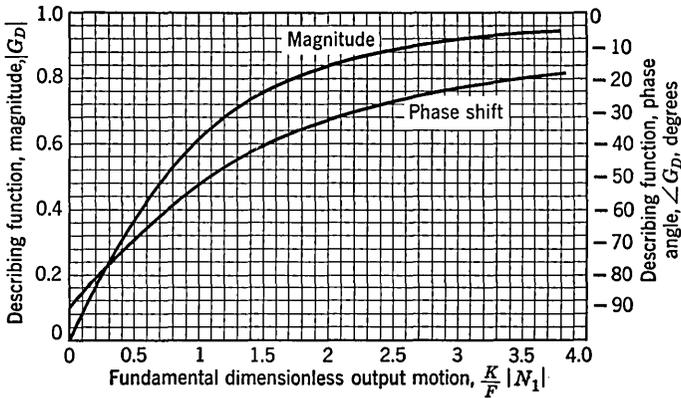


FIG. 19. Describing function for Coulomb friction. See Table 5 for definition of terms.

**Simplifying Complex Nonlinearities**

**Separation of Amplitude and Frequency Sensitive Elements.** It is obvious that the nonlinearities of Table 5 can be simplified by a number of approximations to the point where the simpler amplitude sensitive describing functions can be used. Note that many of the complex nonlinearities consist of a combination of a simple nonlinearity and frequency sensitive element(s). Typical of such a case is the system with backlash shown diagrammatically in Fig. 20a.

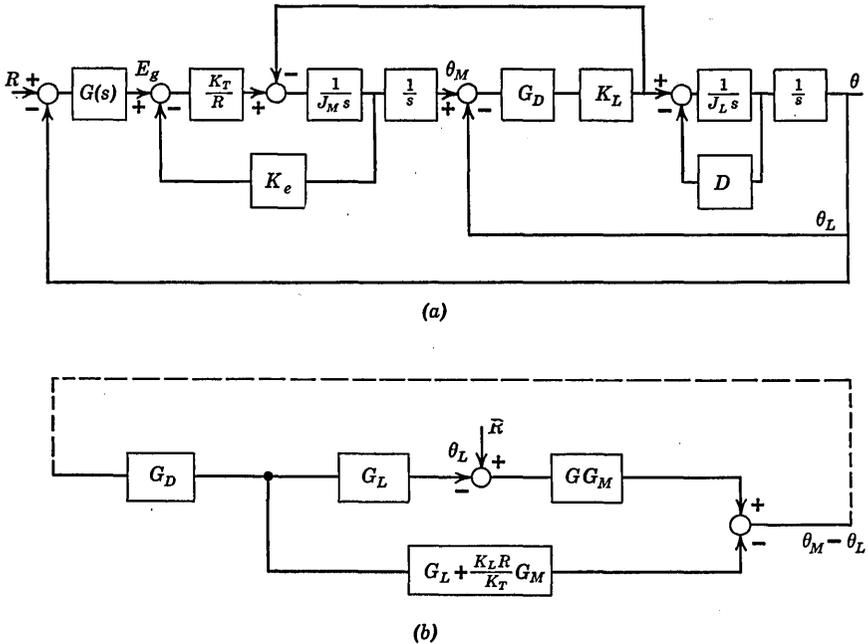


FIG. 20. Block diagram for a simple system with backlash  $G_D$  and  $G(s)$  equal to controller elements: (a) the conventional diagram arrangement; (b) the diagram rearranged so that the amplitude and frequency sensitive portions of the system are separated.

If it is recognized that the input to the nonlinear portion of the system will be essentially sinusoidal, then the describing function for the dead-band (Table 4, Fig. 11) can be used and substituted directly into the block diagram as a gain  $G_D$ . To analyze the system conveniently it is necessary to rearrange the block diagram so that the nonlinearity is separated from the frequency sensitive elements and can be treated by the usual techniques of graphically presenting amplitude sensitive describing functions.

Let

$$G_M = \frac{1}{K_e s \left( \frac{J_m R}{K_T K_e} s + 1 \right)},$$

$$G_L = \frac{K_N / D_L}{s \left( \frac{J_L}{D_L} s + 1 \right)}.$$

Then rearrange the block diagram to appear as in Fig. 20b. In this illustration the amplitude and frequency variant portions of the systems have been separated and the usual type of describing function analysis can be used to determine if the system will be stable and if not what the amplitude and frequency of oscillation will be.

**Method of Equivalent Coefficients.** In complex systems, it becomes difficult and sometimes useless to attempt to separate the amplitude sensitive and frequency sensitive elements in the system so that the method of analysis described earlier in this section under Theory of Describing Functions can be used. For instance in the above example the variables in which one is really interested,  $C$  and  $R$ , are buried in the block diagram. It is thus difficult to determine: (1) whether the overall system will have satisfactory performance, and (2) how to modify the compensation to improve performance.

The technique of using an *equivalent coefficient* is: (a) to recognize that many nonlinearities appear as a gain change in the system, and (b) to combine this gain change with existing gains in the system to form an equivalent gain or coefficient. Once knowing the range of gain to be experienced, the system can be designed to be as insensitive to such a variation as is desirable.

A way to avoid the difficulty in the previous example lies in this approach. It was pointed out that the describing function of the dead space can be considered directly in the analysis. This describing function is in series with the spring constant and can be combined to yield an equivalent spring. Thus, the system will seem to have a very soft spring at low angular displacements and an increasing spring constant (approaching the actual spring constant) as the angular displacement increases. This equivalent coefficient can then be considered as a constant in the remaining analysis. In this case the major effect is the reduction in the load resonant frequency and it becomes necessary to make the system less sensitive to load resonant frequency to avoid difficulties.

It is necessary to consider a number of different spring constants to make sure that no unstable points exist, but this is usually not too difficult a task although it can be somewhat time consuming. (See Ref. 18.)

**EXAMPLE.** As shown in item 3 of Table 5, the equation for backlash from armature voltage to output shaft rate is

$$\frac{s\theta_L}{V_a} = \frac{K'_L K_T / R}{J_M J_L s^3 + \left[ J_M D_L + J_L \frac{K_e K_T}{R} \right] s^2 + \left[ (J_M + J_L) K'_L + \frac{D_L K_e K_T}{R} \right] s + K'_L \left[ D_L + \frac{K_e K_T}{R} \right]}$$

where  $K'_L$  = the equivalent spring constant =  $K_L G_D$ . The value of  $G_D$  is obtained from Fig. 11 with the argument  $B/2(\theta_M - \theta_L)$  rather than  $B/2M$ .

The complete system transfer function including the above equations can then be analyzed for several values of  $K'_L$ . The actual value of  $G_D$  has to be considered only if the magnitude of the input to the backlash is wanted.

## 5. METHODS OF ANALYSIS: PHASE PLANE, GRAPHICAL SOLUTION OF SYSTEM EQUATIONS

### General

This is essentially a heuristic presentation of the phase plane method. As a consequence attention will be directed at the areas of application and only the most rudimentary explanation of the techniques of constructing the diagrams will be provided. At its present state of development this technique has only limited utility in system synthesis; however, phase plane techniques have received some use in the conception and display of schemes for nonlinear compensation. (see Refs. 22 and 23.)

The reader is referred to Refs. 20, 21, and 27 for details beyond those provided here.

**Definitions.** The *phase plane* has the coordinates of velocity (usually the ordinate) and position (usually the abscissa) of the system. The solutions of the differential equations are plotted on this coordinate system.

The locus of a solution to the differential equation is called a *phase trajectory* or simply *trajectory*. A series of solutions or trajectories is referred to as a *phase portrait*.

**Limitations.** Analysis by the phase plane method is limited to:

1. Second order (single degree of freedom) systems.
2. Autonomous systems (time does not appear as a parameter in any of the coefficients of the system).
3. Systems with impulse, step, or ramp inputs or driving functions.

The limitation on the order of the system is severe, but it is possible to approximate a limited number of practical systems by a second order equation for purposes of preliminary analysis. Methods have been proposed for extending the technique to higher order systems but have not received wide use. (See Refs. 24 and 25.)

**Basic Equations.** The basic equations that can be treated by phase analysis are of the form:

$$(6) \quad \frac{d^2x}{dt^2} + f_1(x, \dot{x}) \frac{dx}{dt} + f_2(x, \dot{x})x = 0, \quad \dot{x} = \frac{dx}{dt}.$$

By substituting  $dx/dt = y$ , eq. (6) can be reduced to a set of first order equations:

$$(7) \quad \begin{aligned} \frac{dy}{dt} &= N(x, y), \\ \frac{dx}{dt} &= D(x, y), \end{aligned}$$

and eliminating time by division yields

$$(8) \quad \frac{dx}{N(x, y)} = \frac{dy}{D(x, y)}.$$

### Significant Characteristics of Phase Portrait

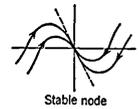
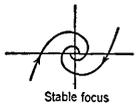
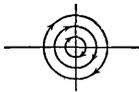
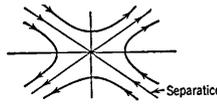
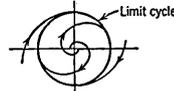
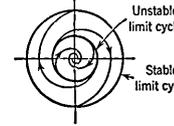
Table 6 describes a few of the significant characteristics of phase trajectories. Identifying these characteristics will be useful to the engineer in interpretation of the phase trajectories.

### Areas of Use of Phase Plane Method

**Analysis.** The graphical techniques of plotting the phase trajectories make the phase plane method particularly useful for systems with second order nonlinear equations of motion. Although the availability of analog computers has greatly reduced the need for such hand methods, there still remains a need for generalizing analysis. The phase plane analysis often can provide this generalization.

**Presentation of Data.** The phase plane has found some use in presentation of analog or actual equipment results. In such cases, the system does not need to be limited to second order. Of course the interpretation becomes more difficult the higher the order of the system. Such plotting techniques have been made even more meaningful when the display is on a cathode ray oscillograph by intensity modulation. Timing pulses can be indicated by brightening or dimming the trajectory.

TABLE 6. SIGNIFICANT CHARACTERISTICS OF PHASE PORTRAIT

Type	Description	Typical Trajectories	Corresponding Conditions in Linear Second Order System
1. Nodal point	The trajectories converge or radiate from the node in such a manner that the direction of the trajectory approaches definite limits as the nodal point is approached. The node is stable if the paths converge on the node and unstable if the paths radiate from the node. This is a singular point, i.e., a point where eqs. (7) are equal to zero.	 Stable node	Stable node, negative real roots Unstable node, positive real roots
2. Focal point	The trajectories converge or radiate from the focus on <i>spiral</i> paths. As for the node, if the paths converge, the focus is referred to as stable; if the paths spiral outward, the focus is referred to as unstable. This is a singular point.	 Stable focus	Stable focus, complex roots with negative real parts Unstable focus, complex roots with positive real parts
3. A center	Closed trajectories about a point. This is a singular point.		Zero damping
4. Saddle point	Trajectories converge toward the saddle point and then diverge except for the special case when the initial conditions are such as to fall on the trajectory that goes into the point (the converging <i>separatrix</i> ). This is a singular point.	 Separatrix	Both roots real, one negative and one positive
5. Limit cycles	A limit cycle describes the oscillation in a nonlinear system. A stable limit cycle is a closed path to which adjacent trajectories converge. When the trajectories diverge from a closed path in the phase plane, the path is called an unstable limit cycle.	 Limit cycle	None
6. Hard oscillations	It is necessary to excite the system beyond a finite bound in order to obtain self-sustained oscillations. The boundary will be an unstable limit cycle.	 Unstable limit cycle	None
7. Soft oscillations	Self-sustained oscillations can be started with an infinitely small excitation. Soft oscillations start from unstable nodes or focal points.	 Stable limit cycle about an unstable focal point	None
		See figure for limit cycle	None

Once the analyst has set up the equations for the phase plane, an analog computer can be used to plot the trajectories. In this manner, one can maintain the generality of the phase plane analysis and avoid the ennui of extensive hand calculation.

**System Synthesis.** Because the graphical presentation often makes interpretation of results easier, the phase plane method has been looked on with favor by many. A number of authors describing work on "optimum controls" have made extensive use of the phase plane in presenting their results. (See Sect. 7 for details.) However, the limitations on the order of the system equations hamper work on any but the simplest systems.

Ku and a number of others have extended the phase plane to *phase space*. (See Refs. 24 and 25.) This is essentially a multidimensional plot allowing solution of higher order equation. Phase space methods have received only very limited use to date.

### Analytical Methods of Constructing Phase Plane

**Direct Method of Solution.** If the equation of motion of the system, eq. (6), can be integrated to obtain time solutions for  $\dot{x}$ , then  $\dot{x}$  as a function of  $x$  can be obtained by eliminating time from the individual solutions, and the relationship between  $x$  and  $\dot{x}$  may be plotted directly.

**Indirect Method of Solution.** If the equations of motion cannot be integrated to obtain time solutions for  $x$  and  $\dot{x}$ , a new differential equation in terms of  $d\dot{x}$  and  $dx$  may be formed and solved to give  $\dot{x}$  as a function of  $x$  directly. The equations in this case reduce to the form of eq. (8).

**EXAMPLE.** *Simple Relay Servo.* Equations (13), Sect. 6, describing the operation of a relay servo for a step input are repeated here for convenience.

$$\frac{d^2x}{d\tau^2} + \frac{dx}{d\tau} = -1, \quad x > 0;$$

$$\frac{d^2x}{d\tau^2} + \frac{dx}{d\tau} = 1, \quad x < 0.$$

Substitution of  $y = dx/d\tau$ , then dividing by  $y = dx/d\tau$ , and recognizing that  $(dx/d\tau)/(dy/d\tau) = dx/dy$  yield:

$$(9) \quad \frac{dy}{dx} = -\frac{1}{y} - 1, \quad x > 0;$$

$$\frac{dy}{dx} = +\frac{1}{y} - 1, \quad x < 0.$$

The variables can be separated in these equations and integrated:

$$(10) \quad \begin{aligned} x &= -\int \frac{y}{1+y} dy, & x > 0; \\ x &= \int \frac{y}{1-y} dy, & x < 0. \end{aligned}$$

These integrals can be found in any good table of integrals, and the function can be plotted on the phase plane for different constants of integration. When plotted, the trajectories of eq. (10) would appear similar to the sketch in Fig. 21.

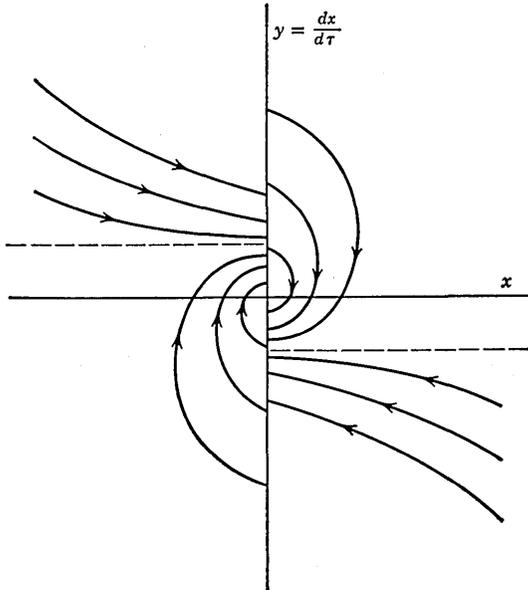


FIG. 21. Phase trajectories for a simple relay servo. Each trajectory is for particular constant of integration of eq. (10).

**Obtaining Time from a Phase Plane Plot.** It is possible to obtain time,  $t$ , from a phase plane plot even though the original characteristic equation of motion cannot be solved for  $x$  and  $\dot{x}$  as functions of time. To do this use the relationship:

$$(11) \quad \tau = \int d\tau = \int \frac{dx}{dx/d\tau} = \int \frac{1}{\dot{x}} dx.$$

Equation (11) shows that if the phase portrait is replotted with  $1/\dot{x}$  as the ordinate and  $x$  the abscissa, the area under the resultant curve re-

presents time. This method makes it possible to obtain plots of  $\dot{x}$  and  $x$  versus time. Graphical methods are also available for obtaining time from the phase portrait. (See Ref. 26.)

### Graphical Methods of Constructing Phase Plane

When the original characteristic equation of motion is nonlinear, the integration of the equation obtained by the above method is difficult or impossible. Graphical methods exist for solving the equation for a direct plot of  $x$  versus  $\dot{x}$ . One of the most useful methods is the *method of isoclines*, described in Refs. 20 and 23. Other graphical methods are also available, e.g., Lienard's, arc-segment procedures.

**Method of Isoclines.** Equation (8) can be written in the form

$$(12) \quad \frac{dx}{dy} = \frac{N(x, y)}{D(x, y)}.$$

Equation (12) is the slope of the phase trajectory. By setting eq. (12) equal to a constant, the equation for the locus of a constant slope can be obtained. One can then strike off lines of the proper slope along the locus. After constructing sufficient loci of constant slope, the phase trajectories can be sketched.

**EXAMPLE.** *Simple Relay Servo.* Equations (9) define the loci of constant slope for the relay servo given in the previous example.

$$(9a) \quad \frac{dy}{dx} = -\frac{1}{y} - 1, \quad x > 0;$$

$$(9b) \quad \frac{dy}{dx} = \frac{1}{y} - 1, \quad x < 0.$$

Equation (9a) set equal to a constant provides the loci in the right half-plane. Equation (9b) describes the left half-plane.

For a  $+45^\circ$  slope eq. (9a) becomes:

$$1 = -\frac{1}{y} - 1, \quad y = -\frac{1}{2}.$$

Several values are tabulated in Table 7. The isoclines are constructed in Fig. 22.

TABLE 7. VALUES OF ISOCLINES FOR SIMPLE RELAY SYSTEM

Slope	Value of $y$	
	Left Half-Plane	Right Half-Plane
0°	1	-1
+30°	0.634	-0.634
-30°	2.36	-2.36
+45°	0.5	-0.5
-45°	$\pm\infty$	$\pm\infty$
+60°	0.366	-0.366
-60°	-1.37	1.37
+75°	0.211	-0.366
-75°	-0.366	0.211
+90°	0	
-90°		0

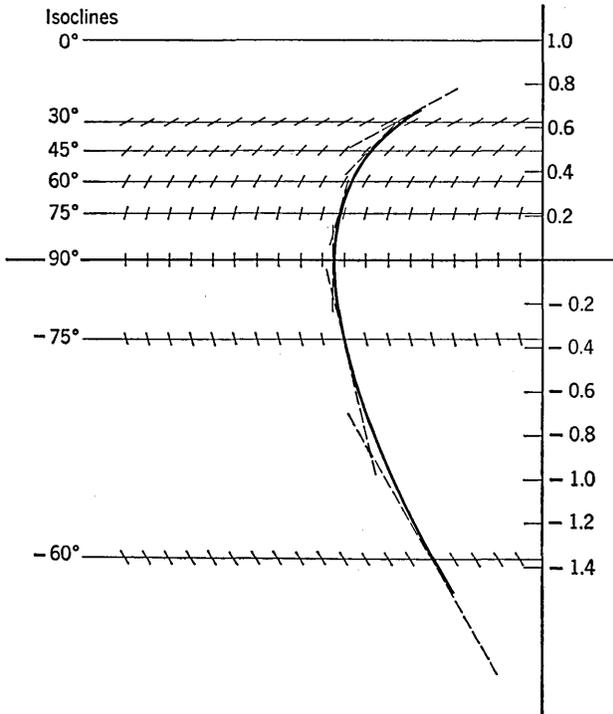


FIG. 22. Construction of a phase trajectory by means of isoclines for a simple relay servo.

## 6. OTHER METHODS OF ANALYSIS

### Differential Equations: Analytical Solutions

An often useful method of analytically obtaining the transient performance of a simple but useful class of nonlinear systems is by *piecewise linearization*. Although the type of nonlinearity that can be treated is restricted, higher order systems can be handled. (See Refs. 31, 32, and 33.) Other analytical methods of obtaining transient solutions are described in Sect. 5 and in Refs. 27, 29, and 30 and their bibliographies.

**Piecewise Linear Systems.** Many nonlinear control systems are linear in well-defined areas of operation. At the boundaries of these linear areas are discontinuities which make the system, when considered as a whole, nonlinear. For such systems the linear differential equations can be solved between boundaries and the boundary conditions matched to obtain a complete solution.

Since it is generally desirable to obtain a solution under steady-state conditions (or at least as steady-state conditions are approached), the process using differential equations becomes quite laborious if there are a number of reversal points. This is true even for a simple second order system, and the process becomes more unwieldy for higher order systems where more than two initial conditions are required at each reversal point.

**Normalized Performance Charts.** Kahn avoided some of the labor in the differential equation approach by using a semigraphical approach that recognizes the fact that the initial conditions at each boundary point are a function of the velocity at the previous boundary and the time between boundaries. (See Ref. 32.) This method loses its value for systems of an order higher than second. Under these conditions more than two dimensions are needed to represent the curves. For a higher order system, it is necessary to have more than one initial condition for each boundary; for *example*, on a third order system, it would be necessary to have initial conditions representing both velocity and acceleration. The higher derivatives would fall in the third dimension.

### Summary of Steps in Piecewise Linear Analysis.

1. Prepare the complete system equations.
2. Break the complete equations into a set of linear equations representing the system operation between discontinuities.
3. Determine the boundary conditions at the discontinuities.
4. Nondimensionalize the equations as much as practical.
5. Rearrange the equation or change the dependent variable to make the dependent variable independent of the input function at the discontinuities.

6. Obtain the solutions to the equations. These may often be best presented graphically.

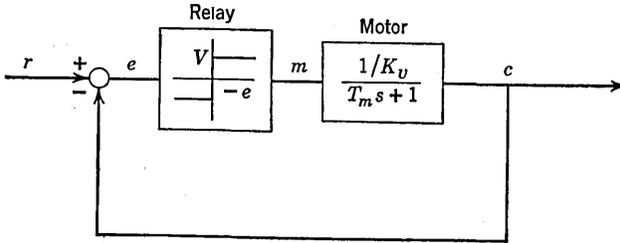


FIG. 23. Block diagram of simple relay servo.

EXAMPLE. *Piecewise Linear Relay Servo.* A typical block diagram for a relay servo is shown in Fig. 23. For the motor of the system of the figure,

$$m = K_v T_m \frac{d^2 c}{dt^2} + K_v \frac{dc}{dt}.$$

From the characteristics for a simple relay

$$m = -V \text{ for } e < 0,$$

$$m = V \text{ for } e > 0.$$

Therefore,

$$K_v T_m \frac{d^2 c}{dt^2} + K_v \frac{dc}{dt} = -V, \text{ for } e < 0;$$

$$K_v T_m \frac{d^2 c}{dt^2} + K_v \frac{dc}{dt} = V, \text{ for } e > 0;$$

$$e = r - c.$$

A typical transient to a step input for the servo of Fig. 23 appears in Fig. 24. The driving voltage in the motor is reversed at each of the zero

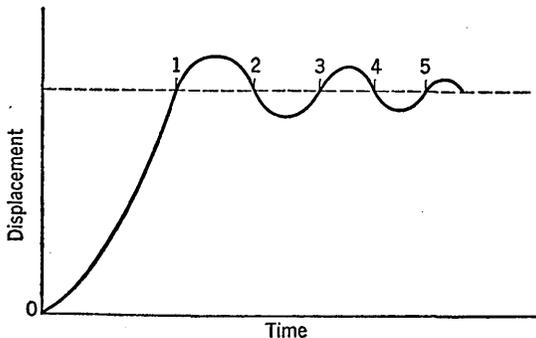


FIG. 24. Typical response of simple relay servo to a step command.

error points 1, 2, 3, 4, 5, etc. Substitution of  $c = r - e$  and normalizing the variables by substituting

$$\tau = t \left( \frac{1}{T_m} \right),$$

$$x = e \left( \frac{K_v}{T_m V} \right)$$

yield for a step input:

$$(13) \quad \frac{d^2x}{d\tau^2} + \frac{dx}{d\tau} = -1, \quad \text{for } x > 0;$$

$$\frac{d^2x}{d\tau^2} + \frac{dx}{d\tau} = 1, \quad \text{for } x < 0.$$

The solutions to these equations are obtained by taking the Laplace transform of eqs. (13) and determining the inverse transform. The transforms of eqs. (13) after the switching at  $\tau_n$  are

$$(14a) \quad s^2X(s) + sX(s) = [-1/s + sx(\tau_n) + x(\tau_n) + \dot{x}(\tau_n)]e^{-\tau_n s}, \quad x < 0;$$

$$(14b) \quad s^2X(s) + sX(s) = [1/s + sx(\tau_n) + x(\tau_n) + \dot{x}(\tau_n)]e^{-\tau_n s}, \quad x > 0;$$

where  $x(\tau_n)$  = value of  $x$  at  $\tau_n$ ,

$$X(s) = \mathcal{L}[x(\tau)],$$

$\dot{x}(\tau_n)$  = derivative of  $x$  with respect to  $\tau$  at  $\tau_n$ ,

$\tau_n$  = normalized time of the  $n$ th switching point.

Notice that with the exception of the first closure of the relay (in the region 0 - 1 of Fig. 24),  $x(\tau_n) = 0$  at the switching point and only  $\dot{x}(\tau_n)$  affects the characteristics of the transient. The velocity just before the relay switches must equal the velocity just after the relay has switched. This is the boundary condition relating the two eqs. (13). To obtain a solution to eqs. (13), it is then necessary to apply the initial conditions at each reversal point.

For eq. (14a) the inverse transform yields, when  $x(\tau_n) = 0$ ,

$$(15) \quad x = -\exp[-\tau + \tau_n] - (\tau - \tau_n) + 1 + \dot{x}(\tau_n)[1 - \exp(-\tau + \tau_n)],$$

$$\tau_{n+1} > \tau > \tau_n,$$

where  $\tau_{n+1}$  = normalized time at the  $n + 1$  switching point.

Equation (15) is dependent only on the time from the last reversal and the initial velocity. The time to reach the next reversal can be calculated by setting eq. (15) equal to zero and solving for the time difference  $\tau_{n+1} - \tau_n$ . This equation is

$$(16) \quad \dot{x}(\tau_n) = \frac{\exp(-\tau_{n+1} + \tau_n) + \tau_{n+1} - \tau_n - 1}{1 - \exp(-\tau_{n+1} + \tau_n)}.$$

The velocity at the next reversal can be obtained from the derivative of eq. (15) at  $\tau_{n+1}$ :

$$(17) \quad \dot{x}(\tau_{n+1}) = \exp(-\tau_{n+1} + \tau_n) - 1 + \dot{x}(\tau_n) \exp(-\tau_{n+1} + \tau_n).$$

Equation (14b) can be solved similarly. Since the system is symmetrical, the equations corresponding to eqs. (16) and (17) will be respectively identical except for sign.

Kahn avoided the labor of obtaining repetitive solutions to eq. (15) by plotting eqs. (16) and (17) as shown in Fig. 25. The transient  $x(\tau)$

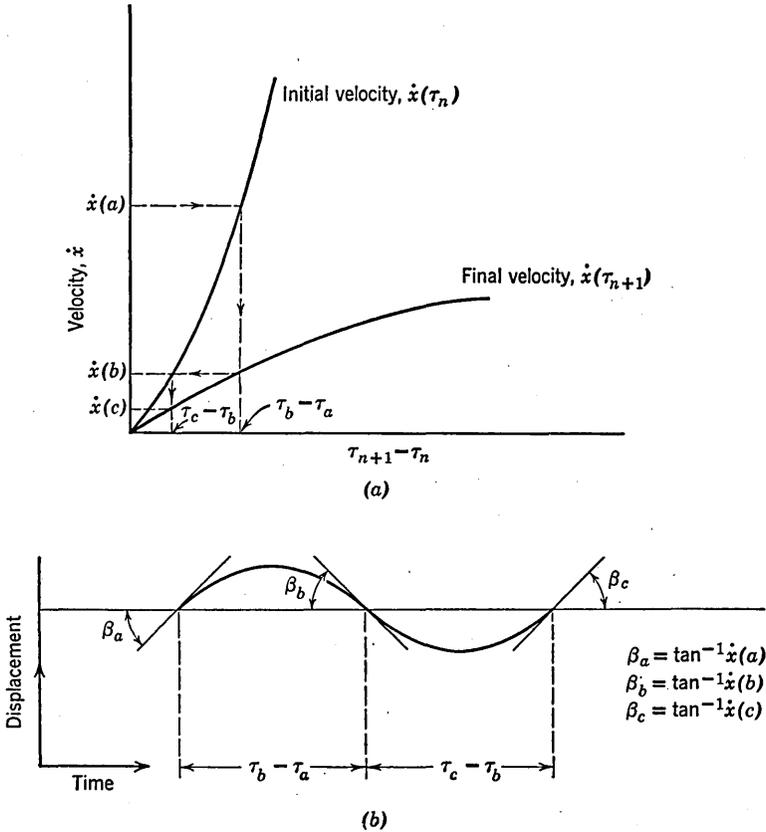


FIG. 25. Plots of piecewise linear relay servo (a) performance charts and (b) method of sketching time response from performance chart data.

can be quickly sketched by laying off the slopes,  $\dot{x}(\tau_n)$ , at the switching times,  $\tau_n$ . The method can be extended to more complex nonlinear functions than described, but practically the system must be limited to second order.

**Methods for Analysis for Variable Linear Systems**

A differential equation with *time-varying coefficients* can be linear; however, the useful frequency response and root locus methods are primarily applicable to *constant coefficient* linear differential equations and cannot be extended simply to the case with time-varying coefficients. This section presents several special techniques which can be used with variable linear systems.

**Equivalent Coefficients for Variable Linear Systems.** This method is useful where the variations are slow with respect to system response.

The designer can reduce the system to a series of constant coefficient cases as follows: Consider the general case of eq. (18).

$$\begin{aligned}
 (18) \quad & (a_{n0} + a_{n1}t + \dots) \frac{d^n x(t)}{dt^n} \\
 & + \dots + (a_{10} + a_{11}t + \dots) \frac{dx}{dt} + (a_{00} + a_{01}t + \dots)x \\
 & = (b_{m0} + b_{m1}t + \dots) \frac{d^m y(t)}{dt^m} + \dots + (b_{10} + b_{11}t + \dots) \frac{dy}{dt} \\
 & \qquad \qquad \qquad + (b_{00} + b_{01}t + \dots)y,
 \end{aligned}$$

where the  $a_{ik}$ ,  $b_{ik}$  are the constants for  $t$  to the  $k$ th power in the coefficient of the  $i$ th derivative.

To investigate the performance at an instant of time,  $t_1$ , substitute  $t_1$  for  $t$  in the coefficients. The resulting constant coefficient linear differential equation can be treated by the usual linear constant coefficient techniques.

$$\begin{aligned}
 (19) \quad & A_n \frac{d^n x(t)}{dt^n} + \dots + A_1 \frac{dx(t)}{dt} + A_0 x(t) \\
 & = B_m \frac{d^m y(t)}{dt^m} + \dots + B_1 \frac{dy}{dt} + B_0 y(t),
 \end{aligned}$$

where  $A_n = a_{n0} + a_{n1}t_1 + a_{n2}t_1^2 + \dots$ ,  
 $B_m = b_{m0} + b_{m1}t_1 + b_{m2}t_1^2 + \dots$ .

**Ultimate Stability of Variable Linear System.** Kirby pointed out a method of exactly determining the ultimate stability of eq. (18) as  $t \rightarrow \infty$ . (See Refs. 34 and 35.) Basically one forms a new equation from the coefficients of the highest order of  $t$  in the coefficient of each derivative, i.e.,

$$(20) \quad a_{nk}s^n + \dots + a_{1k}s + a_{0k} = 0.$$

The stability of eq. (20) can then be determined by the Routh or another stability criterion. To be applicable the order of  $t$  in the coefficient of the highest derivative must be of as high an order as any  $t$  in the coefficients of the remaining derivatives. If the same order of  $t$  does not exist in the coefficient of a derivative, then the value of  $a$  is considered to be zero for the corresponding term in the new equation. It should be noted that the Routh (or similar stability criterion) test for stability then *cannot* be applied directly to eq. (20).

Because it indicates only the ultimate stability as  $t \rightarrow \infty$ , the method is limited.

**Adjoint Computing Method.** The adjoint computing method is valuable in cases where (a) the system is linear time-varying and (b) the performance is wanted at only one instant of time as affected by disturbances at other times.

This is a powerful analytical tool and is particularly well suited for use on analog computers.

The method is described in Chap. 24.

## 7. NONLINEAR SYSTEM COMPENSATION

This section covers (a) compensation of a nonlinear system by linear and nonlinear elements and (b) introduction of intentional nonlinearities to improve the response of an originally linear system.

Because of the difficulty of generalizing a nonlinear system, the majority of the published work has been either on second order systems or in terms of a specific application. For this reason, in the following, emphasis will be placed upon the ideas involved rather than on the details of the method of analysis.

### Compensation of Slowly Varying Nonlinearities

Missiles, engines, and many industrial processes have wide variations of system parameters. These variations are often slow compared with the other variations of the controlled quantities about a particular operating point.

Such changes in system characteristics are usually the result of the variation of an independent variable. This is the case in a ground-launched missile. The aerodynamic coefficients vary with altitude and speed, and since these coefficients determine the transfer function, the missile dynamic characteristics will vary as the missile climbs and accelerates. Although the change in gain from control surface motion to missile attitude can be several thousand to one, these changes may be relatively slow compared with the response of the system controlling the missile attitude.

Conversely, a slow variation in characteristics can be the result of the

change in the dependent variable. For instance in a turbojet engine speed regulator, the fuel to the engine is varied to maintain speed. The transfer function of the engine varies widely with engine speed—as much as 10 to 1 from idle to top speed. However, it normally requires several seconds to accelerate an engine of this type through this speed range which is slow compared with the speed regulator response.

**Technique for Compensation.** For systems which are of the slowly varying nonlinear type, the necessary control characteristics can be determined by the following process:

1. Reduce the system equations to linear, constant coefficient form by the methods of Sects. 2 and/or 6 at a particular point in the range of operation.
2. Synthesize a satisfactory control transfer function at the particular point by the methods of Chap. 23.
3. Repeat the procedure at a sufficient number of points through the operating range to define thoroughly the variation of the control transfer function that is necessary to maintain ideal performance over the range of operation.
4. Determine the simplest method of manipulating the control transfer function and still satisfactorily approximate the values determined in 3.

The change required in the control may be a simple gain change or it may require the modification of time constants. Considerable simplification may result from an ingenious choice of the method for obtaining the control changes. For instance, the variation of the missile characteristics was with altitude and speed. The control characteristics could therefore also be changed with altitude and speed; however, it may prove simpler to program the control changes as a function of time. If the missile path is known in advance, the change in altitude can be rather closely predicted as a function of time. For the engine control, it is obvious after some parameter study that the gain variation of the engine is approximated by an exponential function of fuel flow or a function of compressor pressure. Solutions commonly adopted, therefore, use either a nonlinear fuel valve in the controller or manipulate controller gain with compressor pressure to compensate for the variations in engine gain. The particular method selected depends upon many factors involved in the mechanization of a particular control. (See Ref. 36.)

**Practical Considerations.** If the rate of change of the coefficients approaches the response of the system, the actual system must be analyzed in more detail. The system can become unstable if the proper combination occurs. Of equal importance, if the variation is rapid enough, the system can actually operate satisfactorily through a zone which is unstable if analyzed on a small change fixed system basis. In the latter case,

the analysis based on the assumption of a slowly varying nonlinearity would probably result in a stable system. However, the result of such an analysis would probably be an overdesigned system.

### **Compensation of Rapidly Varying Nonlinearities**

A nonlinearity which changes system characteristics at a rate which is appreciable with respect to the response time of the system is defined as a *rapidly varying nonlinearity*.

All physical systems have limitations and/or restraints. These may take the form of limitations on power, torque, amplifier output, speed, etc., or they may be simply the result of limited precision and appear as backlash, deadband, hysteresis, etc. Relay servos and other systems of this class represent a special case wherein the nonlinearity is introduced intentionally. The purpose of such a system may be, as is the case of the relay servo, to simplify the system, or the nonlinearity may be necessitated because of other system considerations. Regardless of the reason for their presence in the system, if these nonlinearities have a marked change in characteristics with variations of an independent variable in the range of interest, they are rapidly varying and must be considered as an integral part of the analysis of the system.

Rapidly varying nonlinearities can also be a function of the independent variable; however, *the following applies primarily to nonlinearities which are functions of the dependent variable*.

**Compensation with Linear Elements: Use of Describing Function Techniques.** Methods of predicting the necessary system compensation by frequency response techniques which are applicable to linear systems have been extended to nonlinear systems with limited success. There are certain significant differences between the frequency response of a linear and a nonlinear system. For a linear system, the closed loop response is defined by a single set of values and these can be uniquely related to the transient response. Design criterion in terms of the characteristics of the frequency response can therefore be directly interpreted into terms of the transient response. For a nonlinear system in which the nonlinear element has been replaced by the describing function, a different closed loop frequency response is obtained for each value of  $G_D$ , and therefore a single frequency response does not represent the system during a transient which passes through the nonlinear region. Even assuming that the describing function adequately represents the nonlinearity, it is difficult under such conditions to specify a satisfactory design criterion in the frequency domain which is significant in terms of the transient response.

*No generally satisfactory frequency response design criterion for nonlinear systems presently exists.*

However, since the values of the gain and phase margin and  $M_m$  and  $\omega_m$  have received considerable use in linear analysis, a knowledge of the variation of these quantities is a valuable link between the frequency response methods and the describing function methods. Although of more qualitative than quantitative use, it is often valuable to plot these quantities versus the magnitude of the input variable.

EXAMPLE. If the  $M_m$  appears too severe, the frequency locus can be adjusted to give the desired characteristic. Shown in Fig. 26 are the  $\omega_m$

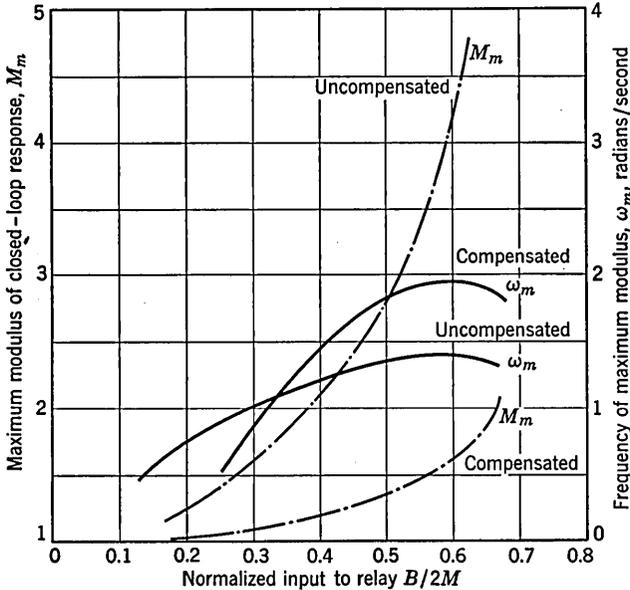


FIG. 26. The maximum modulus  $M_m$  and the frequency of the maximum modulus  $\omega_m$  of a simple relay servo as a function of the argument of the describing function for a relay. The operating conditions and the servo frequency response are given in Fig. 27. The uncompensated characteristics result from setting the system gain sufficiently high to meet sensitivity requirements, i.e., setting  $B$ . Maintaining the system gain and adding a 3 to 1 lead gives the *compensated* characteristics. (Note that in practice there would have to be a 3 to 1 increase in amplification to compensate for the 3 to 1 attenuation that a passive lead network would have.)

and  $M_m$  of a relay servo before and after compensation. Figure 27 shows how  $G(j\omega)$  locus has been changed to improve system response by meeting a criterion of a maximum  $M_m \approx 2$ . The boundary of such an improvement can be quickly obtained by overlaying a Nichols chart with a graph of the  $G_D$  locus and observing the path of the  $G(j\omega)$  locus on the Nichols chart as the origin is moved along the  $G_D$  locus. (See Fig. 28.) If a maximum  $M_m$  criterion is being used, the boundary of this  $M_m$  for all values

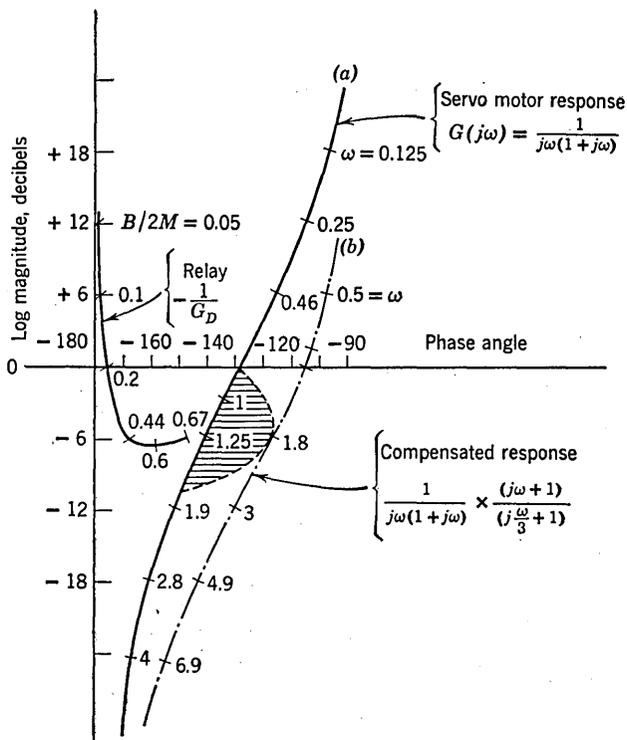


FIG. 27. Log magnitude-angle diagram of a simple relay servo; operating conditions  $V/B = 6$  db;  $H/B = 0.5$ . Curve (a) is the uncompensated system, Curve (b) is the system compensated with a 3 to 1 lead. The cross-hatched region is the necessary modification to the uncompensated system to meet an  $M_m$  criterion of  $M_m < 2$ . For the purposes of illustration a normalized response has been used for the servo motor and only the ratio  $V/B$  has been defined.

of  $G_D$  can be quickly sketched on the  $G(j\omega)$  graph. If sufficient work of this type is performed, templates for several values of  $M_m$  can be built. The necessary compensation networks can be determined either by trial and error or by more elaborate methods discussed in Chap. 23.

**Compensation for Relay Servomechanisms.** Describing function, phase plane, and piecewise linear analyses have all been used extensively to determine the necessary compensation. (See Refs. 11, 18, 24, 31, 32, and 44.) The describing function method is the most useful for higher order systems. The techniques used in the preceding example can be easily extended to higher order systems. *The describing function analysis and experiment normally check within engineering accuracy.* (See Refs. 11 and 41.) A maximum  $M_m < 2$  criterion is generally typical in the design

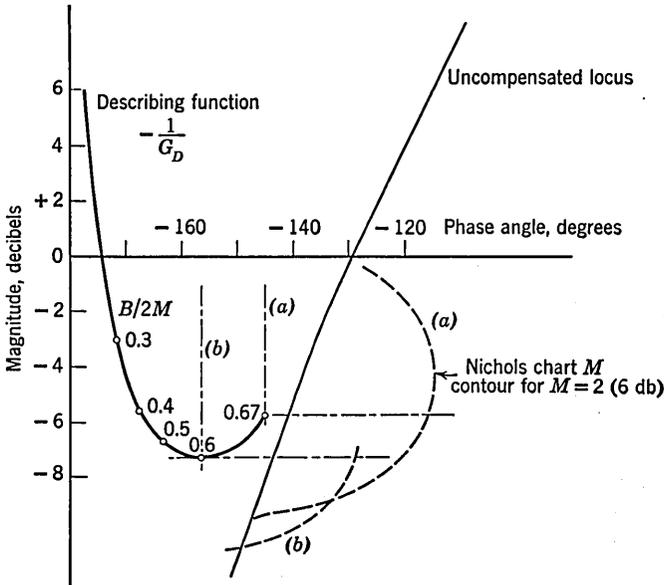


FIG. 28. Log magnitude-angle diagram showing a method of estimating the necessary compensation by overlaying with a Nichols chart. The origin of the Nichols chart is first placed at  $B/2M = 0.67$  and the  $M$  contour is sketched on the log magnitude-angle diagram. This is noted as curve (a), where  $M_m = 2$  has been used for illustrative purposes. The complete area of necessary compensation can be found by moving the origin of the Nichols chart along the describing function locus. Another location is shown at  $B/2M = 0.6$ , and the  $M$  contour is curve (b). The coordinates of the Nichols chart are shown as dotted center lines.

of relay-positioning systems. However, experience with the condition of the particular application may dictate a different value for maximum  $M_m$  or a different criterion. Because of the phase lag of a relay with hysteresis, lead compensation is quite useful. Two forms of such compensation are tandem lead networks and rate feedback. The latter can be obtained either from a tachometer or from motor back electromotive force. Nonlinear compensation can also be used to achieve better performance for a *particular type of input*. See the paragraphs on Optimum Switching Functions later in this section for details. Such compensating networks must be used with care if more than one type of input is to be encountered, because the performance will vary with the form and magnitude of the input function. (See Refs. 42, 43, and 45.)

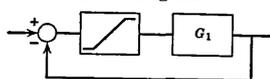
**Compensation for Saturation.** For a large number of systems, it is necessary to follow a relatively smooth input within a very small error. In order to provide economical components, the linear operating range of these components is usually very little beyond that necessary to follow

TABLE 8. TYPICAL TYPES OF SATURATION EFFECTS AND METHODS OF COMPENSATION

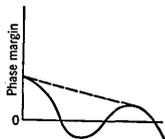
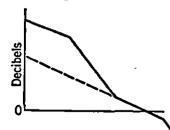
Type of Non-linearity  
Case I, pre-amplifier saturation

System Configuration and Characteristics

Block diagram

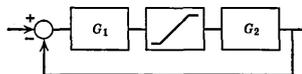


Open loop characteristic



25-54

Case II, power amplifier saturation



$G_1$  = integral compensation

$G_2$  = power element

Same open loop characteristics as Case I

Effect on System Performance

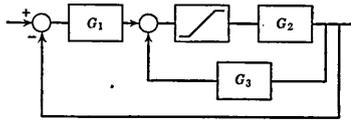
For an unconditionally stable system with saturation, the relative response will be slower for large step inputs. As the step input level is increased for a conditionally stable system, the system will begin to exhibit less stable characteristics until a critical value is reached above which self-sustained oscillations will occur. For moderate saturation, the overshoot may actually be less than for the linear system although the settling time will be increased. Saturation acts like a gain reduction in the system and for saturation to cause system instability, instability must be predicted on a linear basis for reduced gain. Oscillation frequency and amplitude can be estimated within 20 to 30% by describing functions. The presence of noise with the input signal causes an effective increase of saturation beyond that predicted for the input signal by itself. This effect causes an increase in the closed loop phase shift with respect to the input signal. See Ref. 7 and Table I.

Same effect as Case I but a reduction in the overshoot with moderate saturation has not been observed. The degree of saturation (ratio of saturation level to input to element if the system were linear) necessary to start self-sustained oscillations in a system will vary with the location of the nonlinearity in the system. A difference as great as 10 to 1 between the saturation from a step of  $r$  needed in the pre-amplifier and power amplifier has been noted. (See Ref. 18.) In all the cases considered in Ref. 18, it was necessary to have sufficient saturation so that the gain was reduced to the point where there was negative phase margin at gain crossover; however, with preamplifier saturation it was necessary to exceed this level of saturation considerably to cause self-sustained oscillations.

Possible Methods of Compensation

- (a) Eliminate or reduce the integral compensation for large signal inputs; e.g., typical magnitude and phase angle curves are shown at the left. It is obvious that if the gain is reduced sufficiently, the region of negative phase margin will cause instability. However if a nonlinear compensating network is used which for large errors eliminates or reduces the integral compensation, satisfactory performance can be obtained. The dotted line shows the frequency response after such a change. There are a number of methods by which such changes in the compensation can be obtained. (See Refs. 18, 38, and Table 10.) The circuit constants are normally set experimentally.
- (b) Modify the basic system operation for large signal levels. This is essentially an extension of (a). However the basic mode of operation is also changed. Examples are dual-mode servos wherein the mode of controlling the power element is changed with signal level, and two-speed servos wherein the feedback signal gain is lowered. (Actually, in practice, the takeoff is from a different speed shaft which gives rise to the appellation two speed.) (See Refs. 39 and 40.) Often the signal used to switch the feedback signal is used to modify the compensation networks.
- (a) See Case I (a) and (b). Power amplifier saturation is similar to torque saturation for which the dual servo techniques have been developed. (See Refs. 23, 40, and Table 10.)
- (b) Use of tachometer feedback around the saturation is effective. (See Case III.)
- (c) If in place of  $G_1$  the integral compensation can be accomplished by a filter (lead) network around the saturating element, the system can be so designed that, as the system saturates, the compensation automatically becomes less and the system will not become unstable with saturation. (See Ref. 2.)

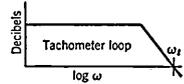
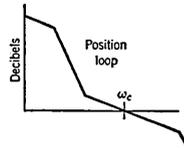
Case III,  
power  
ampli-  
fier  
satu-  
ration



$G_1$  = integral compensation

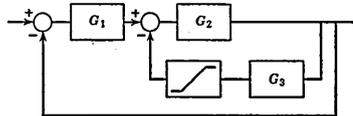
$G_2$  = power element

$G_3$  = tachometer feedback



25-55

Case IV,  
satu-  
ration  
in feed-  
back



Same open loop characteristics as Case III

For a conditionally stable system saturation, the gain is reduced in the tachometer loop lowering the crossover frequency,  $\omega_t$ , which lowers the phase margin at the position loop. There are two possibilities: (1) The phase margin at the normal position loop crossover frequency,  $\omega_c$ , will be lowered until the system becomes unstable at the normal crossover frequency,  $\omega_c$ . (2) The tachometer loop becomes ineffective before (1) occurs, and the position loop gain will be lowered forcing the crossover frequency down into the region where the phase margin goes negative on account of the integral compensation. The effect depends upon the constants of the particular system being considered. For (1) the oscillation frequency will be approximately the crossover frequency. For (2) the frequency of oscillation will be closer to the integral compensation time constants.

Effect similar to Case I.

- (a) Instability is not normally a serious consideration.
- (b) The problem can generally be avoided by achieving the compensation by a filter in the tachometer feedback rather than in tandem elements in  $G_1$ .

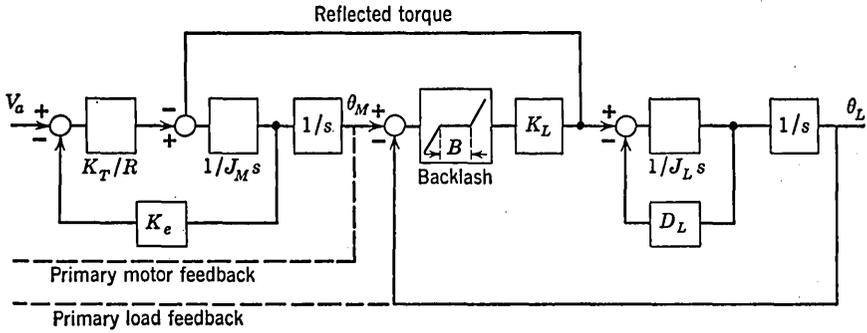
- (a) Eliminate saturation if possible.
- (b) See Cases I(a) and (b) and II(c) above.

the input within the maximum allowable error. When such systems are synchronized on a new operating condition or subjected to violent disturbances, they will inherently be highly saturated. This leads to reduced performance. In addition, because such systems often use integral compensation to achieve high values of low-frequency gain, this can also lead to serious overshoots and the attendant longer settling times or even instability. Normally the requirement on allowable error is not as necessary during the synchronizing period, and a reduction in performance can be tolerated. The major concern is, therefore, that the system should settle rapidly and stably from large signals. The effects of saturation and the methods of compensation for such systems are summarized in Table 8.

**Compensation for Backlash.** The effects of backlash and load resonance are the major limits on the performance that can be achieved with a power servomechanism. The great quantity of published material attests to the serious consideration that has been given to the problem. (See Refs. 13, 14, 15, 18, and 46 to 52.) However, thoroughly satisfactory methods for circumventing the effects of backlash are not available.

The *basic effects* (see also Table 5) are illustrated by the system of Fig. 29. For large input signals, the backlash is quickly taken up and has very little effect upon performance. For low-level signals approaching the magnitude of the backlash, the backlash tends noticeably to disconnect the load from the motor during signal reversals. Heuristically it can be seen that this will cause the load to *lag* farther behind the motor than with a linear system. Conversely because the motor is disconnected from the load, it will accelerate faster than normal in the backlash zone and the motor position will therefore tend to *lead* the normal response. When considered in terms of frequency response, if the primary feedback is from the load, the effect of backlash increases the lagging phase shift and decreases the loop gain; if the primary feedback is from the motor, the effect of backlash on system performance is much less severe and it actually introduces a leading phase shift into the loop. For low signal levels, if the load damping is viscous, the linearized equations of Fig. 29 can be used. For low signal levels if there is appreciable Coulomb friction present, its effects will predominate and the use of hysteresis to represent the backlash is more correct than the equations of Fig. 29. Therefore, it is necessary to evaluate carefully the type and extent of the damping present. If the damping is viscous, the frequency of oscillation caused by backlash will generally be at or higher than the normal linear gain crossover frequency. If the damping is of the Coulomb type, the frequency of oscillation will be lower than the gain crossover frequency. The amplitude in either case will be small (one to several times the backlash angle, normally).

*Methods of analysis.* Phase plane, piecewise linear, and describing function methods of analysis have been used successfully. The describing



Basic linearized equations:

$$\frac{s\theta_L}{V_a} = \frac{(1/K_e)a\omega_L^2\omega_m}{s^3 + [2\delta + \omega_m]s^2 + \left[ \left(1 + \frac{J_L}{J_M}\right)a\omega_L^2 + 2\delta\omega_m \right]s + a\omega_L^2 \left[ 2\delta \frac{J_L}{J_M} + \omega_m \right]}$$

$$\frac{s\theta_M}{V_a} = \frac{(1/K_e)\omega_m[s^2 + 2\delta s + \omega_L^2]}{s^3 + [2\delta + \omega_m]s^2 + \left[ \left(1 + \frac{J_L}{J_M}\right)a\omega_L^2 + 2\delta\omega_m \right]s + a\omega_L^2 \left[ 2\delta \frac{J_L}{J_M} + \omega_m \right]}$$

where  $a = |G_D|$ , magnitude of describing function for deadband;

$$\omega_m = \frac{K_e K_T}{J_M R}, \text{ motor time constant without load;}$$

$$\omega_L^2 = \frac{K_L}{J_L}, \text{ load mechanical resonant frequency;}$$

$$\delta = \frac{D_L}{2J_L}, \text{ load viscous damping.}$$

Other constants are defined in Table 5.

FIG. 29. Typical shunt d-c machine and load with backlash representation. The basic linearized transfer functions are given in a nondimensionalized form.

function methods are the most generally useful for a paper study. However, the complexity of the problem warrants the use of an analog computer for thorough investigations.

As just noted and in Table 5, the representation used for backlash will vary depending upon the constants involved. If the hysteresis representation is chosen, the usual describing function methods of analysis can be used. If the more complex representation of Fig. 29 is chosen, the method of equivalent coefficients, Sect. 4, is recommended.

**General Design Considerations.** In the design of a power servomechanism, the following basic effects should be given consideration:

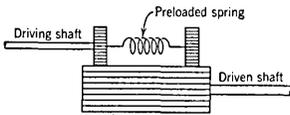
1. Use of tandem integral compensation increases the magnitude of sustained oscillations. The effect can be reduced by the use of dead space.

TABLE 9. USEFUL CORRECTIVE

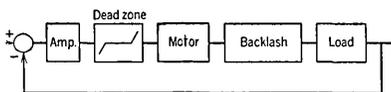
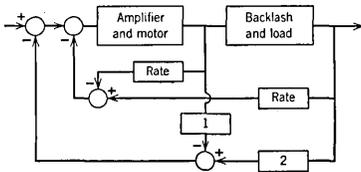
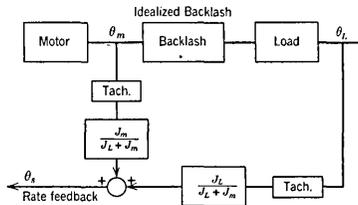
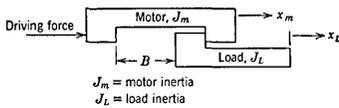
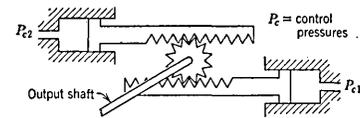
Type of Corrective Measure	Techniques for Backlash Compensation
1. Mechanical design	<p>(a) Improved precision. The backlash can be reduced by improving the grade of gears used and the tolerance on the center distances, by having adjustable center distances, or by numerous other special design and/or assembly procedures.</p> <p>(b) Spring loading. There are various methods for mechanically spring loading the gear trains to take up the backlash. One method used in lightly loaded gear trains is shown at the right. This can be extended to the point where the entire gearing is completely divided and one-half spring loaded against the other half. This takes up the backlash throughout the entire gear train.</p> <p>(c) Split drive. By using two driving motors biased in opposite directions and separate gear trains for each motor, the motors will drive against each other to the extent of the bias and take up any backlash. A hydraulic drive of this type is shown at the right.</p> <p>(d) One speed motor. Eliminate the gearing by driving the load directly from the motor shaft. (See Ref. 52.)</p>
2. Divided reset	<p>(a) Solid motion feedback. Consider the idealized backlash at the right. There must be a point with a displacement somewhere between the displacements of <math>J_m</math> and <math>J_L</math> that responds only to the externally applied forces. The displacement of this point (<i>center of mass</i>) is called the <i>solid motion</i>. All supplementary motions of <math>J_m</math> and <math>J_L</math> relative to the solid motion must then be due to mutual forces that occur upon collision or separation and the momentum of the supplementary motions must be equal and opposite. Instrumenting the solid motion would give a signal which did not contain the backlash effects. This point cannot be physically instrumented but by adding signals from the load and motor in the proper proportion the supplementary motions can be cancelled and only the solid motion will remain. From the principles of conservation of momentum:</p> $\dot{x}_s = \frac{J_M}{J_M + J_L} \dot{x}_m + \frac{J_L}{J_L + J_M} \dot{x}_L,$ <p>where <math>\dot{x}_s</math> is the rate of change of the solid motion. (See Refs. 14, 15, and 49.) A method of instrumenting the technique is shown at the right.</p> <p>(b) Artificial damping. Reducing the load-resonant peak and increasing the apparent load-resonant frequency ameliorates the backlash effects. This can be done by the proper feedback of the relative load and motor motions. This includes position, velocity, and acceleration differences. A configuration containing such a feedback is shown at the right. The circuit constants can be determined by frequency response techniques by assuming the system to be linear as in Fig. 29. Experimental adjustment will be necessary.</p>
3. Network compensation	<p>(a) Dead zone compensation. Backlash oscillations are low amplitude. Use of a dead zone in the error channel opens the loop to low amplitude signals and will stop certain types of backlash oscillation. (See Ref. 48.) The dead zone will be the same order of magnitude as the backlash.</p> <p>(b) Frequency-sensitive networks. Compensating networks, gain changes, parallel-tandem networks, etc., can be used in combination with dead-band or separately to get the desired gain-phase change at low signal levels. (See Ref. 28.) Lead networks are particularly effective in compensating for the lagging phase shift of backlash.</p>

## TECHNIQUES FOR BACKLASH

### Typical Diagrams



Spring tension rotates gears in opposite direction until backlash is taken up. Torque is transmitted through springs to shaft.



### General Remarks

There are many other methods besides the two shown for mechanizing the concepts of (b) and (c). (See Ref. 49.) In general, these methods are costly and increase the friction and wear in the gear train.

Electric and hydraulic models of one-speed motors have been built and tested. The low-speed high-torque requirement makes the electric unit bulky and heavy.

General mechanical design considerations are outlined in the text.

When the feedback (or *reset*) is from (or *divided* between) the motor and the load, it is called *divided reset*. The concept as discussed is highly simplified and in practice the configuration and constants will have to be adjusted to suit the particular case. When  $J_m \gg J_L$ , the motor and the center of mass follow closely and rate feedback from the motor alone is effective in damping the system. Approximate schemes for obtaining the rate feedback from the motor back emf, etc., are often adequate. *Solid* motion position feedback can be obtained in the same manner; however, this position signal can differ from the actual load position by as much as  $BJ_m/(J_m + J_L)$ , and often the addition error cannot be tolerated.

Feedbacks of the type in (b) are very effective. Generally, position feedback alone is sensitive to system parameter variations; rate and position feedback in combination are quite insensitive; acceleration feedback is very sensitive. Position difference feedback increases the resonant frequency, and rate difference feedback increases the damping.

This method has been used for drives with a low-load inertia to motor inertia ratio (referred to the same speed), and with sufficient friction to keep the load from much coasting. Under these conditions, it is effective and the error is small. More complex schemes of sensing the proper time to modify the gain characteristics are possible but usually are not justified.

(See Table 9, item 3.) The effects of integral type compensation achieved by tachometer feedback and a lead filter have not been as thoroughly documented. However, the same general tendency is apparent.

2. Increasing the load mechanical resonant frequency,  $(K_L/J_L)^{1/2}$ , reduces the magnitude of the sustained oscillations.

3. It would be desirable to have mechanical load damping ratios greater than 0.1. These would probably be undesirable on larger drives because of the large power loss involved. However, there are methods of increasing the damping electrically (see Table 9).

4. Primary feedback from the motor gives more stable operation than from the load.

Table 9 summarizes various methods of compensating for backlash.

None of the schemes is perfect in the practical case, but all provide a certain relief from the problem. The final choice usually includes considerations of weight, size, and cost, as well as performance.

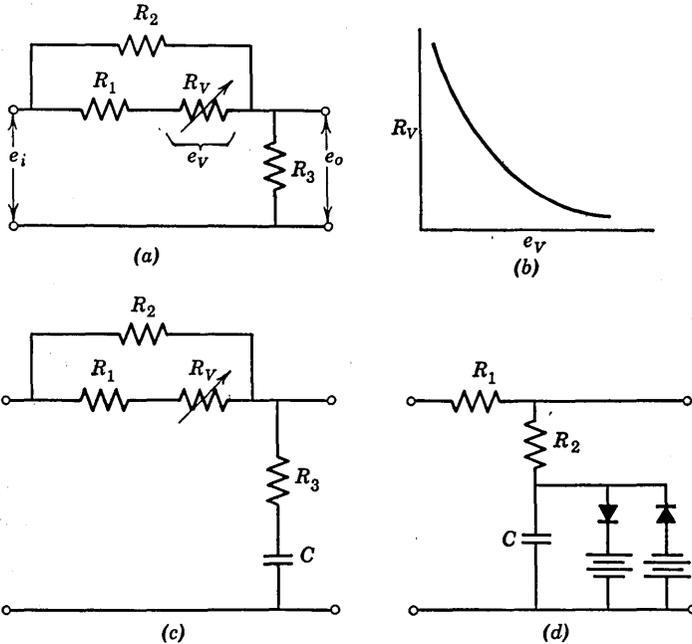


FIG. 30. Typical nonlinear compensating circuits. As shown the circuits vary with the input variable but circuits (a) and (c) can be adapted to vary with an independent variable: (a) nonlinear gain circuit with characteristics that vary with input voltage to increase gain for low-level signals; (b) resistance characteristics of the voltage sensitive resistor  $R_v$ ; (c) nonlinear time constant circuit that reduces the time constant for large input signals; (d) nonlinear time constant circuit that eliminates the time constant and reduces the low-frequency gain for large input signals.

**Nonlinearities to Improve System Response**

Nonlinearities used for improving system performance in general involve methods for (a) reducing the response time and/or (b) minimizing overshoots by more fully utilizing the performance available in the power element(s). Many of these methods accelerate the system rapidly for large errors and increase the relative damping for small errors so that operation is smooth (very stable). Table 10 summarizes several typical methods of nonlinear compensation and refers to typical circuits shown in Figs. 30 and 31.

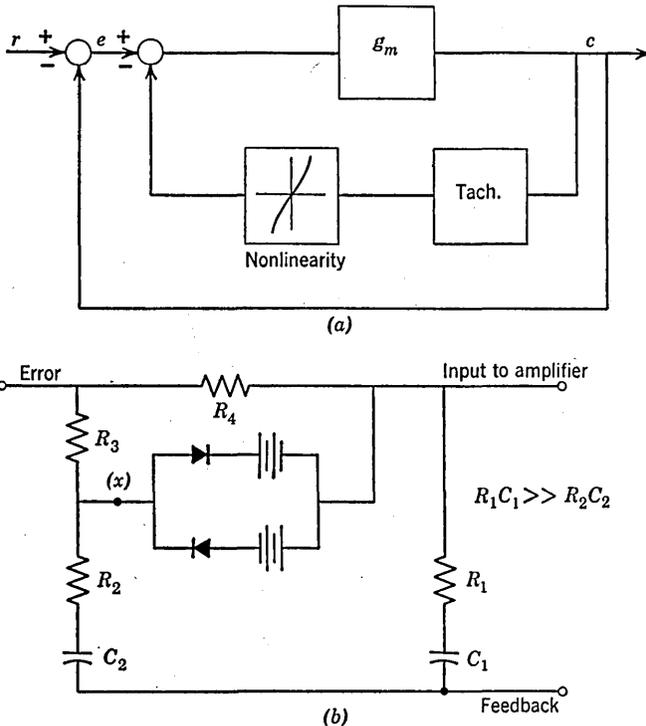
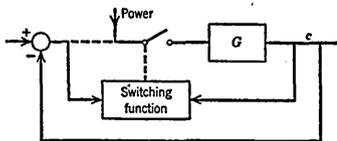


FIG. 31. Typical nonlinear feedback compensating circuits: (a) nonlinear rate feedback to minimize overshoot from large signals; (b) nonlinear stabilizing circuit for switching feedback compensation for large errors or feedback rates. To obtain the proper characteristics it may be necessary to use an isolation amplifier at (x).

Because of the difficulty of specifying the required characteristics mathematically and the impracticality of instrumenting the ideal characteristics for all but the simplest systems, nonlinear compensation is obtained by empirical means in practice.



4. Optimum switching techniques or minimum response time systems



*Optimum switching* is the controlled switching of power to the motor to reduce the error and its derivatives to zero in the minimum possible time, recognizing only the limitations on the performance of the motor. For example, the optimum response to a step input of a second order system with torque limiting is to accelerate at the maximum rate about halfway and then switch and decelerate at the maximum rate for the remaining distance. By proper selection of the switching point the system will arrive at zero error with zero error rate, and if the torque is removed, the system will remain at rest with no further corrective action. See the example in the text. Table 11 gives the optimum switching functions for several second order systems. The number of switching points needed to respond in the minimum time to a step input is  $(n - 1)$  where  $n$  is the order of the system. (See Ref. 24.) Excessive switching at low signal levels can be avoided by having a small deadband at the null. Smoother operation for small signals can be obtained by changing the mode of operation and having a small linear band at null. This has been called *dual mode* operation. (See Ref. 23.)

that for high feedback rates and low errors the switch will open (the diodes stop conduction) and normal stabilization will come into play during synchronization. Under extreme conditions the switch may actually reverse polarity to allow rapid deceleration.

It is difficult to mechanize the optimum switching function for systems higher than the second order. However, the optimum performance can be approached closely without going to the complexity of  $(n - 1)$  switching points. This approximation can be made analytically by deriving a nonlinear function (of one variable) that gives a response that approaches the optimum response. This technique is explained in Ref. 57. The approximation can be arrived at empirically by using the basic second order system switching functions and modifying them by experience and experiment to provide satisfactory performance for higher order systems. The optimum switching technique is not limited to relay servos. The "switching" can be the saturation of some element in the system. In any case the optimum response is obtained only for the designed input; i.e., systems designed for step inputs show poorer response for velocity inputs. Because the optimum response is the minimum time that a power element can make a correction it provides a good basis for rating system performance. The ratio of response time to the optimum response time is a useful index of system performance. (See Ref. 55.)

EXAMPLE. *Optimum Switching Techniques to Obtain Minimum Time Response.* (Refer to Fig. 32.) It is assumed that the amplifier gain is

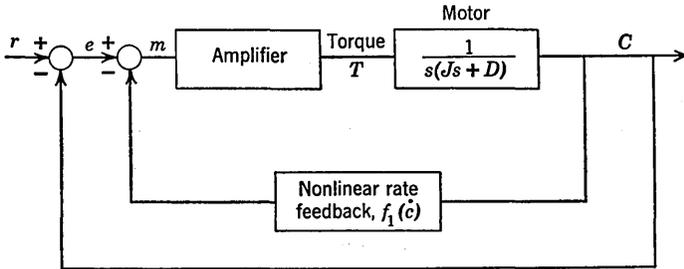


FIG. 32. Nonlinear control system with a very high gain amplifier and torque saturation  $\pm T_m$ .

sufficiently high so that the motor operates with full voltage on it for all but very small errors. The system equations are:

$$\begin{aligned} \pm T_m &= J\ddot{c} + D\dot{c}, \\ (21) \quad m &= e - f(\dot{c}), \\ e &= r - c, \end{aligned}$$

where  $\dot{c} = dc/dt$ ,  $\ddot{c} = d^2c/dt^2$ ,

$$\begin{aligned} +T_m &\text{ for } m > 0, \\ -T_m &\text{ for } m < 0. \end{aligned}$$

For a step input of  $r$ :

$$\begin{aligned} \mp T_m &= J\ddot{e} + D\dot{e}, \\ (22) \quad m &= e - f_1(\dot{e}). \end{aligned}$$

Equation (22) can be solved independent of time to yield a series of trajectories in the phase plane, Fig. 33. The coordinates of this plane are error,  $e$ , and error rate,  $\dot{e}$ .

*There is only one trajectory which passes through the origin, and it will provide the optimum system response if the torque to the motor is reversed when this trajectory is reached.*

From Fig. 33, it is seen that proper choice of the function  $(e, \dot{e}) = e - f(\dot{e})$  will provide the intelligence to perform the necessary switching function. A nonlinear tachometer feedback will then provide the necessary switching information.

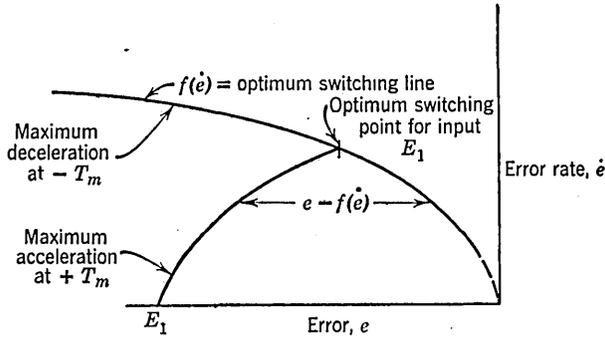


FIG. 33. Phase portrait of the performance of the control system of Fig. 32, showing the optimum switching line where the torque must be reversed to bring the system to rest with no overshoot. When the quantity  $e - f(\dot{e})$  goes to zero, the torque should be reversed.

**Optimum Switching Functions.** The form of the system characteristic equation will dictate what the optimum function should be. Several typical cases as derived in Ref. 56 are given in Table 11.

TABLE 11. TYPICAL OPTIMUM SWITCHING FUNCTIONS FOR SECOND ORDER SYSTEMS WITH LIMITED TORQUE,  $T_m$

System Type	Torque Equation (See Fig. 32)	Optimum Switching Function in the Fourth Quadrant
Undamped	$\pm T_m = J \frac{d^2c}{dt^2}$	$\dot{e} \dot{e}  = \frac{2T_m}{J} e, \quad \dot{e} = \frac{de}{dt}$
Viscous damped	$\pm T_m = J \frac{d^2c}{dt^2} + D \frac{dc}{dt}$	$e = -\frac{T_m J}{D^2} \log_e \left( 1 - \frac{\dot{e} D}{T_m} \right) - \frac{J}{D} \dot{e},$ $\dot{e} = \frac{de}{dt}$
Coulomb damped <sup>a</sup>	$\pm T_m - T_f(\dot{c}) = J \frac{d^2c}{dt^2} + D \frac{dc}{dt}$	$\dot{e} \dot{e}  = 2 \frac{T_m}{J} \left( 1 + \frac{T_f}{T_m} \right) e,$ $\dot{e} = \frac{de}{dt}$

<sup>a</sup>  $T_f(\dot{c})$  is positive for  $c > 0$  and negative for  $c < 0$  and is a constant in either case.

## REFERENCES

1. L. A. MacCall, *Fundamental Theory of Servomechanisms*, Van Nostrand, Princeton, N. J., 1945.
2. J. G. Truxal, *Control System Synthesis*, Chap. 10, McGraw-Hill, New York, 1955.
3. J. C. Lozier, A steady state approach to the theory of saturable servo systems, *I.R.E. Trans. on Automatic Control*, May 1956.
4. K. Klotter, Steady state vibrations in systems having arbitrary restoring and arbitrary damping forces, *Proc. Symposium on Nonlinear Circuit Analysis*, Vol. II, Polytechnic Institute of Brooklyn, New York, 1953.
5. E. Levinson, Some saturating phenomena in servo mechanisms with emphasis on the tachometer stabilized system, *Trans. Am. Inst. Elec. Engrs.*, **72**, Pt. 2, (1953).
6. N. Minorsky, *Introduction to Nonlinear Mechanics*, Edwards, Ann Arbor, Mich., 1947.
7. R. G. Wilson and I. H. Van Horn, *The Effect of Noise on Rate-Limited Systems*, Rept. No. GER2328, Goodyear Aircraft Corp., Feb. 22, 1952.
8. C. A. Ludeke, The generation and extinction of subharmonics, *Proc. Symposium on Nonlinear Circuit Analysis*, Vol. II, Polytechnic Institute of Brooklyn, New York, 1953.
9. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. II, Chap. 10, Wiley, New York, 1955.
10. H. D. Greif, Describing function method of servomechanism analysis applied to most commonly encountered nonlinearities, *Trans. Am. Inst. Elec. Engrs.*, **72**, Pt. 2, 243-248 (1953).
11. R. J. Kochenburger, Frequency response method for analyzing and synthesizing contractor servomechanisms, *Trans. Am. Inst. Elec. Engrs.*, **69**, Pt. 1, 270-284 (1950).
12. E. C. Johnson, Sinusoidal analysis of feedback control systems containing nonlinear elements, *Trans. Am. Inst. Elec. Engrs.*, **71**, Pt. 2, 169-181 (1952).
13. N. B. Nichols, Backlash in a velocity lag servomechanism, *Trans. Am. Inst. Elec. Engrs.*, **72**, Pt. 2, 462-466 (1953).
14. A. Tustin, The effects of backlash and of speed-dependent friction on the stability of closed-cycle control systems, *J. Inst. Elec. Engrs. (London)*, **94**, Pt. 2A, 143-151 (1947).
15. K. N. Satyendra, Describing functions representing the effects of inertia, backlash, and Coulomb friction on the stability of an automatic control system, *Trans. Am. Inst. Elec. Engrs.*, **75**, Pt. 2, 243-248 (1956).
16. R. J. Kochenburger, Limiting in feedback control systems, *Trans. Am. Inst. Elec. Engrs.*, **72**, Pt. 2, 180-192 (discussion), 192-194 (1953).
17. V. B. Haas, Coulomb friction in feedback control systems, *Trans. Am. Inst. Elec. Engrs.*, **72**, Pt. 2, 119-123 (discussion), 123-126 (1953).
18. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. II, Chap. 8, Wiley, New York, 1955.
19. K. Klotter, How to obtain describing functions for nonlinear feedback systems, *Am. Soc. Mech. Engrs.*, IRD Paper No. 56-IRD-5, August 1956.
20. J. J. Stoker, *Nonlinear Vibrations in Mechanical and Electrical Systems*, Interscience Publishers, New York, 1950.
21. T. M. Stout, Basic methods for nonlinear control system analysis, *Am. Soc. Mech. Engrs. Paper No. 56-IRD-9*, July 1956.
22. A. M. Hopkins, A phase-plane approach to compensation of saturating servomechanisms, *Trans. Am. Inst. Elec. Engrs.*, **70**, Pt. 1, 631-639 (1951).

23. D. McDonald, Nonlinear techniques for improving servo performance, *Proc. Natl. Electronics Conference*, Vol. VI, pp. 400-421, National Electronics Conference, Inc., Menasha, Wis., 1950.

24. I. Bogner, and L. F. Kazda, An investigation of the switching criteria for higher order contactor servomechanisms, *Trans. Am. Inst. Elec. Engrs.*, **73**, Pt. 2, 118-126 (discussion), 126-127 (1954).

25. Y. H. Ku, A method for solving third and higher order nonlinear differential equations, *J. Franklin Inst.*, **256**, 229-244 (1953).

26. J. G. Truxal, *Automatic Feedback Control System Synthesis*, Chap. 11, McGraw-Hill, New York, 1955.

27. T. J. Higgins, A résumé of the development and literature of nonlinear control system theory, Am. Soc. Mech. Engrs. Paper No. 56-IRD-4, July 1956.

28. C. H. Shen, H. A. Miller, and N. B. Nichols, Nonlinear integral compensation of a velocity-lag servomechanism with backlash, Am. Soc. Mech. Engrs. IRD Paper No. 56-IRD-3, August 1956.

29. T. M. Stout, A step-by-step method for transient analysis of feedback systems with one nonlinear element, *Trans. Am. Inst. Elec. Engrs.*, **75**, Pt. 2, 378-389 (discussion), 389-390 (1956).

30. J. G. Truxal, Numerical analysis for network design, Approximation Papers, *Trans. I.R.E.*, **PGCT-CT-1**, 4-64, September 1954.

31. H. L. Hazen, Theory of servomechanisms, *J. Franklin Inst.*, **218**, 279-331 (1934).

32. D. A. Kahn, Analysis of relay servomechanisms, *Trans. Am. Inst. Elec. Engrs.*, **68**, Pt. 2, 1079-1088 (1949).

33. J. W. Schwartz, Piecewise linear servomechanisms, *Trans. Am. Inst. Elec. Engrs.*, **72**, Pt. 2, 401-405 (1953).

34. M. J. Kirby, Stability of servomechanisms with linearly varying elements, *Trans. Am. Inst. Elec. Engrs.*, **69**, Pt. 2, 1662-1667 (1950).

35. M. J. Kirby and R. M. Guilianelli, Stability of varying-element servomechanisms with polynomial coefficients, *Trans. Am. Inst. Elec. Engrs.*, **70**, Pt. 2, 1447-1451 (1951).

36. H. Chestnut and R. W. Mayer, *Servomechanisms and Regulating System Design*, Vol. II, Chap. 9, Wiley, New York, 1955.

38. E. S. Sherrard, Stabilization of a servomechanism subject to large amplitude oscillation, *Trans. Am. Inst. Elec. Engrs.*, **71**, Pt. 2, 312-324 (1952).

39. J. C. West, A system utilizing coarse and fine position measuring elements in remote-position-control servo mechanisms, *Proc. I.R.E.*, **99**, Pt. 2, 135-143 (1952).

40. D. McDonald, Multiple mode operations of servomechanisms, *Rev. Sci. Instr.*, **23**, 22-30 (1952).

41. S. K. Chao, Design of a contactor servo using describing function theory, *Trans. Am. Inst. Elec. Engrs.*, **75**, Pt. 2, 223-231 (1956).

42. H. G. Doll and T. M. Stout, Design and analog computer analysis of an optimum third-order nonlinear servomechanism, Am. Soc. Mech. Engrs. Paper No. 56-IRD-10, July 1956.

43. J. C. West and P. N. Nikiforak, The frequency response of a servomechanism designed for optimum transient response, *Trans. Am. Inst. Elec. Engrs.*, **75**, Pt. 2, 234-239 (1956).

44. J. E. Hart, An analytical method for the design of relay servomechanisms, *Trans. Am. Inst. Elec. Engrs.*, **74**, Pt. 2, 83-89 (discussion), 89-90 (1955).

45. R. R. Caldwell and V. C. Rideout, A differential-analyzer study of certain nonlinearly damped servomechanisms, *Trans. Am. Inst. Elec. Engrs.*, **72**, Pt. 2, 165-169 (discussion), 169-170 (1953).

46. A. A. Clark and H. J. Pixley, Effects of non-linearities in multi-loop lead angle prediction systems, Am. Soc. Mech. Engrs. Paper No. 56-IRD-18, July 1956.

47. H. T. Marcy, M. Yachter, and J. Zauderer, Instrument inaccuracies in feedback control systems with particular reference to backlash, *Trans. Am. Inst. Elec. Engrs.*, **68**, Pt. 1, 778-788 (1949).

48. F. J. Ellert, Feedback in contouring control systems, Am. Inst. Elec. Engrs. Second Feedback Control Conference, April 1954.

49. D. C. McDonald, Backlash compensation improves servo system operation, *Instruments and Automation*, **23** [10], 1728-1731 (1955).

50. C. H. Thomas, *Stability Characteristics of Closed-Loop Systems with Dead Band Frequency Response*, 288-305, R. Oldenburger, Editor, Macmillan, New York, 1956.

51. R. L. Hovious, Jitter in instrument servos, *Trans. Am. Inst. Elec. Engrs.*, **73**, Pt. 2, 393-398 (1954).

52. F. M. Bailey, Performance of drive members in feedback control systems, *I.R.E. Trans. on Automatic Control*, **PGAC-1**, May 1956.

53. J. H. Liversidge, Backlash and resilience within the closed loop of automatic control systems, in *Automatic and Manual Control*, A. Tustin, Editor, Butterworths, London, 1952.

54. J. B. Lewis, The use of non-linear feedback to improve the transient response of servomechanisms, *Trans. Am. Inst. Elec. Engrs.*, **71**, Pt. 2, 449-453, (discussion) 453 (1952).

55. R. S. Neiswander and R. H. MacNeal, Optimization of non-linear control systems by means of non-linear feedbacks, *Trans. Am. Inst. Elec. Engrs.*, **72**, Pt. 2, 260-270, (discussion) 270-272 (1953).

56. T. M. Stout, Effects of friction in an optimum relay servomechanism, *Trans. Am. Inst. Elec. Engrs.*, **72**, Pt. 2, 329-335, (discussion) 335-336 (1953).

57. R. E. Kuba and L. F. Kazda, A phase space method for the synthesis of nonlinear servomechanisms, *Trans. Am. Inst. Elec. Engrs.*, **75**, Pt. 2, 282-289 (discussion), 289-290 (1956).

## Sampled-Data Systems and Periodic Controllers

John E. Barnes, Jr.

1. Description and Definition of Sampled-Data System	26-01
2. Methods of Transient Analysis	26-06
3. Sampled-Data System Stability	26-15
4. Sampled-Data System Synthesis	26-20
References	26-32

### 1. DESCRIPTION AND DEFINITION OF SAMPLED-DATA SYSTEM

**Definition of Sampled-Data System.** Systems which operate on data obtained at discrete intervals of time are called sampled-data systems. The information obtained at a particular instant is called the *sample*. Normally the intervals are equally spaced in time and the amplitude of the sample is proportional to the amplitude of the signal.

#### Characteristics of Sampled-Data Systems

**Basic Elements.** Figure 1 shows the basic elements of a sampled-data system: the *sampler* and the *continuous elements*. They may appear in various configurations, and there may be more than one sampler in the system. The output from the sampler is a train of pulses which is denoted by a *starred symbol*; that is, the output of a sampler whose input is  $e(t)$  is written  $e^*(t)$ .

**Linearity.** If the continuous elements are linear, the sampled-data system is linear and the superposition theorem is valid. A sampled-data system has regular time discontinuities, but the techniques of analysis by the use of solutions of the linear constant-coefficient differential equations of the system are directly applicable.

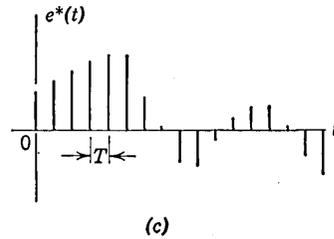
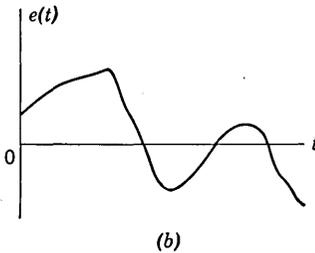
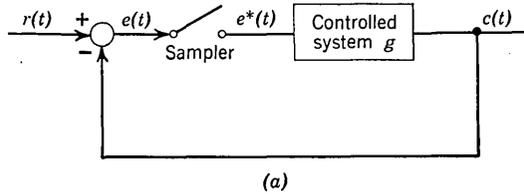


FIG. 1. Sampled-data system and sampler input and output signals: (a) simple sampled-data system; (b) continuous error function; and (c) sampled error function.

**The Sampler.** The sampler acts as a pulse modulator of the input and generates a train of pulses. This action introduces high frequencies into the system which may be attenuated by a linear filter. *The information contained in the input signal may be recovered with reasonable fidelity if the sampling frequency is at least twice the highest frequency component of the input signal.* Figure 2 shows the effect of sampling frequency upon the frequency spectrum of the output of the sampler.

**Use of Samplers.** Sampled-data systems may be used for several reasons:

1. To use a digital computer as part of the controller. The input data *must* be in sampled form.
2. To use simpler, low-powered control elements.
3. To realize the beneficial effect which sometimes accrues when sampled-data systems are used for the process control of plants having inherent dead time.

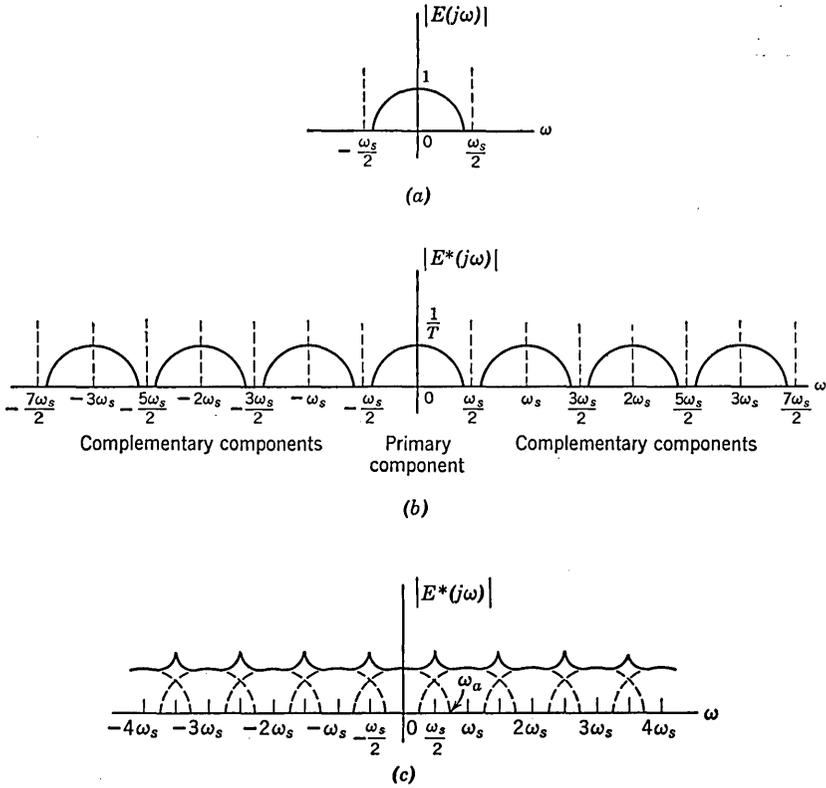


FIG. 2. Sampler transfer characteristics in the frequency domain. (a) Amplitude spectrum of sampler input; (b) amplitude spectrum of sampler output, sampling frequency greater than twice the maximum signal frequency; (c) amplitude spectrum of sampler output, sampling frequency less than twice the maximum signal frequency.  $\omega_s$  = sampling frequency (Ref. 5). Reprinted by permission from J. D. Truxal, *Automatic Feedback Control Systems Synthesis*, Copyright 1955 by McGraw-Hill Book Co.

4. To use pulsed-data information. The input information may be available in discrete samples as in a guided missile control system or as in certain track-while-scan radar systems. Sampled-data systems may be used to advantage where digital sensors are already available.

**Description of Typical Sampled-Data Systems**

**Digital Computer in the Controller.** Figure 3 shows a typical digital system. The sampling and coding unit converts continuous data into pulsed data. The digital computer performs a series of operations on the pulsed data and presents the results in pulsed form to the holding and decoding unit which reconverts the results into (approximately) continuous

signals for use by the continuous control equipment. The feedback may transmit the data in either pulsed or continuous form.

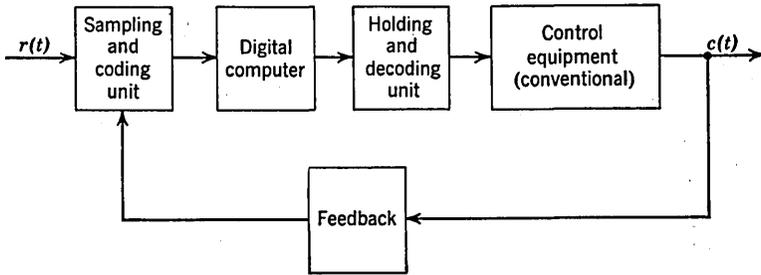


FIG. 3. Typical sampled-data control system. Conventional control equipment is continuous.

In practical operating systems, a typical method of converting from a continuous variable available as a shaft rotation to a binary code number which represents its magnitude and polarity is to use an encoding device such as a circular binary pattern shown in Fig. 4. The circular tracks may

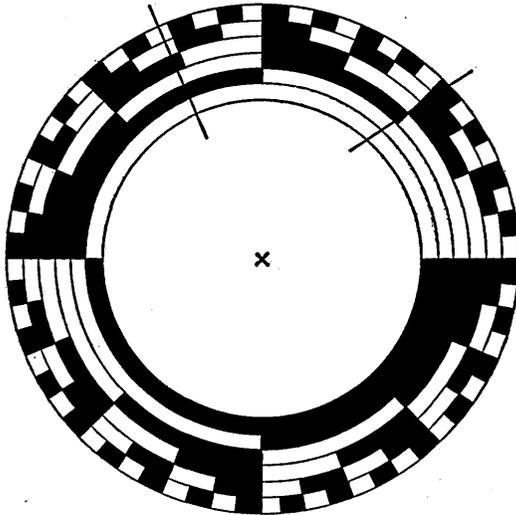


FIG. 4. Circular binary pattern for analog-digital conversion. The lines across the pattern show that accurate angular position of the photocells or brush contacts is necessary to avoid errors in conversion.

be scanned radially with a photoelectric cell or brush pickoffs. The output will be the binary pulse code which represents a particular position of the circular binary pattern; the pattern shown can resolve a circle into  $2^6 = 64$

parts. Although the encoder shown is for angular rotation, devices have been manufactured for conversion of pressures and flows to digital form. Techniques for converting analog voltages to digital form are also available. (See Vol. 2, Chap. 20.)

**Periodic Process Controller.** A typical sampled-data regulator for process control is shown in Fig. 5.

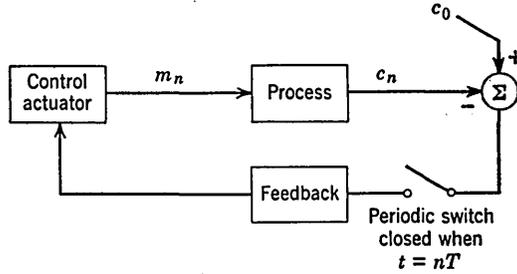


FIG. 5. Typical sampled-data process regulator.

The typical stepwise process controller monitors the controlled variable periodically (every  $nT$ ) and makes a control adjustment at each sensing instant. In process regulation, the usual (and perhaps the most useful) form of control actuator is a servo motor, which serves as a low-pass filter and also serves to reset the error detector. The following description of a periodic controller is taken from Oldenbourg and Sartorius (Ref. 1).

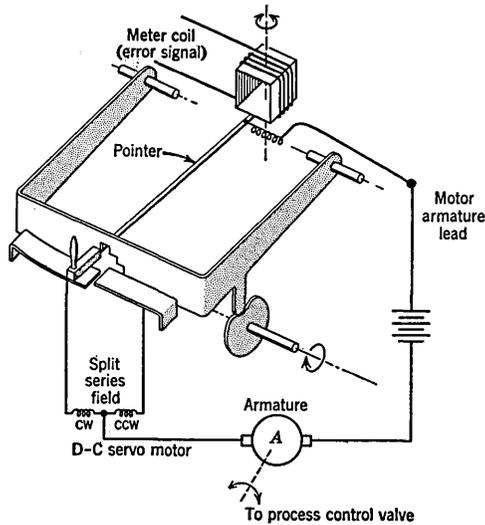


FIG. 6. Schematic form of a periodic controller (chopper bar relay) (Ref. 1).

From a constructional standpoint, the periodic controller operates about as follows. Through a sensing device, such as a meter pointer, the control variable is observed at equal time intervals. Then, by auxiliary power, additional members of the control loop are suitably actuated according to the sensed position of the pointer.

**EXAMPLE.** *The Chopper Bar Controller.* See Fig. 6. As long as the meter pointer stands between the two contact springs the circuit remains broken, even during the sensing instants. It is closed only when the pointer leaves its mid-position. The duration of closure increases with deviation of the pointer. If the contact closure is used to actuate a reversible constant-speed motor, the control action is called *astatic* (never quiet) because, with constant actuating error, the control motor moves *intermittently* across its entire range at an average speed (roughly) proportional to the pointer deviation. Although periodic controllers may have static correspondence between deviation and motor motion, astatic action will be assumed here because of its greater practical significance (Ref. 1).

## 2. METHODS OF TRANSIENT ANALYSIS

### Basic Mathematical Relationships

**Analysis of Sampler.** The output of the sampler (see Fig. 7) is the input modulated by the sampler into a train of pulses:

$$(1) \quad e^*(t) = e(t) \sum_{n=0}^{\infty} u_0(t - nT) = \sum_{n=0}^{\infty} e(nT)u_0(t - nT),$$

where  $u_0(t - nT)$  = the impulse or Dirac delta function occurring at  $t = nT$ , in which

$$u_0(t - nT) \triangleq \lim_{a \rightarrow 0} \frac{u(t - nT) - u(t - nT - a)}{a},$$

$T$  = the sampling period,

$n$  = an integer,

$e(nT)$  = the value of the input at the sampling instant.

The Laplace transform of eq. (1) may be written:

$$(2) \quad E^*(s) = \sum_{n=0}^{\infty} e(nT)e^{-nTs},$$

or eq. (1) may be written in the *frequency response form*:

$$(3) \quad E^*(s) = \frac{1}{T} \sum_{n=-\infty}^{\infty} E(s + jn2\pi/T).$$

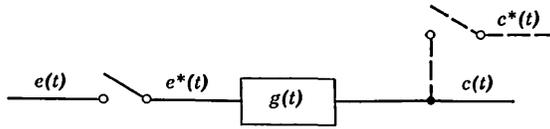


Fig. 7. Showing basic mathematical relationships of a sampler.

NOTE. Equation (3) may be derived by performing the complex convolution of the input  $e(t)$  and the train of unit impulses generated by the sampler, namely

$$(4) \quad E^*(s) = E(s) \otimes \mathcal{L} \left[ \sum_{n=0}^{\infty} u_0(t - nT) \right];$$

where  $\otimes$  is the symbol denoting complex convolution.

Notice that

$$\mathcal{L} \left[ \sum_{n=0}^{\infty} u_0(t - nT) \right] = 1 + e^{-sT} + e^{-2sT} + \dots$$

or in closed form: 
$$= \frac{1}{[1 - \exp(-sT)]}.$$

Because  $\frac{1}{[1 - \exp(-sT)]}$  has only simple poles at  $s = jn2\pi/T$ , the complex convolution reduces to eq. (3).

**Smoothing the Sampled Data.** Normally, the high-frequency components generated by the sampler are removed before the signal reaches the output. Often in sampled-data servo systems, a large portion of the smoothing is accomplished by the components (motors, etc.) between the sampler and the output. Sometimes more smoothing is necessary. One particularly simple low-pass filter is the holding circuit or boxcar generator. In this circuit, the value of a sampling pulse is held until the next pulse arrives, whereupon the circuit assumes the value of the new pulse. The transfer function of such a network is that of a rectangular pulse of unity height and of  $T$  seconds duration, namely

$$(5) \quad G_H(s) = \frac{1}{s} [1 - e^{-sT}].$$

**Response of a Continuous Filter to Sampled Data.** The response of a continuous transfer member  $g(t)$  of Fig. 7 is

$$(6) \quad c(t) = \sum_{n=0}^{\infty} g(t)e(nT)u_0(t - nT).$$

Equation (6) has the Laplace transform

$$(7) \quad C(s) = E^*(s)G(s).$$

Equation (6) is a summation of the filter impulse responses which are excited by each sample and is valid only for the case of *zero initial conditions*. When this condition is not met, a second term must be added to eq. (6) to include the decay of the nonzero initial conditions. Since the system is linear this is not important when considering the stability of the system, but it must be included if a time response is being calculated.

The response of the filter only at the sampling instants is:

$$(8) \quad c^*(t) = \sum_{n=0}^{\infty} e(nT)u_0(t - nT)g(t)u_0[t - (q - n)T],$$

which has the following Laplace transform:

$$(9) \quad C^*(s) = E^*(s)G^*(s),$$

where

$$G^*(s) = \mathcal{L}[g^*(t)] = \mathcal{L}[g(t)u_0(t - nT)].$$

**Sampled-Data System Transfer Function.** From eq. (9), the *sampled-data transfer function* or *pulse transfer* can be defined as

$$(10) \quad G^*(s) = \frac{C^*(s)}{E^*(s)}.$$

An equivalent form in terms of the  $z$ -transform symbolism is indicated in eq. (11). (The  $z$ -transform is defined and illustrated in a later paragraph.)

$$(11) \quad G(z) = \frac{C(z)}{E(z)}.$$

### Laplace Transform Analysis

It is possible to use the equations of the previous section to obtain the complete time response,  $c(t)$ . However, the Laplace transforms are not rational and it requires considerable labor to obtain the complete response. If the response is calculated only at the sampling instants, the transforms can often be written in closed form and the labor of calculation and manipulation is greatly reduced. The  $z$ -transform method is usually used to compute the response at the sensing instants. (See  $z$ -Transform Analysis, Sect. 2.)

**Analysis by Difference Equations**

The analysis of sampled-data control systems leads to characteristic equations, which are difference equations.

**Formulation of the Difference Equations.** Difference equations are discussed in Chap. 4. A simple example will illustrate the analysis of a control process by difference equations.

**EXAMPLE.** (See Fig. 8.) The following simplifying assumptions are made: (a) The displacement of the control means,  $m$ , is linear and unlimited.

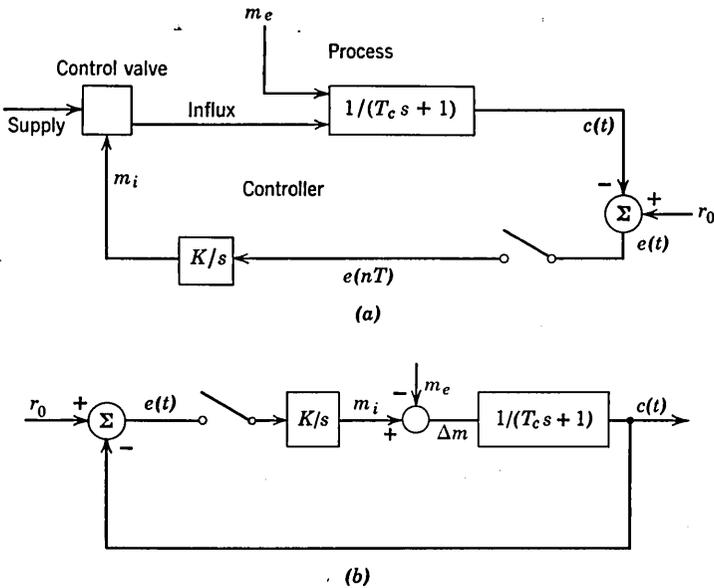


FIG. 8. (a) Simple process control. (b) Elements of process control equivalent to (a).  
 Note.  $e(t) = r_0 - c(t)$ ; this quantity is dimensionless.

(b) The plant has a simple time constant,  $T_c$ . (c) The controller is lag-free, i.e., the sensing and positioning times are negligible.

The disturbance is assumed at the most unfavorable instant (just after sensing). The control means,  $m$ , changes instantly at the sensing instant and remains at its new value throughout the sensing cycle.

The analysis is quite simple. Inside a sensing cycle, the behavior of the plant is continuous and may be described by the linear differential equation

$$\frac{T_c}{T} \frac{de(t)}{d\tau} + e(t) = r_0 - m_i + m_e = r_0 - m$$

where  $T_c$  = the plant time constant, seconds;

$T$  = the sampling cycle, seconds;

$e(t)$  = the controlled variable error, dimensionless:

$e(t) = r_0 - c(t)$ ,  $c(t)$  is the controlled variable,  $r_0$  is the set point, all parameters nondimensional and normalized;

$\tau = t/T$ , dimensionless time;

$m$  = the net value of control means, dimensionless;

$m_i$  = the manipulated control means, dimensionless;

$m_e$  = the disturbance of the control means, dimensionless.

By using the Laplace transformation it can be shown that the solution to eq. (1) at the  $n$ th sensing instant is

$$(12) \quad e_n = De_{n-1} - (1 - D)m_{n-1},$$

where  $e_n$  = value of the controlled variable at the  $n$ th sensing instant,

$e_{n-1}$  = value at the  $(n - 1)$ th sensing instant,

$D = \exp(-T/T_c)$ , the decrement characteristic of the plant and of the sensing cycle,

$m_{n-1}$  = value of the control means at the  $(n-1)$  sensing instant.

Now consider the behavior of the variable  $m$  at the sensing instants. The relationship is assumed to be linear:

$$(13) \quad m_n - m_{n-1} = Ke_n,$$

where  $K$  is the strength of the controller and is called the *specific step*. The minus sign in eq. (13) provides the negative feedback needed for regulation of the variables.

Equations (12) and (13) are the simultaneous difference equations of the control action. They lead to the difference equation of the system, namely,

$$(14) \quad e_{n+2} - [1 + D - K(1 - D)]e_{n+1} + De_n = 0.$$

**Solution of Linear Difference Equations.** The linear homogeneous difference equation may be written:

$$(15) \quad A_0e_{n+q} + A_1e_{n+q-1} + \cdots + A_{q-1}e_{n+1} + A_qe_n = 0.$$

If the roots are not equal, eq. (15) has the solution:

$$(16) \quad e_n = \sum_{i=1}^q a_i z_i^n,$$

where  $z_i$  is a root of the auxiliary equation,

$$(17) \quad A_0z^q + A_1z^{q-1} + \cdots + A_{q-1}z + A_q = 0.$$

The coefficients of eq. (17) are identical with those of eq. (15). Equation (17) is often called the *characteristic equation* of the system.

With *q* distinct roots the general solution is

$$(18) \quad e_n = a_1 z_1^n + a_2 z_2^n + a_3 z_3^n + \dots + a_q z_q^n.$$

If *k* roots are equal the solution is

$$(19) \quad e_n = [a_1 + a_2 n + \dots + a_k n^{k-1}] z_1^n + a_{k+1} z_2^n + \dots + a_q z_{q-k}^n.$$

Thus, the values  $e_n$  at any sensing instant may be computed. The *q* summing constants  $a_i$  are determined by the first *q* of the  $e$ 's at the sensing instants. The characteristic equation with *vanishing roots* has special significance as will be discussed later.

**Use and Limitations of Difference Equation Method.** If the problem is dominated by the sampler, that is, if there is a rather simple control loop whose servomotor is actuated by a periodically applied measurement of the error, this type of analysis is simplest. It is also well to remember that the difference equation method and the *z*-transform method are synonymous. For higher order systems, the difference equation approach becomes laborious, so that the more methodical *z*-transform method becomes advantageous.

### Analysis by *z*-Transform Method

**Usefulness.** The *z*-transform is the shorthand rational way to write the Laplace transform of the linear difference equation. It has the same relationship to linear difference equations as the Laplace transformation bears to linear differential equations. The advantages of the *z*-transform are: (a) it reduces the nonrational Laplace transform of a sampled-data system to a rational transform which facilitates writing transfer functions; (b) it allows definition of the closed loop system response of a sampled-data system (no advantage over difference equations).

**Limitations.** Tables of the more complex *z*-transforms are not readily available and the polynomials must be expanded into partial fractions. A fundamental limitation of *z*-transforms is that the time solutions are defined only at the sensing instants.

**Hidden Oscillations.** Because the time solutions are calculated only at sensing instants, it is possible that the sampling frequency may be lower than the characteristic frequency of the plant being controlled, and oscillations may occur which are not apparent from the *z*-transforms. If such a condition is suspected the *z*-transform can be modified to give the output between the sampling instances and the existences of such oscillations can be checked. (See Refs. 4 and 5.)

**Basic Relationship.** The  $z$ -transform is based on the transformation

$$(20) \quad z = e^{sT},$$

where  $s$  is the Laplace operator and  $T$  is the sampling period. The Laplace transform of the sampled signal will contain  $s$  in the irrational form  $e^{-nsT}$ . Substitution of  $z$  will produce a rational transform in  $z$ . The  $z$ -transform is defined as

$$(21) \quad C(z) = \sum_{n=0}^{\infty} c(nT)z^{-n}.$$

**Table of Useful  $z$ -Transforms.** See Table 1.

TABLE 1. LAPLACE AND  $z$ -TRANSFORMS (Refs. 2, 5)

Row	Column 1 Laplace Transform	Column 2 Time Function	Column 3 $z$ -Transform	Column 4 Description of Time Function
<i>a</i>	1	$u_0(t)$	1	Impulse function at $t = 0$
<i>b</i>	$e^{-nTs}$	$u_0(t - nT)$	$\frac{1}{z^n}$	Impulse function at $t = nT$
<i>c</i>	$\frac{1}{1 - e^{-Ts}}$	$i(t)$	$\frac{z}{z - 1}$	Train of impulses at sampling instants
<i>d</i>	$\frac{1}{s}$	$u(t)$	$\frac{z}{z - 1}$	Step function
<i>e</i>	$\frac{1}{s^2}$	$t$	$\frac{Tz}{(z - 1)^2}$	Ramp function
<i>f</i>	$\frac{1}{s^3}$	$\frac{1}{2}t^2$	$\frac{1}{2}T^2 \frac{z(z + 1)}{(z - 1)^3}$	Quadratic or ac- celeration func- tion
<i>g</i>	$\frac{1}{s + a}$	$e^{-at}$	$\frac{z}{z - e^{-aT}}$	Exponential func- tion
<i>h</i>	$\frac{a}{s^2 + a^2}$	$\sin at$	$\frac{z \sin aT}{z^2 - 2z \cos aT + 1}$	Sinusoidal func- tion
<i>i</i>	$\frac{1}{s - (1/T) \ln a}$	$a^{t/T}$	$\frac{z}{z - a}$	Constant raised to power $t$
<i>j</i>	$\frac{b}{[s - (1/T) \ln a]^2 + b^2}$	$a^{t/T} \sin bt$	$\frac{za \sin bT}{z^2 - 2az \cos bT + a^2}$	Sine wave multi- plied by $a^{t/T}$
<i>k</i>	$\frac{s - (1/T) \ln a}{[s - (1/T) \ln a]^2 + b^2}$	$a^{t/T} \cos bt$	$\frac{z(z - a \cos bT)}{z^2 - 2az \cos bT + a^2}$	Cosine wave mul- tplied by $a^{t/T}$
	$F(s + a)$	$e^{-at}f(t)$	$F(e^{-aT}z)$	Effect of multipli- cation by $e^{-at}$

**Methods of Inverting  $z$ -Transformation.**

*Real Inversion Integral.* The real inversion integral for the  $z$ -transform is

$$(22) \quad c(nT) = \frac{1}{2\pi j} \oint C(z)z^{n-1} dz,$$

where the line integration is made of a sufficiently large radius to enclose all roots of the integrand.

*Partial Fraction Expansion.* The  $z$ -transform is factored into components so that each term in the expansion can be obtained from Table 1. The usual methods for partial fraction expansion are applicable. (See Chap. 20.)

*Power Series Expansion.* From eq. (21)

$$(23) \quad C(z) = \sum_{n=0}^{\infty} c(nT)z^{-n} = \frac{c(0)}{z^0} + \frac{c(T)}{z^1} + \frac{c(2T)}{z^2} + \dots,$$

which thus expands the  $z$ -transform of a variable in an inverse power series in  $z$ . The coefficient  $c(nT)$  of  $z^{-n}$  is the value of the variable at the  $n$ th sensing instant, and the coefficients can be used directly to plot the time function *at the sampling instants*.

**$z$ -Transform Block Diagram Algebra.** The  $z$ -transform describes the transfer function of two variables *at the sensing instants only*. Figure 9a illustrates the transform  $R(z)$ . Figure 9b shows the transform

$$(24) \quad C_b(z) = R(z)G_1(z),$$

and Fig. 9c

$$(25) \quad C_c(z) = R(z)G_1(z)G_2(z).$$

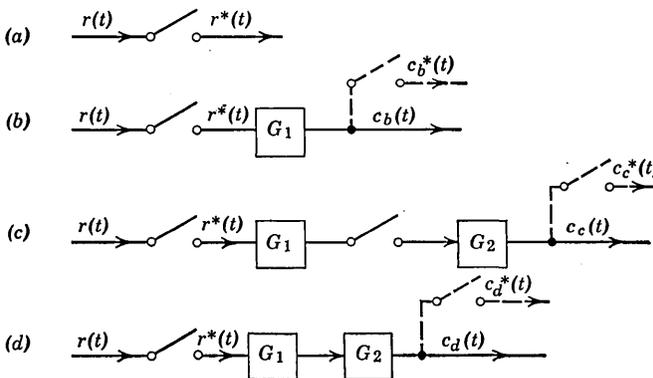


FIG. 9. Basic  $z$ -transform relationships.

In words, if each transfer member is separated from others by *synchronous* samplers, the  $z$ -transforms cascade, i.e., they can be multiplied. But notice that if the transfer members *are not separated by a chopper*, the  $z$ -transform cannot be obtained by multiplying together the  $z$ -transforms of the component members. In continuous systems where coupling exists between transfer members a similar difficulty is encountered. For *example* Fig. 9d has the transform:

$$C_d(z) = R(z)G_1G_2(z).$$

Consider

$$G_1(s) = \frac{1}{s+1}; \quad G_1(z) = \frac{z}{z - \exp(-T)},$$

and

$$G_2(s) = \frac{1}{s+2}; \quad G_2(z) = \frac{z}{z - \exp(-2T)}.$$

Then

$$\frac{C_c(z)}{R(z)} = \frac{z^2}{[z - \exp(-T)][z - \exp(-2T)]},$$

whereas

$$\begin{aligned} \frac{C_d(z)}{R(z)} &= \frac{z}{[z - \exp(-T)]} - \frac{z}{[z - \exp(-2T)]} \\ &= \frac{z[\exp(-T) - \exp(-2T)]}{[z - \exp(-T)][z - \exp(-2T)]}. \end{aligned}$$

NOTE.

$$G_1(z)G_2(z) \neq G_1G_2(z).$$

The difference is that in the first case,  $G_2$  is driven by a train of impulses, whereas in the second, it is driven by the linear response of  $G_1$  to its own input pulses. A helpful concept is that the  $z$ -transform of a chain of transfer members must be derived from chopper to chopper in the circuit. Table 2 shows some control loops, their Laplace transforms and their  $z$ -transforms. The output  $c$  may be assumed to be sampled by an imaginary chopper (synchronized with the real one), resulting in  $c(nT)$ , although this imaginary chopper must be disregarded in traversing the complete control loops.

TABLE 2. OUTPUT TRANSFORMS FOR BASIC SAMPLED-DATA SYSTEMS<sup>a</sup>

	System	Laplace Transform of Output $C(s)$	$z$ -Transform of Output $C(z)$
1		$R^*(s)$	$R(z)$
2		$GR^*(s)$	$GR(z)$
3		$G(s)R^*(s)$	$G(z)R(z)$
4		$\frac{G(s)R^*(s)}{1 + HG^*(s)}$	$\frac{G(z)R(z)}{1 + HG(z)}$
5		$\frac{G^*(s)R^*(s)}{1 + H^*(s)G^*(s)}$	$\frac{G(z)R(z)}{1 + H(z)G(z)}$
6		$G(s) \left[ R(s) - \frac{H(s)RG^*(s)}{1 + HG^*(s)} \right]$	$\frac{RG(z)}{1 + HG(z)}$
7		$\frac{G_2(s)RG_1^*(s)}{1 + HG_1G_2^*(s)}$	$\frac{G_2(z)RG_1(z)}{1 + HG_1G_2(z)}$

<sup>a</sup> This table is reprinted from an article by Ragazzini and Zadeh (Ref. 2) with the permission of the authors.

### 3. SAMPLED-DATA SYSTEM STABILITY

#### Stability Criteria of Difference Equations and $z$ -Transforms

The solution of the characteristic difference equation with nonequal roots is

$$(26) \quad c_n = c(nT) = \sum_{i=1}^m a_i z_i^n.$$

Now if  $c(nT)$  is to remain finite even for large  $n$ , then

$$(27) \quad z_i^n < 1 \quad \text{or} \quad |z_i| < 1.$$

In words, the inequality (27) states that, for stability of the sampled-data

system, the roots (or zeros) of its characteristic difference equation must lie inside the unit circle with center at the origin. *The unit circle in the z-plane is the periodic limit of stability corresponding to the Routh-Hurwitz stability criteria.*

**The Routh-Hurwitz Stability Criteria.** This is used in linear control theory and may be applied to sampled-data systems by using the conformal transformation

$$(28) \quad s = \frac{z + 1}{z - 1}.$$

This transformation changes the unit circle in the z-plane into the left half of the s-plane as shown in Fig. 10. If the Hurwitz conditions are applied

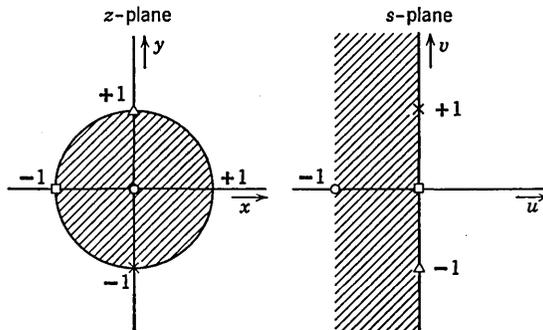


FIG. 10. The linear transformation  $s_w = (z + 1)/(z - 1)$ , used for deriving stability conditions from the difference equation of control.

to the characteristic equation (subjected to the transformation of eq. 28), the conditions can be found which cause the roots of the transformed equation to lie in the left half of the w-plane. Hence, the roots of the characteristic equation must lie within the unit circle in the z-plane. As an *example* consider the second order characteristic equation

$$(29) \quad A_0 z^2 + A_1 z + A_2 = 0.$$

If the transformation (eq. 28) is used, the transformed equation is

$$(30) \quad B_0 s^2 + B_1 s + B_2 = 0,$$

where  $B_0 = A_0 + A_1 + A_2$ ,

$$B_1 = (A_0 - A_2),$$

$$B_2 = A_0 - A_1 + A_2.$$

In the simple quadratic case, the Hurwitz criterion requires for stability only that all these coefficients have the same sign. Therefore, the periodic

stability limit of the second order eq. (29) is defined by the *dual* condition (if  $B_0 > 0$ ):

$$(31) \quad \begin{aligned} B_1 > 0 & \text{ or } A_0 - A_2 > 0, \\ B_2 > 0 & \text{ or } A_0 - A_1 + A_2 > 0. \end{aligned}$$

The above procedure can be extended to higher order systems.

**Use of Frequency Response Methods to Determine Stability**

**Nyquist Diagram.** *Exact Graphical Procedure.* The Nyquist diagram can be drawn by considering  $G(z)$  rather than  $G^*(s)$ . The complex plane plot is made by allowing  $z$  to vary along a unit circle in the  $z$ -domain. The gains and phase at any frequency are found by locating the point on the unit circle ( $z$ -domain) corresponding to this angular frequency. The interpretation of the Nyquist diagram follows conventional lines.

**EXAMPLE.** Simple sampled-data system ( $T =$  sampling period) (this example is after a similar example by Truxal, Ref. 5):

$$\begin{aligned} G(s) &= \frac{K}{s(s+1)} = K \left[ \frac{1}{s} - \frac{1}{s+1} \right], \\ G^*(s) &= \left[ \frac{z}{z-1} - \frac{z}{z-e^{-T}} \right] K = \frac{Kz(1-e^{-T})}{(z-1)(z-e^{-T})}, \end{aligned}$$

for  $\omega_s = 4$  rad/sec,

$$\begin{aligned} \frac{2\pi}{\omega_s} = T &= \frac{6.28}{4} = 1.57; \quad e^{-1.57} = 0.208, \\ G^*(s) &= \frac{0.792Kz}{(z-1)(z-0.208)}. \end{aligned}$$

The unit circle in the  $z$ -plane would appear as in Fig. 11. The vectors are shown for 1 rad/sec. At  $\omega = 1$  rad/sec,

$$\begin{aligned} z &= 1/90^\circ, \\ G^*(j1) &= \frac{0.792K1/90^\circ}{1.414/135^\circ \cdot 1.0216/101.8^\circ} \\ &= 0.86 \frac{K}{T} / -146.8^\circ; \end{aligned}$$

likewise

$$G^*(j2) = 0.52 \frac{K}{T} / 180^\circ.$$

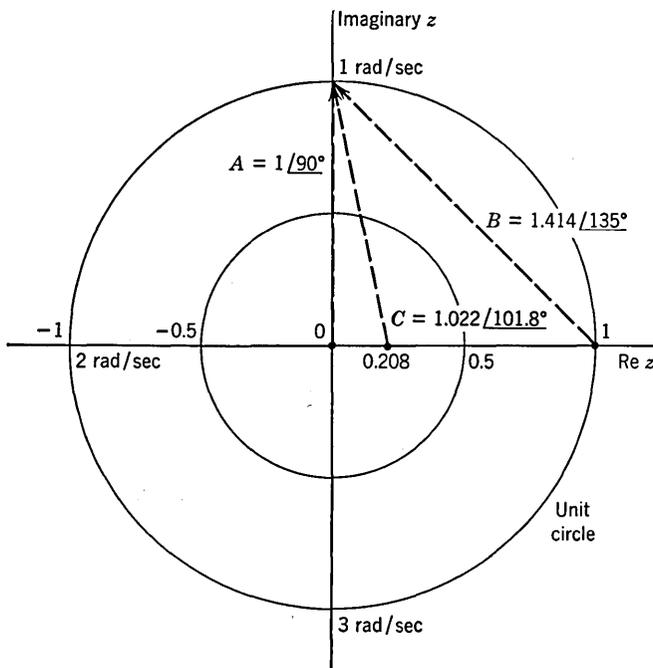


FIG. 11. Pole-zero configuration for  $G(z)$ , with vectors shown for calculation of Nyquist diagram at  $\omega = 1$  rad/sec

$$G(s) = \frac{K}{s(s + 1)},$$

and  $\omega_s = 4$  rad/sec.

Continuing the above for other frequencies, a Nyquist plot,  $G^*(j\omega)$ , similar to that shown in Fig. 12, could be produced.

**Comparison with Amplitude Modulation. Graphical Approximate Nyquist Plot.** Linvill applied the Nyquist diagram to sampled systems based on an approximation for the starred open loop transfer function:

$$G^*(s) = \frac{1}{T} \sum_{n=-\infty}^{\infty} G(s + jn\omega_s).$$

If  $G(s)$  is a good low-pass filter,  $G^*(s)$  will contain only two or three significant terms. For example  $G^*(j1)$  is the vector addition

$$G^*(j1) = (1/T)[G(j1) + G(j1 - j\omega_s) + G(j1 + j\omega_s) + \dots].$$

If the sampling frequency is 4 rad/sec,

$$G^*(j1) = [G(j1) + G(-j3) + G(j5) + G(-j7) + \dots](1/T).$$

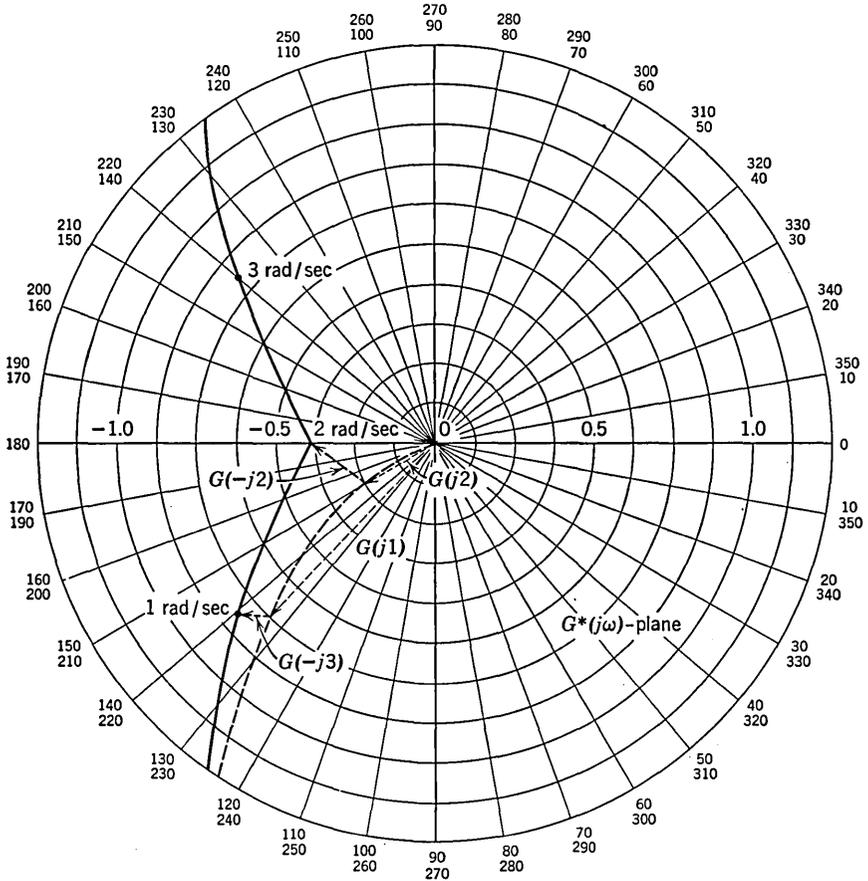


FIG. 12. Nyquist diagram for  $G^*(s)$  with  $G(s) = K/s(s + 1)$ ,  $K/T = 1$ , and  $\omega_s = 4$  rad/sec (constructed by using two-term approximation) (Ref. 5). Radial scale numbers are in terms of  $T/K$ ; circle spacing is  $0.125T/K$ .

All terms except  $G(j1)$  and  $G(-j3)$  would be small if  $G(s)$  is an effective low-pass filter.

The graphical construction on the  $G^*(j\omega)$  plane is shown in Fig. 12.

NOTE. The value of  $G^*(j\omega_s/2)$  is purely real. At frequencies above  $\omega_s/2$  the Nyquist diagram continues into the upper half-plane until it reaches infinity at the sampling frequency. The only part of the diagram of interest in stability considerations is the section corresponding to frequencies lying between zero and  $\omega_s/2$ .

The *example* illustrates that sampling, by itself, increases the phase lag for a given gain. From the Nyquist diagram the maximum gain for a

stable system is read directly. If only the two terms are used in the series expansion of  $G^*(j\omega)$  the allowable  $K/T$  is 2.5, if all are used, the  $K/T$  is 1.94. The gain in the first case is 3.93; consideration of the rest of the terms reduces the allowable gain to 3.05, since in this example,  $T = 2\pi/4 = 1.57$ .

#### 4. SAMPLED-DATA SYSTEM SYNTHESIS

##### Design Procedure Using $z$ -Transforms

This section is based upon material from Ref. 4.

The typical sampled-data system of Fig. 13 will be used for *illustration*. The error unit embodies both the analog-digital transducer, which peri-

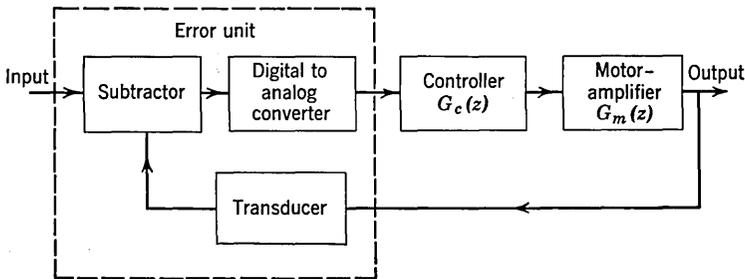


FIG. 13. The basic digital servo system.

odically expresses the angular position of the output shaft as a number in binary code form and the digital subtractor which takes the difference of this number and the incoming one. The characteristics of the servo motor-amplifier combination differ for different applications. They are assumed to be known and invariant, so that the problem is to synthesize a suitable controller. The following  $z$ -transforms will be used:

$G_c(z)$  =  $z$ -transform of the controller,

$G_m(z)$  =  $z$ -transform of the motor-amplifier combination,

$G_0(z)$  =  $z$ -transform of the open control loop,

$G(z)$  = the closed loop  $z$ -transform.

$$(32) \quad G(z) = \frac{z\text{-transform of the forward path}}{1 + G_0(z)}$$

**Performance Criteria.** It is useful to assess the performance in terms of the responses to specific driving functions such as a step function, a steady velocity or acceleration, a sinusoidal input of various frequencies or

a random noise. Any or all such tests may be applied and, since improvement in one respect is often accompanied by deterioration in another, it will be necessary to compromise. Such overriding factors as the demand for zero velocity lag must take precedence. The servo amplifier may overload if it is fed a series of discontinuous pulses representing samples. Its input must be reasonably smooth, and the correction due to one error number will not be complete before the next is begun.

An equivalent system is shown in Fig. 14. The system delay,  $\lambda T$ , is now shown with the motor amplifier. The controller has two parts, the first of which is characterized by its  $z$ -transform,  $G_1(z)$ , and it modifies the sequence of correction samples supplied to it. The modified sequence is

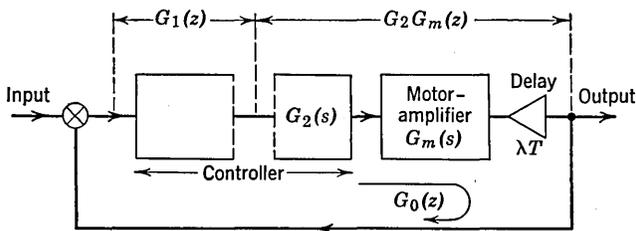


FIG. 14. A system equivalent to that of Fig. 13.

smoothed by the second part, characterized by its transfer function  $G_2(s)$ , which provides a continuous signal for driving the servo amplifier. This subdivision is unlikely to correspond to any physical separation of the components. The composite expression  $G_1(z) * G_2(s)$  may be called the "operational instruction" of the controller. The  $*$  symbol is used in this case to separate the sampled and continuous portions of the operational instruction and indicates that the information input to the continuous elements is in sampled form.

Knowing  $G_m(s)$  and the performance requirements, it should be possible to specify  $G_2(s)$ . For *example*, the motor amplifier may have the simple transfer function  $1/[s(1 + T_m s)]$ , where the time constant,  $T_m$ , is probably smaller than the sampling interval. To avoid sudden changes in velocity,  $G_2(s)$  need only be  $1/s$ .

The next step is the determination of a suitable  $G_0(z)$ , taking into account all the overriding factors. The fact that  $G_0(z)$  may be expressed as the ratio of two polynomials  $N(z)/D(z)$  is also used.

**Physical Realizability.** The order of  $N$  must be at least one less than that of  $D$ .

**Poles at  $z = 1$ .** To have zero static error, the function  $G_0(z)$  must possess at least one simple pole at  $z = 1$ . A second order pole at  $z = 1$

would provide zero velocity lag and a third order pole, a zero acceleration lag characteristic.

**Cancellation of Poles and Zeros.** The characteristic equation of the system is  $D(z) + N(z, \Delta) = 0$ . The parameter  $\Delta$  indicates the need for checking the values of the system variables between sensing instants. The system will be unstable if any root lies on or outside the unit circle ( $|z| = 1$ ). One may be tempted to arrange by adjustment of parameters for the cancellation of a zero by a pole so as to eliminate the root which would otherwise lie outside the unit circle. It is better to increase the sampling frequency. This point cannot be emphasized too strongly, particularly because it is tempting to deal with the special case of no system delay, but this can result in instability that would then be revealed only when the behavior between sampling instants is investigated. See Ref. 5.

**Design Constants.** The suggested method of synthesizing a system is to match the characteristic equation with one known to give satisfactory performance. Lawden *et al.* (Ref. 6) has used equations of the form  $(z - a)^n = 0$ , although when  $n$  is a small number, it may be desirable to depart from this form. Oldenbourg and Sartorius (Ref. 1) show that minimum control area (see Condition for Minimal Control Area, later in this section) results from the case of vanishing roots, namely  $z^n = 0$ . Examples relevant to continuous systems may well be suitable for sampling systems. The procedure is to arrange for  $N(z)$  and  $D(z)$  to include between them a number of constants which are adjustable in the design stage. This number should be equal to the order of the characteristic equation. It is always possible to do this, because two additional constants are picked up each time the order of the characteristic equation is increased by one. The characteristic equation  $z^n = 0$  leads to minimum control area, but if there is noise present, very little smoothing is provided; as a result, the servo amplifier may be transiently overloaded or driven into saturation. The characteristic equation  $(z - 0.4)^n$  has been used by some authors to provide smooth and satisfactory performance in the presence of noise. Analysis of a representative second order sampled-data system by Jury (Ref. 7) leads to the results of Fig. 15, which shows the constant overshoot loci in the  $z$ -plane. It can be shown that these loci can be used for higher order systems. Note that a system which has no overshoot has its characteristic roots on the positive real axis. The values of the roots must be less than unity.

The simplest expression which meets all the above requirements is the expression  $G_0(z)$ . Dividing it by  $G_2G_m(z)$  gives  $G_1(z)$ , the first part of the operational instruction.

**The Operational Instruction.** It remains to decide how standard components may be assembled into a system having the required operational instruction,  $G_1(z) * G_2(s)$ . The  $s$ -part must describe the properties

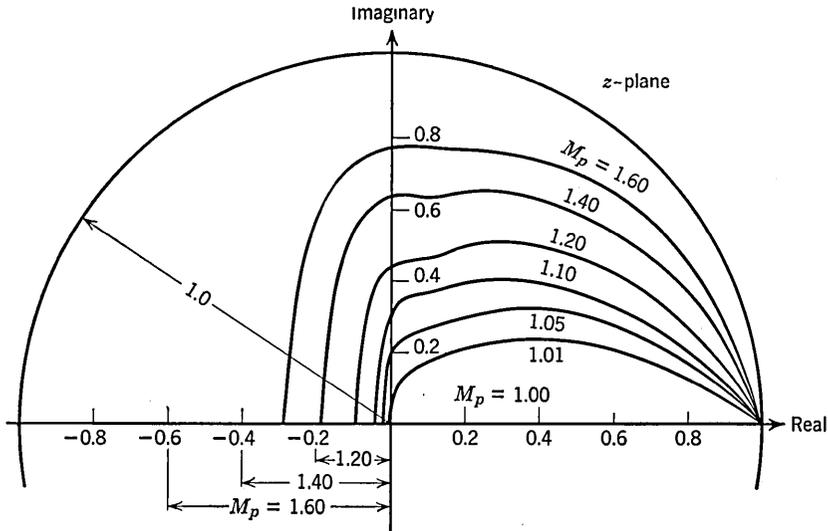


FIG. 15. Constant overshoot loci in the  $z$ -plane for a unit step input,  $M_p$  = ratio of values transient peak to steady state (Ref. 7).

of the digital-analog converter included in the error unit. This converter may perform the function of a clamp, which has the operational instruction  $(1 - e^{-sT})s^{-1}$ . Other functions of  $s$  may be obtained by the usual synthesis procedures and may lead to further terms in the  $z$ -part. This usually leaves an expression in  $z$  which is required for the rest of the operational instruction. Generally, such an expression is of the form

$$(33) \quad \frac{A_0 + A_1z^{-1} + A_2z^{-2} + \dots + A_rz^{-r}}{1 + B_1z^{-1} + B_2z^{-2} + \dots + B_rz^{-r}}$$

This function may be synthesized in many ways; one way of constructing its physical counterpart is with the aid of  $r$  delay elements (each equal to  $T$ ). The output is obtained by the summation of delayed components proportional to the coefficients in the numerator. The correct denominator is obtained by negative feedback of the delayed components proportional to the coefficients in the denominator.

**EXAMPLE. Synthesis of a Simple Analog System.** Figure 14 shows the simple system to be synthesized, and the following assumptions are made regarding it:

(a) The servo motor and amplifier are constructed so that the rate of rotation of the motor is proportional to the voltage applied to the amplifier. The transfer function is  $G_m(s) = K/s$ .

(b) The transducer samples and introduces a delay,  $T$ , the effect of which is to multiply the  $z$ -transform by  $z^{-1}$ .

(c) There must be zero static error; hence  $G_0(z)$  must include the factor  $(z - 1)$  in its denominator.

(d) There shall be no sudden changes in output velocity. The law of motion shall be quadratic between sampling instants. Hence  $G_2(s) = s^{-2}$ . Therefore,  $G_2(s)G_m(s) = Ks^{-3}$  and  $G_2G_m(z) = KT^2z(z + 1)/2(z - 1)^3$ , the  $z$ -transform being found directly from Table 1.

**The  $z$ -Transform of the Loop.** This is

$$(34) \quad G_0(z) = \frac{KT^2G_1(z)(z + 1)}{2(z - 1)^3}$$

It contains the required factor  $(z - 1)$  in the denominator. It must also contain adjustable design constants such that the characteristic equation can be forced into one known to be suitable such as  $(z - a)^n = 0$ . The simplest expression for  $G_1(z)$  which adds two further constants without increasing the order is

$$(35) \quad G_1(z) = (z - 1)^2/(z^2 + B_1z + B_2).$$

The operational instruction for the controller is therefore

$$(36) \quad \frac{(z - 1)^2}{(z^2 + B_1z + B_2) * s^{-2}}.$$

The  $z$ -transform of the operational instruction is obtained by replacing  $s^{-2}$  with its  $z$ -transform,  $Tz(z - 1)^{-2}$ , which reduces eq. (36) to  $Tz(z^2 + B_1z + B_2)^{-1}$ . The characteristic equation is

$$(37) \quad z^3 + (B_1 - 1)z^2 + (B_2 - B_1 + 1/2KT^2)z + (1/2KT^2 - B_2) = 0.$$

The simplest third order equation this can be identified with is  $z^3 = 0$ . Choice of the characteristic equation  $z^3 = 0$  is known to produce the most rapid recovery from a transient disturbance. (See Condition for Minimal Control Area, later in this section.) If  $KT^2 = B_1 = 1$  and  $B_2 = 1/2$ ,  $G_0(z)$  can be written

$$(38) \quad G_0(z) = (z + 1)(2z^3 - z - 1)^{-1}.$$

The closed loop  $z$ -transform in response to a pulse is

$$(39) \quad G(z) = G_0(z)[G_0(z) + 1]^{-1},$$

or

$$G(z) = \frac{1}{2}z^{-2} + \frac{1}{2}z^{-3}.$$

**Design Procedure Using Frequency Response.** The design procedures of linear techniques discussed in previous chapters are applicable. Nyquist and Bode diagrams may be used after the transfer function has been obtained. In working with the Nyquist diagram, it can be seen that lead compensation will increase the bandwidth of the system. If the sampling frequency is not increased, no additional high-frequency information will be passed. This illustrates a difficulty which one may encounter in the synthesis of sampled data systems.

**Design Procedure Using Root Locus Techniques.** As in frequency response techniques, the root locus could be used as an aid for synthesizing sampled-data systems. However, it can be shown that the desired root locus is the positive real axis in the  $z$ -plane. As previously mentioned, the characteristic equation  $z^n = 0$  is known to lead to the fastest recovery of the system from a disturbance. If noise is present  $(z - a)^n = 0$  is the desired characteristic, where  $a$  is a number between zero and 1. As Jury has shown (see Fig. 15), the loci of constant overshoot also illustrate that the positive  $z$ -axis is the desired place to locate the roots of the characteristic equation. The circumference of the circle in the  $z$ -plane having unity radius is the periodic limit of stability.

In summary, the sampling frequency controls the bandpass and thus the speed with which the system can transmit information. The roots of the characteristic equation should be placed as near the origin as permissible. Placement at the origin is known to produce the liveliest system; if noise is present, the roots must be moved along the positive real axis in the  $z$ -plane toward  $z = 1$ . It should be noted that in many practical cases the above simple criterion for performance will have to be modified for one or more practical reasons. In such cases the approach suggested is to use the above rules for the first approximation and then to introduce the other considerations.

### Performance Charts for Typical Sampled-Data Systems

**Performance Index: Control Area.** To evaluate the results of computations and to choose the most favorable conditions of operation, the concept of control effectiveness, measured by the smallness of the control area, is very useful. For continuous controllers, the control area is defined as

$$(40) \quad F = \int_0^{\infty} e(t) dt.$$

For sampled-data controllers the calculation is not so simple, except in one case, when the control process has the initial value of zero at the first

sensing instant. In this case (which leads to the largest control area) the control area is

$$(41) \quad \frac{F}{T} = \sum_{n=0}^{\infty} e_n, \quad T = \text{sampling period.}$$

It can be seen the control area is the error-time integral.

**Condition for Minimal Control Area.** The characteristic equation with vanishing roots, namely  $z^n = 0$ , has the least control area. Such a system can be shown to recover most quickly from a disturbance. However, if there is noise present, or if such a characteristic equation is physically unrealizable, the characteristic equation  $(z - a)^n = 0$  is used. It is used in the presence of noise to provide smoothing of the impulses, and it is used in the second case so that normal system components may be employed.

**Second Order System with Dead Time and with No Compensation.** (See Fig. 16.) The characteristic equation for this system is

$$(42) \quad z^2 + [K(1 - D/L) - (1 + D)]z + D - KD(1 - 1/L) = 0,$$

where  $L = \exp(-T_L/T_c)$  and  $D = \exp(-T/T_c)$ .

If  $T < T_L < 2T$ , the equation becomes

$$(43) \quad z^3 - (1 + D)z^2 + [D + K(1 - D^2/L)]z + KD(D/L - 1) = 0.$$

Dead time does not increase the order of the characteristic equation as long as  $T_L < T$ . When  $T < T_L < 2T$ , the order of the equation is increased from 2 to 3. Further increase of the dead time (or shortening of the sensing time) leads to successively higher order characteristic equations, and it can be shown that the equation becomes transcendental for the case of the continuous controller. It can be shown that the controlled variable never oscillates if all the roots lie on the positive axis of the  $z$ -plane between 0 and +1. This fact is used to force eq. (42) to have a double positive root less than unity. The control factors leading to this aperiodic limit are shown in Fig. 16. It is not possible to cause the roots to vanish (i.e.,  $z^n = 0$ ) unless compensation is added. Adding compensation introduces two arbitrary constants into the characteristic equation. The two extra constants can be used to design the system characteristic for vanishing roots.

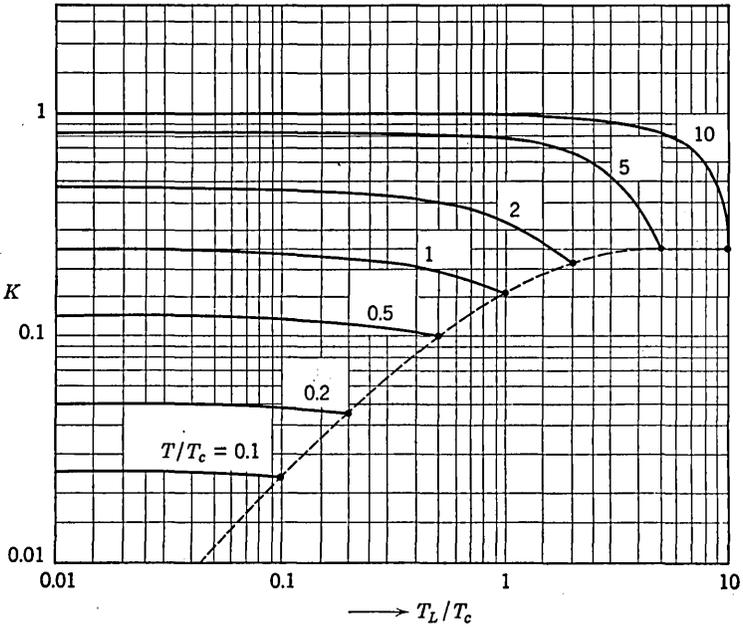
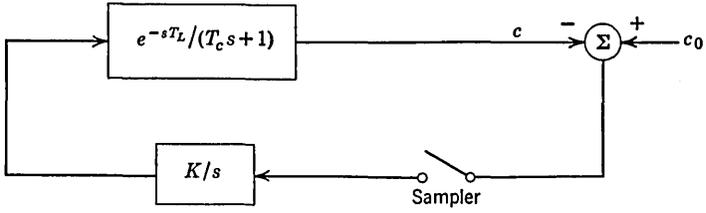


FIG. 16. Excursion-dependent periodic control on a plant with first order time constant and dead time; aperiodic limit (Ref. 1).



This limit arises when  $A_1/A_0 = 0$  and  $A_2/A_0 = 0$ . These conditions are called *optimal* because they lead to least control area. They are valid only in the range  $0 < T_L/T_c < T/T_c$ . The control area increases steadily with  $T/T_c$  so that if one is free to choose  $T/T_c$ , the most favorable operating conditions are obtained when  $T = T_L$ . The action following a step disturbance inside the control loop is shown in Fig. 18. The parameters for optimal response are shown in Fig. 17.

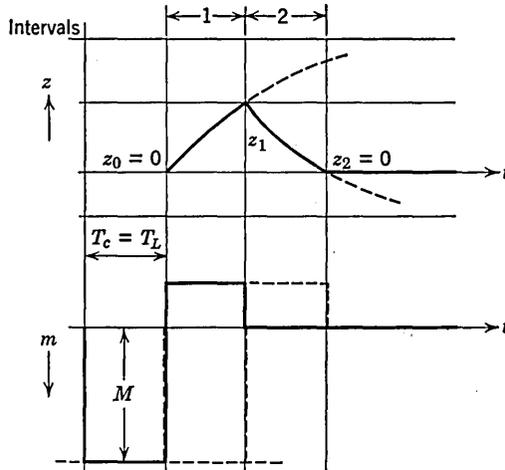


FIG. 18. Example of a difference equation of second order with vanishing characteristic values (Ref. 1).

**Third Order System with Dead Time and with Delayed Rate Compensation.** The system of Fig. 19 leads to the following characteristic equation:

$$(45) \quad A_0 z^3 + A_1 z^2 + A_2 z + A_3 = 0,$$

where  $A_0 = 1$ ,

$$A_1 = K(1 - D/L + \rho Q) - 1 - D - Q,$$

$$A_2 = K[D/L(1 + Q) - \rho Q(1 + D) - D - Q] + D + DQ + Q,$$

$$A_3 = DQ[K(1 + \rho - 1/L) - 1],$$

$$Q = \exp(-T/T_R), \text{ and } L, D, M, F \text{ are defined above (see eq. 44).}$$

Here again optimal response is possible with its finite control process similar to that shown in Fig. 18. The parameters for optimal response are shown in Fig. 19.

**Comparison of Continuous and Sampled-Data Controllers.** If the process has no dead time, sampled-data control is decidedly less favor-

FEEDBACK CONTROL

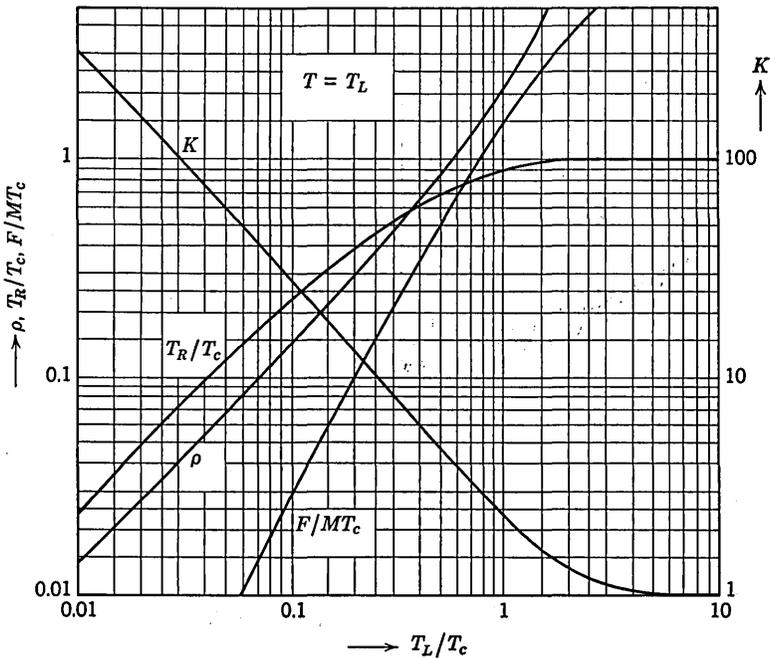
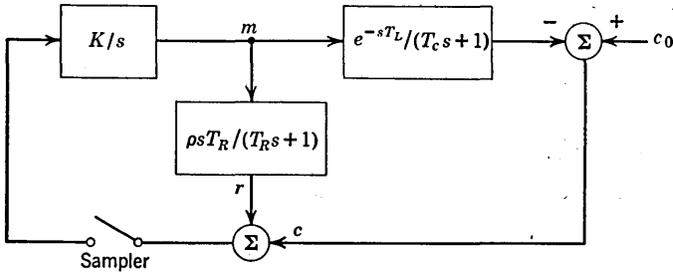


FIG. 19. Excursion-dependent periodic control with retractile followup, on a plant with first order time constant and dead time; parameters for optimal response (Ref. 1).

able than the corresponding continuous control. Figure 20 shows the relationships which are present when there is dead time in the plant. If the control areas of sampled-data and continuous controllers *without stabilization* are compared at the aperiodic limit, the two upper solid lines of Fig. 20 are obtained. For small values of the dead time, these two curves can hardly be distinguished from one another. Decidedly different relations are present, however, if the controller includes a stabilizing device, since

then a control response which terminates in a finite time can be had with a sampled-data controller. A comparison of control areas, Fig. 20, shows that sampled-data control gives appreciably better results. To be sure, the combination of parameters which causes vanishing roots of the characteristic equation is not possible for arbitrarily small dead times

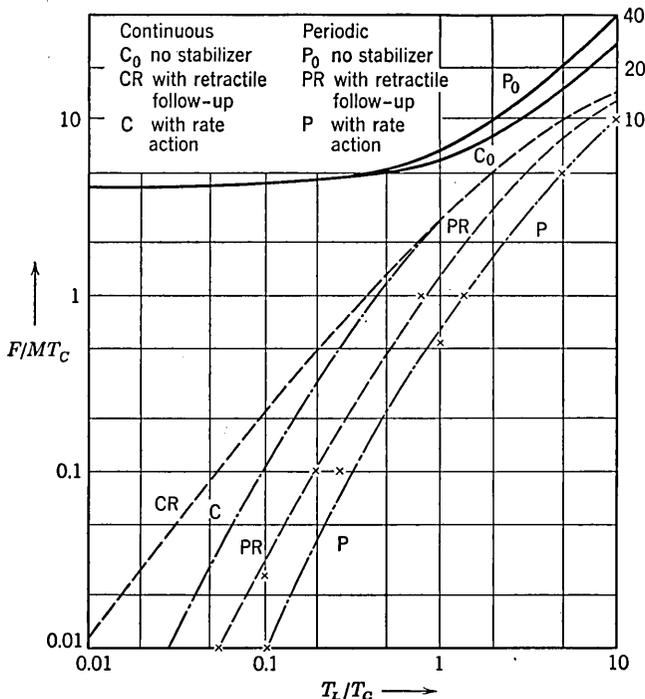


FIG. 20. Comparison of optimal control areas, continuous versus periodic controllers, on a plant with first order time constant and dead time (Ref. 1). Basic plants similar to those of Figs. 16, 17, and 19.

because, among other things, the gain required becomes smaller than is practically possible. It should be entirely acceptable to have somewhat overdamped response when the dead time is small.

**Summary.** Aside from constructional considerations, the sampled-data controller is most appropriately used when the control loop has considerable dead time and the controller is provided with a stabilizing device. Without the latter the behavior of a continuous controller is basically more favorable. However, with short dead times or a short sensing cycle, the differences between the two forms are so slight that constructional considerations alone can dictate the choice. It need hardly be explained

that the frequency and the form of the disturbance can be decisive in the choice of one or the other type of controller. The length of the sensing interval,  $T$ , is one of the most important considerations in the design of sampled-data systems. If possible it should be chosen equal to the dead time,  $T_L$ . If the specifications on the controlled response do not permit such a long sensing cycle, the sensing cycle must be shortened at the expense of less damping or the dead time must somehow be reduced. If the dead time is very short, continuous controllers may be indicated, but the performance of sampled-data controllers should also be entirely acceptable.

### REFERENCES

1. R. C. Oldenbourg and H. Sartorius, *Dynamik Selbsttätiger Regelungen*, R. Oldenbourg Verlag München, 1. Auflage 1944. 2. Auflage 1951. The American translation is *The Dynamics of Automatic Controls*, Am. Soc. Mech. Engrs., New York, 1948.
2. J. R. Ragazzini and L. A. Zadeh, The analyses of sampled data systems, *Trans. Am. Inst. Elec. Engrs.*, **71**, Pt. 2, 225-232 (1952).
3. W. K. Linvill, Sampled-data control systems studied through comparison with amplitude modulation, *Trans. Am. Inst. Elec. Engrs.*, **70**, Pt. 2, 1779-88 (1951).
4. R. H. Parker, A servo system for digital data transmission, *Proc. Inst. Elec. Engrs.*, **75**, Pt. B, 52-64 (1956).
5. John G. Truxal, *Automatic Feedback Control Systems Synthesis*, McGraw-Hill, New York, 1955.
6. C. H. Smith, D. Lawden, A. Bailey, Characteristics of sampling servo-systems, in *Automatic and Manual Control*, A. Tustin, Editor, pp. 377-408, Butterworths, London, 1952.
7. E. I. Jury, Synthesis and critical study of sampled data control systems, *Trans. Am. Inst. Elec. Engrs.*, **75**, Pt. 2, 141-149 (1956).

# INDEX

- A-c control systems, 20-79  
  components treated as d-c components, 20-83  
  design, 20-83
- A-c servomechanisms, see *A-c control systems*
- A-c systems, carrier frequency shift, 23-52  
  compensating networks, 23-48
- Accuracy, of control, 19-06  
  dynamic, 20-70  
  error coefficients, 20-74  
  static, 20-70
- Adams and Bashforth method, 14-59
- Adjoint, of control system, 24-13  
  integral equation, 6-06  
  matrix, 14-37  
  of square matrix, 3-08, 3-09
- Algebra, fundamental theorem, 2-02, 7-15  
  of sentences, 11-05
- Algebraic equations, 2-01  
  roots of, 2-03, 2-04
- Allocation models, 15-31
- Amplifiers, power (table), 20-24
- Amplitude ratio, maximum, 22-03
- Analog data, continuous and discontinuous, transmission, 18-04
- Analogies, 20-07  
  d-c and a-c, 20-83  
  elements (table), 20-08  
  impedances (table), 20-55  
  and operation, 11-01, 11-02, 12-01
- Argument principle, 2-05, 7-11, 7-14
- Assignment problem, see *Operations research, linear programming*
- Attenuation diagrams, see *Bode diagrams*
- Autocorrelation function, 17-05, 17-24, 24-02  
  calculation, 24-06  
  computers, 24-10  
  of noise, 17-08  
  nonstationary inputs, 24-09, 24-10  
  range of  $\tau$ , 24-07  
  relation to spectral density, 24-06  
  sample length, 24-08  
  stationary inputs, 24-06
- Autocorrelation spectrum, 17-07, 17-14
- Backlash, 25-30, 25-34  
  compensation, 25-56, (table) 25-58  
  corrective techniques (table), 25-58  
  gears, 25-58
- Bandpass frequency, 22-03, 22-09
- Bandwidth, 19-16  
  effect of zeros, 23-15
- Baseband signal, in data transmission, 18-11
- Bayes theorem, 12-03
- Bendixson theorem, 5-19
- Bernoulli distribution, 13-04
- Bessel, equation, 7-24  
  function, 7-24  
  inequality, 6-05
- Beta function, 7-26
- Binary coding, see *Information theory, codes*
- Binary operation, sets, 1-08
- Binary relations, sets, 1-05, 1-07
- Binit, 16-08

- Binomial coefficient, 4-07
- Binomial distributions, 12-16, 13-04
- Bit, 16-08
- Block diagrams, 20-56
  - algebra, 20-61, 26-13
  - feedback control, 20-56
  - manipulations, 20-61, (table) 20-62
  - $z$ -transform, 26-13
- Bode diagrams, 21-29
  - application, 21-40
  - basic building blocks, 21-31
  - construction, 21-31
  - in drawing Nyquist plots, use of, 21-43
  - Nyquist criterion rephrased, 21-31
  - sampled data systems, 26-25
  - stability analysis, 21-30
- Boolean algebra, 11-01
  - algebra of sentences, 11-05
  - atom, 11-10
  - complete, 11-10
  - distributive, 11-10
  - implication, 11-06
  - and logic, 11-05
  - postulates, 11-02, 11-04, 11-10
  - propositional functions, 11-05
  - quantifiers, 11-06
  - relation to set theory, 11-04
  - sets, 1-10
  - Sheffer stroke operation, 11-10
  - Stone representation, 11-09
  - symbols (table), 11-02
  - symmetric difference, 11-03
- Boolean functions, 11-08
  - canonical form, 11-08
  - minimal polynomial, 11-08
- Boundary value problem, 14-61, 14-72
  - conformal mapping, 10-09
  - integral equations, 6-03
- Boxcar generator, 26-07
- Bridged-T network, 23-49
- Brock and Murray method, 14-60
- Butterworth-Thomson filters, 17-28
  - transitional, 17-30
- Canonical form, 1-08
  - Boolean functions, 11-08
  - matrix, 3-10, 3-12
  - partial differential equation, 14-68
- Carrier frequency shift, sensitivity to, 23-52
- Carrier systems, 20-79
  - modulators and demodulators, 20-80
  - suppressed, 20-82
- Cartesian product, sets, 1-04
- Casorati, theorem of Weierstrass and, 7-12
- Cauchy inequalities, 7-08
  - integral formulas, 7-07
  - integral theorem, 7-06
  - residue theorem, 7-13
  - Riemann equations, 7-05
- Cauchy method, 14-66
- Cauer's method of synthesis, 17-26
- Cayley-Hamilton theorem, 14-32
- Center, phase portrait, 25-38
- Central limit theorem, 12-18
- Channel, see *Information theory, channels*
- Channel capacity, see *Information theory, channels*
- Characteristic equation, 8-03, 20-28
  - difference equations, 4-04
  - differential equations, 5-05, 5-13
  - roots and stability, 20-53
  - sampled data system, 26-11
  - and stability, 21-03
- Characteristic functions, of the random variable, 12-14
- Characteristic polynomial, 3-11
  - for a matrix, 14-29
- Characteristic value, 3-11
  - integral equations, 6-02
- Check codes, 16-35; see also *Codes; Information theory*
- Chi squared ( $X^2$ ) distribution, 13-05, (table) 13-19
- Chopper bar controller, 26-05
- Circuit simplification, 20-07
  - aids, 20-10
- Classical methods of analysis, 20-28
- Clippinger and Dimsdale method, 14-60
- Closed loop, advantages, 19-12
  - frequency response, 22-03
  - from open loop response, 21-72
  - and open loop roots, relation between, 22-15
  - poles, 22-04
  - response, approximate, 21-80
  - roots, iterative process, 22-16
- Closed loop poles, relation to system characteristics (table), 23-12

- Closed loop poles and zeros, determination of open loop poles and zeros from the, 23-15
- Closed loop pole-zero location, 19-20
- Codes, see also *Information theory*, *codes*
- error correcting, 18-11
  - error detection, 18-10
  - for transmission, 18-07
  - reflected binary, 18-07
- Coefficients, equation, second order systems (table), 20-46
- Communication theory, 16-01; see also *Information theory*
- Compensation, 23-01; see also *Networks*, *d-c*, *compensation* and *Nonlinear systems*, *compensation*
- a-c systems, 23-48
  - backlash, 25-56, (table) 25-58
  - control systems, 23-01
  - d-c components, 23-18
  - d-c electric networks (table), 23-29
  - dead zone, 25-58
  - feedback, 23-21
  - load disturbances, 23-09
  - mechanical networks, 23-42
  - phase lag and lead, 23-18
  - rate feedback, 23-21
  - rate and lead network, 23-24
  - saturation, 25-53 (table), 25-54
  - typical nonlinear methods of (tables), 25-61, 25-62
- Complement, sets, 1-03
- Complex functions, 7-02
- analytic, 7-04
  - analytic continuation, 7-16
  - convergence, 7-08 [7-21]
  - defined by linear differential equations, definite integrals, 7-03, 7-15
  - elliptic, 7-18
  - examples of, 7-03
  - harmonic, 7-04
  - hypergeometric equation, 7-23, 7-25
  - identities, 7-04
  - integral theorems, 7-05
  - Laurent series, 7-08
  - mapping, 7-03
  - poles, 7-12
  - power series, 7-08
  - removable singularity, 7-12
- Complex functions, residues, 7-13
- Riemann surfaces, 7-17
  - singularities, 7-12, 7-13
  - zeros, 7-11
- Complex plane, 7-02
- plots, 20-54
  - typical, 20-67
- Complex variables, 7-01
- formulas, 7-02
- Component equations (table), 20-02
- Components, analogous (table), 20-08
- error detectors (table), 20-18
  - power amplifiers (table), 20-24
- Computer, analog, noise studies, 24-13
- autocorrelation, 24-11
  - for feedback control analysis, 19-20
- Computer, digital, controller, 26-03
- storage requirements and number of operations, differential equations, ordinary (table), 14-63
  - differential equations, partial (table), 14-87
  - eigenvalue problem (table), 14-48
  - matrix inversion (table), 14-27
  - systems of linear equations (table), 14-27
- Conformal mappings, 10-01, (table) 10-06, 21-12
- application to boundary value problems, 10-09
  - by elementary functions, 10-06
  - equivalence, 10-04
  - geometric meaning, 10-03
  - linear fractional transformations, 10-05
  - Schwarz-Christoffel, 10-08
  - theorems, 10-02
- Continuous elements, 26-01
- Contours of constant  $M$  and  $\alpha$ , 21-73
- Control accuracy, 19-06
- Control area, 19-06
- Control precision, 19-06
- Control ratio, 19-06
- peak, 19-16
- Control systems, see also *Feedback control*; *Nonlinear systems*; *Sampled data systems*; *Servomechanisms*
- a-c systems, 20-79
  - adjoint, 24-13
  - analysis, classical method of, 20-28

- Control systems, block diagrams, 20-56;  
 see also *Block diagrams*  
 compensation, 23-01; see also *Com-  
 pensation*  
 design in the presence of noise, 24-15  
 equations, examples, 20-05  
 first order, response, 20-36  
 fundamentals, 20-01  
 open-closed loop, 23-54  
 optimization, 24-18  
 representation, 20-01  
 response, by correlation technique,  
 20-27  
 response characteristics, 22-02  
 second order parameters, 20-44, 20-46  
 response, 20-39  
 stability, see *Stability*  
 types 0, 1, and 2, 20-67, 23-02  
 log magnitude diagrams, 23-01  
 representative plots, 20-67  
 typical applications (table), 20-69
- Controllers, continuous and sampled-data,  
 see also *Sampled data system*  
 comparison of, 26-29  
 periodic, 26-05, 26-25
- Controller synthesis, 19-13
- Convergence, power series, 7-08
- Convolution, Laplace, 9-08
- Convolution integral, application, 20-53  
 approximated, 20-53
- Corner frequency, 19-06, 22-09
- Correlation, and signal structure, 17-09  
 functions, see *Autocorrelation; Cross-  
 correlation*
- Correlation technique, system response,  
 20-27
- Cote's formulas for integration (table),  
 14-11
- Coulomb friction, 25-33, 25-56  
 simplified, 25-31
- Cramer rule, 3-09, 14-20
- Crank-Nicholson method, 14-84, 14-87
- Cross-correlation, function, 17-08, 17-24,  
 24-03  
 spectrum, 17-09
- Crossover frequency, phase margin at,  
 approximation, 22-43
- Curve fitting, 14-06; see also *Numerical  
 analysis, curve fitting*
- Curve fitting, least squares, 13-14  
 procedure, transient, 23-04
- Curve matching, 20-27
- Damping, absolute, 19-15  
 factor, and degree of stability, 23-06  
 ratio, 19-15, 20-44  
 approximation, 22-41  
 oscillatory transients, 20-50  
 viscous, dashpot, 20-02
- Darlington, Tchebysheff-, filters, 17-32
- Dashpot, 20-02
- Data transmission, 18-01  
 analog, 18-04, 18-21, 18-29  
 noise and error, 18-21  
 codes, see *Codes*  
 decoder, 18-08  
 digital, 18-05, 18-30  
 noise and error, 18-21  
 distances, 18-02  
 echoes, 18-22  
 and equalization, relationship, 18-24  
 tolerance, 18-26  
 encoding, 18-07  
 equalization, 18-22  
 error, influence of noise on, 18-21  
 standards, 18-05  
 facilities, 18-02  
 impairment, 18-18  
 modulation, 18-11  
 amplitude, 18-13  
 vestigial sideband, 18-14  
 frequency, 18-15  
 pulse code, 18-15  
 multiplexing, 18-11  
 auxiliary signal, 18-16  
 frequency division, 18-15  
 time division, 18-15  
 noise, see *Noise*  
 nominal effective band, 18-12  
 power density spectrum, see *Power  
 density spectrum*  
 prediction, 18-10  
 pulse shape, 18-12  
 quadrature components, 18-14  
 vestigial sideband transmission,  
 18-28  
 real time, 18-04  
 ripple, 18-25  
 tolerance, 18-26

- Data transmission, signal-to-noise ratio, 18-14
    - critical, 18-21
    - symbols, 18-01, 18-06
    - teletypewriter, 18-17
    - transfer functions, 18-23
    - vestigial sideband, 18-14, 18-28
  - D-c generator, transfer function, 20-14
  - D-c motor, backlash, 25-57
    - transfer function, 20-14
    - with inertia load, 20-06
  - D-c systems, compensating components, 23-18
  - Dead band, describing function, 25-22
  - Dead time, 19-06, 19-14
    - sampled data systems, 26-26
  - Dead time and rate stabilization, sampled data systems, 26-28
  - Dead zone compensation, 25-58
  - Decibel, gain- conversion, 21-44
  - Decoders, 16-07
  - Decoding, data transmission, 18-07
    - methods, 18-10
  - Decrement factor, 19-15
  - Delay time, 22-03
    - describing function, 25-23
  - Delta function, 8-06, 9-07
  - Demodulators, modulators and, 20-80
  - Descartes rule, 2-04
  - Describing functions, 19-20, 25-13
    - accuracy, 25-15
    - amplitude sensitive nonlinearities (table), 25-22
    - backlash (d-c shunt motor driving load), 25-30
      - simplified (high load damping), 25-31
    - compensation, 25-50
    - complex nonlinearities (table), 25-30
    - convergent point, 25-18
    - Coulomb friction, simplified, 25-31
    - dead band, 25-22, 25-24
    - delay time, 25-23
    - divergent point, 25-18
    - frequency variant, 25-20
    - granularity, 25-23, 25-28
    - hysteresis, 25-22, 25-25, 25-27
    - limitations, 25-15
    - log-magnitude-angle plane representation, 25-17
    - method of equivalent coefficients, 25-35
  - Describing functions, motor, acceleration limiting, 25-30
    - velocity limiting, 25-30
  - negative deficiency (types 1 and 2), 25-22, 25-26
  - Nichols charts, 25-51
  - Nyquist diagrams, inverted, 25-16
  - rapidly varying nonlinearities, 25-50
  - relation to frequency response, 25-51
  - relay, 25-23, 25-27
  - saturation, 25-22, 25-24
  - simplifying complex nonlinearities, 25-34
  - stability criteria, 25-15, 25-17
  - theory, 25-14
  - typical loci (table), 25-19
  - variable gain, 25-23, 25-29
- Design charts, relating open loop frequency response and transient response, 22-18
  - Design procedure, feedback control, 19-12
  - Detectors, error (table), 20-18
  - Determinants, 3-08
  - Difference, sets, 1-03
  - Difference-differential equations, 9-20
  - Difference equations, 4-01, 14-70
    - analysis of sampled data systems, 26-09
    - difference operator, 4-03
    - Laplace method, 4-08
    - rules for particular solutions (table), 4-06
    - solution of linear, for sampled data systems, 26-10
  - Differential equations, 5-01
    - characteristic equation, 5-05, 5-13
    - complex Fuchsian type, 7-23
    - dependent variable missing, 5-09
    - first order and first degree, 5-02
    - first order but not of first degree, 5-07
    - Fourier transforms, 8-16
    - graphical methods, 5-15
    - Heaviside operators, 8-05
    - independent variable missing, 5-09
    - integral transforms, 8-07
    - integro, 9-18
    - isoclines, method of, 5-15
    - Laplace transforms, 9-10
    - linear, 5-04

- Differential equations, numerical methods, 5-14  
 operational methods, 5-07, 8-05  
 ordinary, 5-01, 14-55  
 partial, 5-20; see also *Numerical analysis, differential equations, partial*  
 Picard, method of, 5-14  
 power series, 5-10  
 relaxation methods, 5-22  
 simultaneous, linear, 5-12  
   Heaviside operators, 8-05  
   Laplace transforms, 9-14  
 singular points, 5-16, 7-22  
 step-by-step integration, 5-14  
 successive approximations, 5-14  
 Taylor series solution, 5-10  
 undetermined coefficients, method of, 5-05  
   variation of parameters, 5-06  
 Differential operators, 8-01  
 Differentiation, numerical, 14-08  
   pure, 21-31, 21-33  
 Digital data, transmission, 18-05, 18-21, 18-30  
 Digital servo system, 26-20  
 Dimsdale and Clippinger method, 14-60  
 Dipoles, 23-15  
 Dirichlet problem, 5-20, 10-10, 14-72, 14-76  
 Disturbances, see *Noise*  
 Division, synthetic, 2-02  
 Donahue's analytical procedure, 21-52, (tables) 21-54, 21-57  
 Double description theorem, 3-16  
 Duhamel theorem, 8-06  
 Duty cycle, 19-18  
 Dynamic error coefficients, 19-17, 20-71  
 Dzung criterion, 21-10, 21-72  
 Echoes, in data transmission, 18-22  
 Eigenfunctions, expansion theorem, 6-07  
 Eigenvalue problem, 14-28  
 Eigenvalues, 3-11, 6-02  
   integral equations, 6-02  
   problem, variational interpretation, 6-08  
   theorems, 6-06  
 Eigenvector, 3-11, 14-28  
 Electrical elements (table), 20-08  
   transfer functions (table), 20-14  
 Electrical systems, equations, 20-02  
 Element, 1-01  
 Encoding, data transmission, 18-07  
 Enskog, method of, 6-13  
 Envelope delay distortion, 18-27  
   tolerance, 18-28  
 Erlang models, 15-76, 15-78  
 Error, calculation, 20-74  
   characteristics, type 0, 1, 2 systems (table), 20-69  
   coefficients, 20-70, (table) 20-73  
     acceleration, 20-71  
     calculation, 20-74  
     dynamic, 20-71, 20-74, 20-78, 23-03  
     position, 20-70  
     relative usefulness, 20-77  
     static, 20-70, 23-01  
     steady-state, 20-72  
     velocity, 20-70  
   correctors, (table), 20-26  
   detectors (table), 20-18  
   sinusoidal applied, 20-71  
 Error coefficients, 19-17, 23-12  
   and control system types, 23-02  
   and zeros, 23-15  
   criteria, 19-17  
   dynamic, 23-03, 23-04  
   ratio, 19-06  
   static, 23-01  
 Error-correcting codes, 18-11  
 Error criteria, mean square, 24-15  
 Error-detecting codes, 18-10  
 Error functions, 7-26  
 Euler constant, 7-25  
 Euler-Mascheroni constant, 7-26  
 Euler method, 14-57, 14-58, 14-59, 14-67  
 Evans, W. R., root locus method, 21-46  
 Existence theorem, 5-02  
 Extraneous signals, 19-13, 23-09; see also *Noise*  
 Extrapolation formulas (tables), 14-57, 14-58  
 Factor theorem, 2-02  
 Farkas lemma, 3-16  
 Feedback compensation, 25-62  
   d-c systems, 23-21  
 Feedback control, analysis (table), 19-20  
   computers, 19-20

- Feedback control, basic elements, 19-04  
 block diagrams, 20-57  
 common performance specifications (table), 19-14  
 computers, 19-20  
 controller, 19-06  
 definitions of terms (table), 19-04  
 design steps, 19-12  
 optimization, 19-19  
 performance specifications (table), 19-14  
 symbols (table), 19-01  
 symbols alternate (table), 19-11  
 synthesis methods (table), 19-19  
 system, 19-06  
 systems analysis, 20-01  
 systems types, 20-66  
 terminology (table), 19-01
- Feedback controller, 19-06
- Figures of merit, static error coefficients, 23-02
- Filters, Butterworth-Thomson, 17-28  
 transitional, 17-30  
 continuous, response to sampled data, 26-07  
 correlation in a derivative, 17-20  
 design data, 17-31  
 discrete data, 17-22  
 factorization problem, 17-14  
 with lag, 17-19  
 minimum phase condition, 17-11  
 network, 17-12, 17-14  
 network synthesis, 17-25  
 Caer method, 17-26  
 optimization, alternates to Wiener criterion, 17-24  
 optimum, 17-02, 17-13, 17-19, 24-16  
 decision theory, 17-24  
 phase response, of shaping network, 17-10  
 power spectrum, see *Power density spectrum*  
 prediction, 17-13  
 example, 17-16  
 nonlinear, 17-25  
 time, 17-14, 17-18  
 realizable network, 17-11  
 symbols, 17-04  
 Tchebysheff-Darlington, 17-32  
 transfer response, amplitude, 17-12
- Filters, transfer response, of the optimum, 17-13
- First order systems, Laplace transform pairs (table), 20-34  
 responses, 20-36
- Floyd's procedure, 22-44
- Focal point, 25-38
- Fourier, coefficients, 8-11  
 cosine, integral, 8-15  
 series, 8-11  
 transform, 8-15  
 integral, 8-09, 8-15  
 series, 8-10  
 sine series, 8-11  
 transform, 8-15  
 transforms, 8-09, 8-16  
 convolution, 8-12, 8-16  
 exponential, 8-16  
 finite, 8-10, 8-12  
 inverse, 8-12  
 properties, 8-16 [18-23]
- Fourier transform, data transmission, Fredholm, integral equation, 6-01, 6-06
- Frequency domain, noise, 24-12
- Frequency modulation, 18-15
- Frequency response, 19-07, 19-20  
 approximations (table), 22-41  
 bandpass frequency, 22-03, 22-09  
 closed loop, 22-03  
 corner frequency, 22-09  
 maximum, ratio of output to input, 22-08  
 maximum amplitude ratio, 22-03  
 open loop, 22-03 [22-18  
 and transient response, design charts, related parameters, 22-08  
 relation to transient response, 22-01  
 graphical techniques, 22-43  
 numerical techniques, 22-43  
 open loop design charts, 22-18  
 sampled data systems, 26-17  
 Samulon method, 22-48  
 Shannon sampling theorem, 22-48  
 transient response from, 22-44  
 from transient response, 22-47
- Frequency-sensitive networks, 25-58
- Friction, simplified Coulomb, describing function, 25-31
- Fuchsian type differential equation, 7-23
- Functions, 1-06

- Functions, Bessel, 7-24  
   beta, 7-26  
   cosine integral, 7-27  
   delta unit impulse, 8-06, 9-08  
   error, 7-26  
   exponential integral, 7-27  
   gamma, 7-25  
   Green, 6-04  
   Hankel, 7-24  
   incomplete gamma, 7-26  
   Jacobian elliptic, 7-18  
   Jacobi theta, 7-20  
   Legendre, 7-24  
   logarithmic integral, 7-26  
   Mathieu, 7-25  
   propositional, 11-05  
   Reimann zeta, 7-27  
   scalar product of, 6-05  
   sine integral, 7-27  
   transcendental, 7-25  
   unit, 8-06, 9-06  
   Weierstrass, 7-20  
   Weierstrassian analytic, 7-16  
   Weierstrass sigma, 7-20  
   Weierstrass zeta, 7-20  
   Whittaker, 7-25
- Gain, attenuation, 19-07  
   constant, 21-31, 21-32  
   crossover, 19-07  
   -decibel conversion, 21-44, 21-45  
   intermodulation effect on, 25-06  
   magnitude, 19-07  
   margin, 19-07, 19-16, 21-19
- Galvanometer, transfer function, 20-14
- Games, see *Operations research, games*
- Gamma function, 7-25  
   incomplete, 7-26
- Gauss formula (table), 14-12
- Gaussian distribution, 12-16, 12-18  
   entropy of, 16-43  
   errors, transmission, 18-06
- Gaussian noise, 16-44, 18-18  
   white, 18-19
- Gauss quadrature, 6-11
- Gears, backlash, 25-58
- Gill method, 14-59
- Gräffe method, 2-06
- Gram Schmidt orthogonalization method, 14-18, 14-27
- Granularity, describing function, 25-23, 25-28
- Graphical solution, determination of poles and zeros, 23-16
- Gray code, 18-07
- Green function, 6-04
- Guillemin method, 20-79
- Gyros, rate, 23-21
- Gyroscope, transfer function, 20-15
- Hamilton-Cayley theorem, 3-11
- Hankel function, 7-24
- Hartley, 16-08
- Hartley method, 14-49, 14-50
- Heat equation, 5-20, 14-70, 14-82
- Heaviside calculus, appraisal, 8-07  
   expansion formula, 9-10  
   theorem, 8-04  
   operator, 8-01  
   inverse, 8-02  
   power series, 8-04  
   unit function, 8-06, 9-06  
   step response to, 8-06
- Hermitian matrices, 3-14, 14-38, 14-42
- Heun formula, 14-58, 14-59, 14-67
- Hilbert matrix, 14-14
- Holding circuit, 26-07
- Homogeneous equation, 20-28
- Huffman coding, 16-12
- Hurwitz criterion, 21-71
- Hurwitz-Routh criterion, 2-05
- Hydraulic elements (table), 20-08  
   transfer functions (table), 20-16
- Hydraulic networks (table), 23-45
- Hysteresis, describing function, 25-22, 25-25
- If* operation, 12-01
- Impedance, complex (table), 20-55
- Implication, 11-06
- Impulse function, 9-07  
   Floyd procedure, 22-44  
   response to, 9-15
- Impulse response, optimum, 24-17
- Inequalities, linear, theorem, 3-17  
   system of linear, 3-14
- Inertia, rotational systems, 20-02
- Information theory, 16-01  
   amount of information, 16-08  
   bits and binitis, 16-08

- Information theory, channels, 16-06
- binary, 16-29
  - capacity of, 16-24, 16-32, 16-43
  - coding theorems, 16-26, 16-34
  - continuous noisy, 16-41
  - discrete, 16-06
  - discrete noiseless, 16-24
  - discrete noisy, 16-32, 16-26
  - Gaussian noise, 16-44
- codes, 16-04
- binary, 16-11
  - block coding, 16-22
  - codebooks, 16-04
  - codewords, 16-13
  - coding delay, 16-05, 16-38
  - coding theorems, 16-12, 16-17, 16-26, 16-34
  - error-free coding, binary erasure channel, 16-35
  - error probability, 16-38
  - Huffman coding, 16-12
  - minimax coding, 16-17
  - parity check, 16-36
  - pulse-code modulation, 16-40
  - Shannon-Fano coding, 16-12
  - transliteration, 16-05
- decoders, 16-06
- distribution of information, 16-09
- entropy, 16-11, 16-43
- messages, 16-03, 16-18
- minimax coding, 16-17
- mutual information, 16-26, 16-28, 16-42
- quantization, 16-39
- redundancy, 16-24
- sampling, 16-39
- theorem, 16-40
- segmentation, 16-03
- self-information, 16-08, 16-20, 16-21, 16-41
- sources, 16-03
- continuous, 16-39
  - controlled, 16-12, 16-16
  - discrete, 16-19
  - Markov, 16-20
  - rate of, 16-11, 16-21
  - simple discrete, 16-08
  - uncontrolled, 16-16
- Szilar-Kraft inequality, 16-13
- Integral compensation, saturation, 25-54
- Integral equations, 6-01
- Integral equations, adjoint, 6-06
- approximation methods, 6-10
  - boundary value problems, 6-03
  - degenerate kernels, method of, 6-11
  - eigenvalues, 6-02
  - theorems, 6-06
- Enskog, method of, 6-13
- expansion theorem, eigenfunctions, 6-07
- kernel, 6-02
- Rayleigh-Ritz quotient, 6-09
  - Ritz-Galerkin method, 6-12
  - steepest descent, method of, 6-16
- Integrals, approximation of, 6-10
- Integral transforms, 8-07
- convolution, 8-08
  - inverse, 8-08
- Integration, numerical, 14-11
- pure, 21-31, 21-33
- Integro-differential equations, 9-18
- Interpolation, numerical, 14-02
- Intersection, sets, 1-03
- Invariant, 1-08
- Inventory models, 15-21; see also *Operations research, inventory models*
- Isoclines, method of, 25-41
- Jacobian elliptic functions, 7-18
- Jacobi method, 14-42, 14-48
- Jacobi theta function, 7-20
- Jordan-Gauss method, 14-16, 14-27
- Jordan normal form, 3-11
- Jump resonance, 25-04
- Kirchhoff law, 20-05
- Laasonen method, 14-84, 14-87
- Lag, see also *Phase lag*
- quadratic, 21-31, 21-36
  - simple, 21-31, 21-34
- Lagrange formula, 14-02
- Lagrangian multipliers, 6-13, 15-12, 24-19
- modified, 15-13
- Lame equation, 7-24
- Laplace equation, 5-20, 14-70, 14-77
- Laplace method, difference equations, 4-08
- Laplace-Stieltjes transform, 8-18
- Laplace transforms, 8-17, 9-01, (table) 9-04
- asymptotic behavior of, 9-21

- Laplace transforms, convolution, 9-08  
 differential equation application, 9-10  
 feedback control, 20-30  
 inverse, 9-09, (table) 9-12, 20-31  
 pairs, 20-31, (table) 20-32  
   first order system (table), 20-34  
   second order system (table), 20-35  
 partial fraction expansion, 20-34  
 sampled data systems, 26-06, 26-08,  
   (table) 26-12  
 two-sided, 8-18  
 of unit functions, 9-06  
 and  $z$ -transforms, output, 26-15, (table)  
   26-12
- Laplacian operator, 6-03
- Laurent expansion, 7-11
- Laurent series, 7-10
- Lead, see also *Phase lead*  
 quadratic, 21-36  
 simple, 21-31, 21-34
- Lead-lag network, 20-16
- Least squares, 13-14
- Lebesgue integral, 12-09
- Legendre, elliptic integrals of, 7-19  
 equation, 7-23  
 function, 7-24  
 transform, 8-19
- Level changes, in data transmission,  
 18-29  
 tolerance in digital system, 18-30
- Levenberg method, 14-08
- Lewis servo, 25-62
- Liebmann extrapolated method, 14-73,  
 14-87
- Limit cycles, 5-19, 25-04, 25-38
- Linear equations, systems, 14-13; see  
 also *Numerical analysis, linear  
 equations, systems of*
- Linear programming, see *Operations re-  
 search, linear programming*
- Linear system, 19-07, 21-02
- Liouville theorem, 7-08
- Load disturbances, 23-09
- Load resonance, 25-56
- Load sensitivity, 19-15
- Log magnitude and phase diagram  
 (Bode diagram), 19-07
- Log magnitude-angle charts, 23-05; see  
 also *Nichols charts*
- Log magnitude-angle diagram, 19-07
- Log magnitude diagrams, 23-01  
 synthesis, 23-01, 23-08  
 transient curve fitting, 23-04
- Logarithmic plots, 20-56
- Loop gain, 19-07
- Loop input, 19-07
- Loop ratio, 19-08
- Mapping, see *Conformal mapping*
- Markov sources, 16-20
- Mathieu equation, 7-25
- Matrices, 3-01  
 adjoint, 3-09  
 canonical form, 3-10, 3-12  
 diagonal, 3-11  
 echelon form, 3-05, 3-06  
 equivalence, 3-09  
 Hermitian, 3-14  
 inverse, 3-07  
 inversion, 14-13; see also *Numerical  
 analysis, matrix inversion*
- Jordan normal form, 3-11  
 orthogonal, 3-13  
 rank, 3-07  
 similarity, 3-10  
 symmetric, 3-13
- Maximum operating conditions, 19-18
- Maximum principle, 7-08
- Maximum system error, 19-17
- Measurement devices, errors (table),  
 20-18
- Mechanical elements (table), 20-08  
 transfer functions (table), 20-13
- Mechanical-hydraulic networks (table),  
 23-45, 23-48
- Mechanical networks (table), 23-42,  
 23-48
- Mechanical-pneumatic networks (table),  
 23-46, 23-48
- Mesh analysis, nodal and, 20-11
- Messages, 16-03  
 effective number of, 16-18
- Milne procedure, 14-58
- Minimal polynomial, 11-08
- Minimax principle, 15-102
- Minimum error criterion, 19-16
- Minimum response time systems, 25-63
- Models, see *Operations research*
- Modulation, see *Data transmission,  
 modulation*

- Modulators and demodulators, 20-80  
 Monte Carlo method, 13-17, 15-18, 15-97  
 Morera theorem, 7-06  
 Morris escalator method, 14-18, 14-27  
 Motors, a-c, 23-48  
     acceleration limiting, describing function, 25-30  
     backlash, 25-57  
     optimum switching technique, 25-64  
     split drives, 25-58  
     velocity limiting, describing function, 25-30  
 Multiloop feedback systems, 23-26  
 Multiloop systems, 21-28, 21-69  
 Multiplexing, see *Data transmission, multiplexing*  
 Murray and Brock method, 14-60  
 Mutual information, 16-26, 16-42  
     distribution, 16-28  
  
 Nat, 16-08  
 Natural frequency, 20-44  
     damped, 22-10  
     undamped, 20-44  
 Natural period, 20-45  
     undamped, 20-44  
 Negative deficiency, describing function, 25-22, 25-23  
 Networks, a-c, bridged-T, 23-49  
     carrier frequency shift, 23-52  
     compensation, 23-48  
     parallel-T, 23-50  
     tachometers, 23-53  
     d-c, compensation, 23-18  
         electric (table), 23-29  
         hydraulic (table), 23-45  
         mechanical (table), 23-42  
     minimum phase, 17-12, 17-14, 21-30  
     pneumatic (table), 23-46  
     shaping, 17-10  
     synthesis, see *Filters*  
 Neumann problem, 10-10, 14-72  
 Neumann series, 6-15  
 Neville procedure, 14-03  
 Newton difference formulas, 14-08, (table), 14-09  
 Newton interpolation formulas, 14-04  
 Newton law, 20-05  
 Newton method, 2-04, 15-16  
 Nichols charts, 21-76, 23-05, 23-24  
  
 Nichols charts, multiple, 21-78  
     use with describing function, 25-51  
 Nodal and mesh analysis, 20-11  
 Nodal point, 25-38  
 Noise, 17-02, 19-13, 24-01  
     additive, 16-42  
     autocorrelation of, 17-08  
     data transmission, types, 18-18  
     error, high frequency, 23-08  
     error criteria, 24-15  
     Gaussian, 18-18  
     generators, 24-14  
     impairment of margin, 18-23  
     influence on error, 18-21  
     measurement of, 24-06  
     Rayleigh distribution, 18-20  
     sampled data systems, 26-25  
     system response, 24-11  
         computer methods, 24-13  
         frequency domain methods, 24-12  
         time domain methods, 24-11  
     systems design in the presence of, 24-15  
     Wiener-Hoft equation, 24-18  
 Nonlinearities, see also *Backlash; Compensation*  
     amplitude sensitive (table), 25-22  
     common types (table), 25-04  
     complex (table), 25-30  
     essential, 25-01  
     to improve system response, 25-61, (table) 25-62  
     intermodulation effect on gain, 25-06  
     jump resonance, 25-04  
     limit cycle, 25-04  
     optimum switching functions (table), 25-65  
     parasitic, 25-02  
     rapidly varying, 25-02  
     simplifying complex, 25-34  
     slowly varying, 25-02, 25-48  
     subharmonic generation, 25-05  
     typical complex (table), 25-30  
 Nonlinear methods of compensation (table), 25-62  
 Nonlinear problem, 14-84  
 Nonlinear systems, 25-01  
     common phenomena (table), 25-04  
     compensation, 25-48, 25-61  
         relay servomechanisms, 25-52  
     dynamic effect, 25-02

- Nonlinear systems, jump resonance, 25-04  
 saturation effects, typical types (table),  
 25-54  
 synthesis, 25-03  
 variable linear, 25-47
- Nonlinear systems analysis, see also *De-  
 scribing functions*; *Phase plane  
 analysis*  
 adjoint computing method, 25-48  
 analytical solutions, 25-43  
 linearization, 25-07  
 graphical, 25-10  
 useful algebraic approximations  
 (table), 25-09  
 normalized performance charts, 25-43  
 perturbation theory, 25-07  
 piecewise linear, 25-43  
 stability, ultimate, 25-47  
 Taylor series, 25-08  
 typical loci for amplitude sensitive  
 (table), 25-19  
 variable linear, 25-47
- Nonstationary processes, 24-02, 24-18
- Normal distribution, 13-05  
 function (table), 13-18
- Norton theorem, 20-11
- Not operation, 12-01
- Numbers, complex, 7-01  
 strong law of large, 12-12  
 table of random, 15-19  
 weak law of large, 12-07
- Numerical analysis, curve fitting, 14-06  
 Levenberg method, 14-08  
 differential equation, ordinary, 14-55  
 Adams and Bashforth method, 14-59,  
 14-63  
 Brock and Murray method, 14-60  
 computer storage requirements  
 (table), 14-63  
 Dimsdale and Clippinger method,  
 14-60, 14-63  
 Euler method, 14-57, 14-58, 14-59  
 extrapolation formulas (table), 14-57,  
 (table) 14-58, 14-63  
 extrapolation of zero grid size, 14-60,  
 14-63  
 fourth order method, 14-59, 14-63  
 Gill method, 14-59, 14-63  
 Heun formula, 14-53, 14-59, 14-63
- Numerical analysis, differential equation,  
 ordinary, Milne procedure, 14-58  
 number of operations (table), 14-63  
 Runge-Kutta method, 14-59, 14-60,  
 14-63  
 Simpson rule, 14-58  
 trapezoidal formula, 14-58
- differential equation, partial, 14-64  
 Cauchy method, 14-66  
 classification, 14-68  
 Euler method, 14-67  
 Heun method, 14-67  
 replacement by difference equations,  
 14-70
- differential equation, partial elliptic,  
 14-71  
 computer storage and time require-  
 ments, 14-87  
 Dirichlet problem, 14-72, 14-76  
 iteration method, 14-72  
 Laplace equation, 14-77  
 Liebmann method, 14-73, 14-87  
 Neumann problem, 14-72  
 Peaceman and Rachford method,  
 14-75, 14-87  
 relaxation method, 14-72  
 Richardson method, 14-73, 14-87
- differential equation, partial hyper-  
 bolic, 14-77  
 computer storage requirements,  
 14-87  
 explicit method, 14-78, 14-87  
 implicit method, 14-80, 14-87  
 roundoff errors, 14-80  
 triple diagonal systems, 14-81  
 truncation errors, 14-80, 14-87  
 von Neumann criterion for conver-  
 gence, 14-82
- differential equation, partial parabolic,  
 14-82, 14-87  
 computer storage requirements  
 (table), 14-87  
 Crank-Nicholson method, 14-84  
 explicit method, 14-83, 14-87  
 implicit equation, 14-84, 14-87  
 Laasonen method, 14-84, 14-87  
 truncation error, 14-87
- differentiation, 14-08  
 Newton difference formulas, 14-08,  
 (table) 14-09

- Numerical analysis, differentiation, Stirling formulas, 14-09, (table) 14-10
- eigenvalue problem, 14-28
- adjoint, 14-37
  - bounds on eigenvalues, 14-44
  - Cayley-Hamilton theorem, 14-32
  - computer storage requirements, 14-48
  - eigenvalues of special matrices, 14-43
  - escalator method, 14-33
  - Hermitian matrices, 14-38, 14-42
  - Jacobi method, 14-42, 14-48
  - number of operations for finding
    - eigenvalues and eigenvectors, 14-48
  - Sourian-Frame algorithm, 14-29
  - Sturm sequence, 14-36
  - triple diagonal method, 14-34, 14-48
- integration, 14-11
- Cote formula (table), 14-11
  - Gauss formula (table), 14-12
  - Simpson rule, 14-11
  - trapezoidal rule, 14-11
- interpolation, 14-01
- Lagrange formula, 14-02
  - Neville procedure, 14-03
  - Newton formula, 14-04
- linear equations, systems of, 14-20, 14-27
- computer storage requirements, 14-26
  - conjugate gradient method, 14-21, 14-24, 14-27
  - Cramer rule, 14-20
  - elimination method, 14-20, 14-27
  - gradient method, 14-23, 14-27
  - number of operations, 14-26
  - relaxation method, 14-22, 14-27
  - Seidel method, 14-21, 14-27
- matrix inversion, 14-13, 14-27
- digital computer storage requirements, 14-26
  - Gram Schmidt orthogonalization method, 14-18, 14-27
  - Jordan-Gauss method, 14-16, 14-27
  - modified matrix method, 14-19, 14-27
  - Morris escalator method, 14-18, 14-27
  - number of operations, 14-26
  - partition method, 14-17
- statistical analysis of experiments, 14-48
- balanced incomplete blocks, 14-51
- Numerical analysis, statistical analysis of experiments, factorial designs, 14-49, 14-54
- Hartley method, 14-49, (tables) 14-50
  - variance, 14-49, (tables) 14-51, 14-54
- Nyquist criterion, 21-09
- abbreviated, 21-18
  - applications, 21-15
  - conformal mapping, 21-12
  - diagram, 19-08
  - sampled data systems, 26-17, 26-19, 26-25
    - use in system compensation, 23-24, 23-26
  - disadvantage, 21-28
  - multiloop systems, 21-28
  - physical meaning of, 21-11
  - practical considerations in plotting diagrams, 21-17
  - principles, 21-13
  - rephrased in terms of Bode diagrams, 21-31
  - use of Bode diagrams in drawing plots, 21-43
- Open-closed loop control, 23-54
- Open loop, vs. closed loop, 19-12
- frequency response, 22-03
  - poles and zeros from the closed loop, 23-15
  - response, closed loop response from, 21-72
  - roots, relation between closed loop roots and, 22-15
  - transfer functions, polar plots of some common, 21-21
- Operational mathematics, 8-01
- Operational research, see *Operations research*
- Operations research, 15-01, 15-04
- allocation models, 15-31
  - bidding problems, 15-100
  - competitive problems, 15-99
  - control of the solution, 15-120
  - dynamic programming, 15-31
  - effectiveness, measure of, 15-08
  - executive problems, 15-03
  - games, 15-99
    - four-by-four, 15-113

- Operations research, games, minimax  
 principle, 15-102  
 mixed strategy, 15-102, 15-104  
 one-person, 15-101  
 rectangular games, theorems, 15-106  
 single strategy, 15-102  
 three-by- $n$ , 15-113  
 three-by-three, 15-111  
 two-by- $n$ , 15-109  
 two-by-two, 15-107  
 two-person, zero-sum, 15-101, 15-102,  
 15-104  
 zero-sum,  $n$ -person, 15-115
- implementation, 15-123
- inventory models, 15-21  
 dynamic, 15-30  
 elementary, 15-22  
 multistorage points, 15-30  
 with price breaks, 15-28  
 with restrictions, 15-30
- Lagrangian multipliers, 15-12
- linear programming, 15-31  
 assignment problem, 15-73  
 dual problem, 15-67  
 games, 15-114  
 geometric interpretation of, 15-63  
 short cut, 15-70  
 simplex method, 15-33, 15-41, 15-65
- models, 15-08, 15-10  
 data reduction, 15-119  
 sampling, 15-117  
 solutions, 15-10  
 testing, 15-115  
 types, 15-10
- Monte Carlo method, 15-18, 15-96
- Newton method, 15-16
- problem formulation, 15-05
- queuing theory, 15-73
- random walk problems, 15-18
- replacement models, 15-86  
 items that deteriorate, 15-86  
 items that fail, 15-89  
 Monte Carlo method, 15-96
- sampling, 15-117
- simplex method, 15-33
- transportation problem, 15-46  
 alternate evaluation method, 15-60  
 alternate optimum programs, 15-54  
 northwest corner rule, 15-48
- Operations research, transportation prob-  
 lem, solution of maximization  
 problems, 15-57  
 variations, 15-58
- waiting time models, 15-73  
 Erlang, 15-76, 15-78  
 holding time, 15-78  
 Lindley, 15-76  
 multiple channels, 15-82  
 Poisson input, 15-75  
 priority discipline, 15-79  
 sequencing models, 15-83  
 single channel, 15-77
- Operators, difference, 4-03  
 Heaviside, 8-01  
 integral, 8-07  
 power series, 8-04
- Optimization, criterion of, 17-13  
 feedback control, 19-19
- Optimum switching, functions (table),  
 25-65  
 techniques, 25-63, 25-64
- Or operation, 11-01, 11-03, 12-01
- Oscillation frequency, approximation,  
 22-42
- Oscillations, hard and soft, 25-38
- Output transforms for basic sampled  
 data systems, 26-15
- Overshoots, 19-10, 19-14  
 first and successive, 23-07
- Parallel-T network, 23-50
- Parity check, 16-36, 18-10
- Peaceman and Rachford method, 14-75,  
 14-87
- Peak overshoot, approximations, 22-41
- Performance index, 19-16  
 sampled data systems, 26-25
- Periodic controllers, 26-01, 26-05, 26-25;  
 see also *Sampled data systems*
- Phase angle, servomechanism scale, 21-36
- Phase crossover, 19-08
- Phase lag, compensation networks, 23-18  
 electric (table), 23-34  
 mechanical (table), 23-43  
 pneumatic (table), 23-47  
 network, 20-16
- Phase lag-lead, compensation networks,  
 23-20  
 electric (table), 23-38

- Phase lag-lead, mechanical (table), 23-43, 23-44  
     pneumatic components, 23-47
- Phase lead, compensation networks, 23-18  
     electric (table), 23-30  
     feedback, 23-25  
     mechanical (table), 23-42  
     pneumatic components (table), 23-46  
     network, 20-16
- Phase margin, 19-08, 19-16, 21-18  
     at crossover frequency, approximation, 22-43
- Phase plane analysis, 5-15, 25-36  
     construction methods, 25-39  
     isoclines, method of, 25-41  
     phase trajectory, 25-36, 25-38  
     relay servo, 25-40  
     singular points (table), 25-38
- Phase portrait, 25-36  
     significant characteristics of (table), 25-38
- Phase space, 25-39
- Physical laws, 20-04
- Picard, method of, 5-14
- Pneumatic components (table), 23-46
- Poisson distributions, 12-16, 13-04
- Poisson equation, 14-72, 14-73
- Poisson input, 15-75
- Polar plots, inverse, 21-20  
     of some common open loop transfer functions, 21-21
- Poles, closed loop, relation to system characteristics, 23-11, (table) 23-12  
     complex, dominant pair, 22-04, 22-07
- Pole-zero location, synthesis, 23-11
- Polynomials, 2-01  
     characteristic, 3-11
- Power amplifiers (table), 20-24
- Power amplifier saturation, 25-54
- Power density spectrum, 17-07, 24-04  
     autocorrelation, 17-07, 17-14  
     baseband signal, 18-11  
     cross-correlation, 17-09  
     phase response, shaping network, 17-10  
     pulse, 18-12  
     relation to autocorrelation, 24-06  
     shaped impulses, 17-09  
     vestigial sideband transmission, 18-14
- Power elements, selection of, 19-13
- Power series, 7-08
- Preamplifier saturation, 25-54
- Prediction, 17-03, 17-16; see also *Filters*  
     data transmission, 18-10  
     discrete data, 17-04  
     symbols, 17-04
- Primary feedback ratio, 19-08
- Probability, 12-01; see also *Statistics*  
     almost certain, 12-13  
     averages, 12-07, 12-13  
     Bernoulli distribution, 13-04  
     binomial distribution, 12-16, 13-04  
     central limit theorem, 12-13, 12-18  
     conditional, 12-03  
     continuous random process, 12-18  
     covariance function, 12-18  
     density, 12-09  
     distribution functions, 12-04  
     distribution of sums and averages, 12-13  
     expectation, 12-09, 13-03  
     expected value, 12-06, 12-07  
     independence, 12-04  
     joint distribution, 12-10  
     mean, 12-09, 13-03  
     normal distribution, 12-16, 12-18, 13-05  
     Poisson distributions, 12-16, 13-04  
     postulates, 12-01  
     random variables, 12-04, 13-02  
         bounded, 12-08  
         simple, 12-07  
     sentences, 12-01  
     stationary process, 12-18  
     strong law of large numbers, 12-12  
     theorems, 12-02  
     variance, 12-11, 12-12, 13-03  
     weak law of large numbers, 12-07
- Pulse code modulation, 16-40, 18-15
- Quadratic lag and lead, 21-31, 21-36
- Queuing theory, see *Operations research, waiting time models*
- Rachford, Peaceman and, method, 14-75, 14-87
- Ramp inputs, first order system, 20-34  
     second order system, 20-35
- Random inputs, see *Noise*
- Random processes, 24-02  
     stationary, 24-04
- Rate action, controller, 26-28

- Rate and lead network feedback, 23-24
- Rate feedback compensation, 23-21
- Rate gyros, 23-21  
inertial navigation, 23-10
- Rate network, a-c, 23-49
- Rayleigh distribution, of noise, 18-20
- Rayleigh-Ritz quotient, 6-09
- R-c networks (table), 23-29
- Real time, data transmission, 18-04
- Redundancy, 16-24
- Reflected binary code, 18-07
- Regression curve, 13-13  
function, 13-14
- Relations, 1-05  
binary, 1-05, 1-07  
equivalence, 1-07  
order, 1-09  
product of, 1-06
- Relaxation methods, 5-22, 14-22, 14-27
- Relay, describing function, 25-23
- Relay servo, compensation, 25-51, 25-52  
phase trajectories, 25-40, 25-42  
piecewise linear, 25-44
- Remainder theorem, 2-02
- Replacement models, 15-86
- Reset, component, integrating, 23-53  
divided, 25-58  
servo, 23-54
- Residues, 7-11
- Resolution, 19-18
- Response, see *Frequency response*;  
*Transient response*
- Response time, 19-08
- Richardson method, 14-73, 14-87
- Riemann mapping theorem, 10-04  
surfaces, 7-17  
theorem of, 7-12  
zeta function, 7-27
- Ripple, 18-25
- Rise time, 19-08, 19-14, 22-03  
approximations, 22-43
- Ritz-Galerkin method, 6-12
- Rolle theorem, 2-04
- Root locus, 19-20, 21-46  
angle condition, 21-47  
asymptote, 21-51  
common plots, 21-60  
compensation, 23-27  
construction aids, 21-58  
construction theorems, 21-49
- Root locus, construction theory, 21-47  
diagram, 19-09  
Donahue procedure, 21-52, (tables)  
21-54, 21-57  
interpretation of results, 21-68  
magnitude condition, 21-47  
mechanical angle summer, 21-58  
multiloop systems, 21-69  
practical considerations in drawing,  
21-52  
procedure, 21-48  
relative merits, 21-71  
sampled data systems, 26-25  
Spirule, 21-58
- Roots, of characteristic equation, and  
stability (table), 20-53  
closed loop, iterative process for de-  
termining, 22-16  
graphical method of determination,  
22-16  
of polynomials, 2-03, 2-04
- Roots of system equations, dynamic  
error coefficients and, 20-78
- Rotational systems, equations, 20-02
- Rouché theorem, 7-15
- Routh criterion, 21-05  
special cases, 21-07
- Routh-Hurwitz criterion, 19-20
- Routh-Hurwitz stability criteria, sam-  
pled data systems, 26-16
- Runge-Kutta method, 14-59, 14-60, 14-63
- Saddle point, 25-38
- Sampled data systems, 26-01  
Bode diagrams, 26-25  
constant overshoot loci, 26-23  
control area, 26-25  
controllers, comparison of continuous  
and, 26-31  
dead time, 26-26  
and rate stabilization, 26-28  
with delayed rate, 26-29  
design procedure, 26-20  
digital computer in the controller, 26-03  
Laplace transform, 26-06, 26-08  
linearity, 26-02  
minimum control area, 26-22, 26-26  
Nyquist diagram, 26-17, 26-25  
operational instruction, 26-21  
output transforms (table), 26-15

- Sampled data systems, performance,  
   charts, 26-25  
   criteria, 26-20  
   index, 26-25  
   regulator, 26-05  
   root locus, 26-25  
   smoothing, 26-07  
   stability criteria, 26-15  
     frequency response, 26-17  
     Routh-Hurwitz, 26-16  
   synthesis, 26-20  
     simple analog system, 26-23  
   transfer functions, 26-08  
   variables between sensing instants,  
     26-22  
   z-transform method, see *z-transforms*
- Samples, 13-02, 13-06, 26-01  
   analysis of, 26-06  
   variance, 13-07
- Sampling, 16-39  
   Shannon theorem, 16-40, 24-07
- Samulon method, 22-48  
   error, 22-49
- Saturation, compensation, 25-53, (table)  
   25-54  
   describing function, 25-22  
   in feedback, 25-54  
   power amplifier, 25-54  
   preamplifier, 25-54  
   servomechanisms, relay, 25-51  
   typical types of (table), 25-54
- Schedule and trim control, 23-54
- Schwarz-Christoffel mappings, 10-08
- Schwarz constants, 6-15  
   inequality, 6-05
- Second order systems, equation coef-  
   ficients (table), 20-46  
   parameters (table), 20-44  
   time responses (table), 20-35, 20-39
- Segmentation, 16-03
- Seidel method, 14-21, 14-27
- Self-information, of a symbol, 16-21  
   of continuous signals, 16-41  
   Markov process, 16-20
- Selsyns, 23-48
- Sentences, algebra of, 11-05  
   probability, 12-01
- Series, Fourier, 8-10  
   Laurent, 7-10  
   Neumann, 6-15
- Series, power, 7-08  
   Taylor, 7-09
- Series networks, a-c compensation, 23-48
- Servomechanisms, 19-09  
   a-c systems, 20-79, 23-48  
   error coefficients (table), 20-73  
   error correctors (table), 20-26  
   error detectors (table), 20-18  
   power amplifiers (table), 20-24  
   system type, 20-67; see also *Control*  
     *system, types*
- Sets, 1-01  
   binary operations, 1-08  
   binary relations on, 1-05  
   Cartesian product, 1-04  
   complement, 1-03  
   difference, 1-03  
   empty, 1-03  
   examples of, 1-01  
   intersection, 1-03  
   lattice, 1-09  
   of points, 1-02, 1-10  
   power set, 1-03  
   subsets, 1-02  
   union, 1-03
- Settling time, 19-09, 19-14, 22-03  
   approximations, 22-41
- Shannon-Fano coding, 16-12
- Shannon sampling theorem, 22-48, 24-07
- Sheffer stroke operation, 11-02, 11-10
- Signals, extraneous, see *Noise*
- Signal flow diagram, 20-57
- Signal-to-noise ratio, data transmission,  
   18-14
- Simplex technique, linear programming,  
   15-33  
   minimization problems, 15-45
- Simpson rule, 6-11, 14-11, 14-58
- sin  $x/x$  values (table), 22-52
- Singular points, 5-16, 7-22
- Singularities, 7-11
- Sinusoidal driving function, 20-54
- Smoothing, 17-01; see also *Filters*  
   and decoding, data transmission, 18-08  
   noise, 17-02  
   symbols, 17-04  
   Wiener theory, 17-02, 17-13
- Sources of information, see *Information*  
   *theory, sources*
- Souriau-Frame algorithm, 14-29

- Spectrum, autocorrelation and cross-correlation, see *Power density spectrum*
- Spirule, 21-58, 22-16
- Spring loaded gears, 25-58
- Stability, 19-09, 21-01
  - absolute, 20-29
  - Bode diagrams, see *Bode diagrams*
  - boundary theory, 21-72
  - classical approach, 21-02, 21-04
  - conditional, 21-19
  - contours of constant  $M$  and  $\alpha$ , 21-73
  - criteria,  $z$ -transforms, 26-15
  - degree of, 23-06
  - describing function, 25-15
  - Dzung criterion, 21-10, 21-72
  - gain margin, 21-19
  - Hurwitz criterion, 21-71
  - inverse polar plots, 21-20
  - margin, 21-72
  - multiloop systems, 21-28
  - Nichols charts, 21-76
  - Nyquist criterion, see *Nyquist criterion*
  - phase margin, 21-18
  - relation to characteristic equation, 21-03
  - relation to roots of characteristic equation (table), 20-53
  - root locus method, see *Root locus*
  - Routh criterion, see *Routh criterion*
  - sampled data systems, 26-15
  - unconditional, 21-19
  - unstable system, 21-01
  - Wall criterion, 21-72
- Stabilization, by internal feedback, 23-16
- Stabilization networks, for d-c compensation (table), 23-29
- Stabilizing networks, transfer functions (table), 20-15
- Standard deviation, 12-12, 13-03
- Static accuracy, 20-70
- Static error coefficient, 19-17, 20-70
- Stationary processes, 24-02, 24-15
- Statistics, 13-01; see also *Probability*
  - analysis of experiments, see *Numerical analysis, statistical analysis of experiments*
  - Bernoulli distribution, 13-04
  - binomial distribution, 13-04
  - bivariate distributions, 13-13
  - Statistics, computation, 13-08, (table) 13-09
    - confidence intervals, 13-10, 13-12, 13-15
    - curve fitting by least square, 13-14
    - distribution of sample moments, 13-09
    - expectation, 13-03
    - goodness of fit, 13-16
    - hypothesis testing, 13-10
    - maximum likelihood, method of, 13-16
    - mean, 13-03
    - mean deviation from the mean, 13-04
    - median, 13-04
    - midrange, 13-04
    - mode, 13-04
    - moments, 13-03
    - Monte Carlo method, 13-17
    - normal distribution, 12-16, 13-05, (table) 13-18
    - Poisson distribution, 13-04
    - probable error, 13-04
    - random variables, 13-02
    - regression curve, 13-13
    - relation to probability, 13-01
    - sample space, 13-02
    - sequential analysis, 13-16
    - standard deviation, 13-03
    - student  $t$  distribution, 13-06, (table) 13-20
    - unbiased estimate, 13-07
    - variance, 13-03, 13-07
    - $X^2$  distribution, 13-05, (table) 13-19
- Steady state, data for evaluating transfer functions, 20-23
  - equation terms (table), 20-55
  - errors (table), 20-71, 20-72
  - quasi-static assumption, 20-23
  - response, 20-01, 20-55
  - solution, 20-54
- Steady-state condition, 21-11
- Steady-state error, 19-09, 19-14
  - coefficient, 19-17
- Step function, 9-07
  - design charts, comparison of frequency response and transient response, 22-21
  - time responses (table), 20-48
- Step input, first order system, 20-34
  - second order system, 20-35
- Stieltjes integral, 12-09
- Stirling formula, 14-09, (table) 14-10

- Stirling numbers of the second kind, 4-07
- Stone, representation of Boolean algebra, 11-9  
 theorem, 11-10
- Student  $t$  distribution, 13-06
- Sturm sequence, 14-36
- Sturm theorem, 2-04
- Subharmonic generation, 25-05
- Summing points, 20-61
- Superposition, 20-10, 21-02  
 principle, 5-06, 8-06
- Suppressed carrier system, 20-82
- Symbols, alternate, 19-11  
 feedback control, 19-01  
 smoothing and filtering, 17-04
- Synthesis, 20-03  
 controller, 19-13  
 dominant pair of complex conjugate poles, 23-11
- Guillemin method, 20-79
- log magnitude diagrams, 23-01, 23-08
- network, see *Filters*
- Nyquist diagram, 23-27
- pole-zero location technique, 23-11
- root locus, 23-28
- selection of methods, 19-19
- System analysis, see *Control systems*
- System characteristics, additional poles and zeros, 23-13  
 relation to closed loop poles, 23-12  
 two complex conjugate poles, 23-12
- System response to noise, 24-11
- System stability, see *Stability*
- Systems, physical (table), 20-02  
 physical laws, 20-04
- Szilar-Kraft inequality, 16-13
- Tachometers, 23-21, 23-25  
 a-c, 23-48
- Tandem compensation, 25-62
- Taylor series, 7-09  
 differential equation solution, 5-10
- Tchebysheff-Darlington, filters, 17-32
- Tchebysheff inequality, 12-12
- Temple quotients, 6-15
- Terminology, feedback control, 19-01
- Thevenin theorem, 20-10
- Thomson, Butterworth-, filters, 17-28
- Time constant, 19-09  
 equivalent, approximation, 22-42
- Time constant, for second order system, 20-45
- Time delay, pure, 21-39
- Time domain, noise, 24-11
- Time to first peak, 22-02, 22-06
- Time parameter ratio, 20-45
- Time to peak, 22-10  
 approximation, 22-41
- Time response, approximate, 22-04  
 approximations, 22-41  
 convolution integral, 20-53  
 first order system, 20-36  
 second order system, 20-35, 20-39  
 transient modes (table), 20-48
- Time sequence, quasi-stationary, 17-02  
 stationary, 17-02
- Time series, nonstationary, optimum filter design, 17-24
- Total overshoot, 19-10
- Transfer function, 9-10, 9-19, 19-10,  
 (tables) 20-12, 20-13, 20-14, 20-16  
 in data transmission, 18-23  
 d-c motor, 25-57  
 electrical elements (table), 20-14  
 experimental evaluation, 20-23  
 hydraulic elements, 20-16  
 mechanical elements (table), 20-13  
 polar plots of some common open loop, 21-21  
 root locus plots of common, 21-60  
 sampled data systems, 26-08  
 typical, 20-12
- Transfer locus, 19-10
- Transfer response, amplitude, 17-12  
 filters, 17-26  
 of the optimum filter, 17-13
- Transformations, inverse, 3-07  
 linear, 3-03  
 rank, 3-03, 3-07
- Transformers, stabilizing, 23-21
- Transient error, 19-10
- Transient overshoot, 19-10, 19-14
- Transient response, 19-08, 20-01  
 approximations (table), 22-41  
 coefficients of terms, 22-12  
 data for evaluating transfer functions, 20-23  
 delay time, 22-03  
 effect of poles and zeros, 22-09  
 effect of significant real poles, 22-11

- Transient response, Floyd's procedure, 22-44  
 frequency response from, 22-47  
 from frequency response, 22-44  
 and open loop frequency response, design charts, 22-18  
 parameters, 22-02  
 relation to frequency response, 22-01  
 graphical techniques, 22-43  
 numerical techniques, 22-43  
 open loop design charts, 22-18  
 rise time, 22-03  
 settling time, 22-03, 22-10  
 $\sin x/x$  (table), 22-52  
 time to the first peak, 22-02, 22-06, 22-10  
 unit step input, 22-09
- Transients, input signals, 23-04  
 oscillatory, 20-50  
 time responses (table), 20-48
- Translation systems, equations, 20-02, (table) 20-08
- Transmission, of data, see *Data transmission*
- Transportation problem, see *Operations research, transportation problem*
- Trapezoidal formula, 14-11, 14-58
- Trapezoidal rule, 6-11
- Ultimately controlled variable, 19-10
- Union, sets, 1-03
- Unit function, 8-06, 9-06
- Unit impulse, first order system, 20-34  
 second order system, 20-35
- Unit impulse function, 8-06, 9-08  
 response to, 9-20
- Unstable system, 21-01
- Valve-pistons, transfer function, 20-16
- Variable gain, describing function, 25-23
- Variance, 12-11, 13-03
- Variance, analysis of, 14-49, 14-54
- Vector, components, 3-04  
 convexity, 3-15  
 coordinates, 3-04  
 half-space, 3-14  
 space, 3-01  
 subspace, 3-02
- Vibrating string, 5-21, 14-70, 14-77
- Volterra integral equation, 6-02, 6-15
- von Neumann criterion for convergence, 14-82
- Waiting time models, see *Operations research, waiting time models*
- Wall criterion, 21-72
- Wave equation, 5-20, 6-03
- Weierstrass, and Casorati, theorem of, 7-12  
 $\mathcal{O}$ -function, 7-20  
 sigma function, 7-20  
 zeta function, 7-20
- Weierstrassian analytic function, 7-16
- Weighting function, 9-19
- Whittaker function, 7-25
- Wiener-Hopf equation, 24-18
- Wiener theory of filtering, 17-02, 17-13
- Wronskian determinant, 5-07
- Wye-delta transformation, 20-10
- Zeros, 7-11  
 effect on bandwidth, 23-15  
 effect on error coefficients, 23-15
- $z$ -transforms, 26-11, (table) 26-12  
 block diagram algebra, 26-13  
 closed loop, 26-24  
 design procedure, 26-20  
 hidden oscillations, 26-11  
 inverse, 26-13  
 Laplace and (table), 26-12  
 output (table), 26-15  
 stability criteria, 26-15