

N O T I C E

THIS DOCUMENT HAS BEEN REPRODUCED FROM
MICROFICHE. ALTHOUGH IT IS RECOGNIZED THAT
CERTAIN PORTIONS ARE ILLEGIBLE, IT IS BEING RELEASED
IN THE INTEREST OF MAKING AVAILABLE AS MUCH
INFORMATION AS POSSIBLE

DIGITAL FILTERING AND PROCESSING

by

TRANSFORM TECHNIQUES

VOLUME I

FINAL REPORT

30 June 1968

CONTRACT NAS8-20308

NON-LINEAR AND LINEAR DIGITAL
FILTERING TECHNIQUES

Prepared by
E. H. GERMAN JR.

for

GEORGE C. MARSHALL
SPACE FLIGHT CENTER
Huntsville, Alabama

THE BENDIX CORPORATION
Communications Division
Baltimore, Maryland 21204

For Information

contact

D. H. MacDONALD
Government Marketing Department
Phone 1-823-2200, Ext. 626

or

E. H. GERMAN JR.
Advanced Research Department
Phone 1-823-2200, Ext. 715

THE BENDIX CORPORATION
COMMUNICATIONS DIVISION
BALTIMORE, MARYLAND 21204

ABSTRACT

A complete programming and engineering level discussion of the various aspects of digital filtering is given. Relations between analog and digital signals and systems are discussed for the instance where data is derived from analog situations. Limitations on simulation of analog systems are presented. Various transform techniques are given and compared. Linear, time invariant digital filters are defined, conditions for realizability given, and forms of realization enumerated. A wide variety of synthesis methods for both recursive, and non-recursive digital filters are given with advantages and limitations. Filtering with the 2 component complex modulation is defined and complex filter synthesis methods developed. Descriptions of basic statistical concepts are presented. Also discussed are methods of statistical spectrum analysis and linear statistical function estimation. Numerous applications to digital systems are described. The modern fast discrete Fourier transform is described in detailed and integrated, where applicable, into the text.

PREFACE

This document, in two volumes, is the final report on the NASA contract NAS8-20308, "Non-Linear and Linear Digital Filtering Techniques". Volume I is devoted to the general theory of digital filtering and processing. Volume II is primarily a compendium of many machine programs of techniques developed in Volume I along with guidelines for proper usage.

A main intent of Volume I has been the expression of the research under the contract within a context that will lead to clarity for a relatively wide audience, not just those persons who already have a basic mastery of the general subject. With this desire, certain consequential constraints are placed on the contents. Mathematical rigor has had to be relaxed to allow ideas to be grasped more easily; though with the present burgeoning and rearranging of knowledge one might wonder if rigor quite deserves the attention that it often holds. Considerable material that is relatively well-known and available elsewhere; such as transform theory, elements of the theory of Fourier series, and statistical concepts has been included of necessity to make the presentation as self-supporting as possible.

Another basic approach in the development of this volume has been the emphasis on linear digital filters (difference equations) and numerical processing as a subject worthy of distinct treatment, a discipline by itself. With the wide utilization of modern digital computers of high speed in the handling of large quantities of data, the rapid growth of digital data transmission, and the impact of micro-electronic and integrated circuit technology on signal and data processing, this is not an unwarranted attitude. Much of the work on digital filtering may be found scattered in works on sampled-data control systems, numerical analysis, and statistical estimation theory. Emphasis on digital filter theory and synthesis, per se, has only appeared prominently in the technical journals of the past decade. It is hoped that the collection of pieces into one document will provide a sum greater than that of the individual pieces.

Central to the text is usage of the continuous Fourier, discrete Fourier, Laplace, and z transforms to allow algebraic manipulation of various processes and enable interpretation of sampled-data filtering and processing in terms of frequency content and spectral shaping which often has more direct physical meaning than working in the "time" domain. When dealing with sampled continuous processes, the "bandlimited function" approach is used extensively.

Some original work is believed to be incorporated between the covers of this book: modification of the z transform filter technique to include a usually ignored constant, development of low-pass to band-pass and band-reject transformations directly in the digital domain, development of digital quadrature (or "complex") filtering relations paralleling those for continuous processes, and synthesis of quadrature, or complex, filters from analog prototypes.

The fast discrete Fourier transform algorithm has been integrated into computational procedures, where applicable, to enhance their utility.

As much of the material has been used by the author in day-to-day practice for many years it has been quite difficult to assign credit to particular researchers for many of the techniques used. A bibliography has been included, citing important techniques, where the reader may look for further information on particular subjects.

Two topics of interest not included within the covers are recursive statistical filters and data and filter coefficient quantization and truncation effects. The first was not included as it was somewhat outside of the scope of the book; the second is a serious omission and has been caused by lack of time.

The author wishes to thank those people, who have so graciously supplied assistance and encouragement; John Kmetzo and Knowles Little for their efficient programming of many techniques and intelligent and diligent questioning and criticism of work in progress which led to a much greater clarity of ideas; and to John Martin, the group head, who has allowed the author the freedom within an industrial organization, that is so necessary with work of this kind.

The NASA technical representative Ronald J. Graham of the NASA, Huntsville Computation Laboratory is deserving of special thanks for his encouragement and for the freedom of research within the contract scope that he has allowed.

Also, the author wishes to acknowledge his appreciation for the attractive cover and graphs made by Mrs. Shirley Brown and to JoAnn Schepf and Barbara Lochner for their valuable assistance in typing this volume and organizing the many clerical aspects associated with it.

Finally, the author extends his heartfelt appreciation to his wife, Beverly A. German for her forbearance and encouragement which have made this endeavor possible.

TABLE OF CONTENTS

	Page
1. INTRODUCTION	1-1
2. SAMPLED OR "DIGITAL" DATA REPRESENTATION	2-1
2.1 FOURIER TRANSFORMATION (ANALOG)	2-1
2.1.1 Bandlimited Functions and Sampling	2-2
2.1.2 "Aliasing" Due to Sampling at a Rate Less Than Twice the "Bandlimit"	2-6
2.1.3 Subharmonic Sampling	2-9
2.1.4 Bandlimited Functions and "Impulse Sampling"	2-11
2.2 LAPLACE TRANSFORMATION	
2.2.1 Generalization of Bandlimited Function Representation to Functions Which Are Not Fourier Transformable	2-17
2.3 THE z TRANSFORM	2-20
2.3.1 Initial and Final Value Theorems	2-22
2.4 THE INVERSE z TRANSFORM	2-23
2.4.1 Direct Inverse Transformation	2-23
2.4.2 Inversion by Long Division	2-24
2.4.3 Inverse z Transform as an "Impulse Response"	2-25
2.5 GEOMETRICAL INTERPRETATION OF CONNECTION AMONG FOURIER, LAPLACE, AND z TRANSFORMS	2-26
2.5.1 The "s" To " z " Place Mapping Induced by Sampling	2-27
2.6 THE DISCRETE FOURIER TRANSFORM	2-30
2.6.1 Definition of the Discrete Fourier Transform	2-31
2.6.2 Interpretation of the "Time" and "Frequency" Indices " j " and " n "	2-35
2.6.3 Relationships Between Various Forms of the Discrete Fourier Transform	2-36
2.6.4 Effect on the Transform When an Output Sequence is Real	2-41
2.6.5 Relation Between Continuous and Discrete Fourier Transforms	2-43
2.6.6 Fourier Series and the Discrete Fourier Transform	2-47
2.6.7 Use of the Discrete Fourier Transform for the Evaluation of Fourier Series	2-49
2.7 THE CONVOLUTION THEOREM: CONTINUOUS AND DISCRETE WAVEFORMS	2-50
2.7.1 Continuous Convolution Theorem	2-50
2.7.2 Discrete Convolution Theorem	2-51
2.8 TABULAR COMPARISON OF VARIOUS CONTINUOUS AND DISCRETE FOURIER TRANSFORM RELATIONS	2-54
3. DIGITAL FILTER CONCEPTS	3-1
3.1 DEFINITION OF A LINEAR TIME INVARIANT, SAMPLED DATA OR "DIGITAL" FILTER	3-1
3.2 DESCRIPTION OF DIGITAL FILTERS BY THE z TRANSFORM	3-2
3.3 FREQUENCY BEHAVIOR OF THE z TRANSFER FUNCTION	3-4
3.3.1 Complex Frequency Response	3-4
3.3.2 Transfer Function Definition Without Recourse to Complex Numbers	3-5
3.3.3 Transfer Function Time (Or Phase) Reference	3-8

TABLE OF CONTENTS (Continued)

	Page
3.4 DEFINITION OF THE "IMPULSE RESPONSE" (AND FILTER DEFINITION FROM THE "IMPULSE RESPONSE")	3-9
3.4.1 Filter Input-Output Relation in Terms of the Impulse Response (Convolution)	3-10
3.4.2 Filter Definition and Stability From the Impulse Response	3-11
3.4.3 Filter Stability Via the Difference Equation	3-14
3.5 SOME PROPERTIES OF DIGITAL FILTERS	3-15
3.6 SERIES FILTERS	3-16
3.7 REALIZATION METHODS	3-18
3.7.1 Direct Programming	3-18
3.7.2 Factorization	
3.7.3 Partial Fraction	3-20
3.7.4 Simple Product	3-21
3.7.5 Discrete Fourier Transform Realization	3-22
3.7.5.1 The "Select-Saving" Method	3-24
3.7.5.2 The "Overlap-Adding" Method	3-24
3.7.5.3 Determination of N to Minimize the Number of Computations	3-25
3.7.5.4 Reduction of Operations for Real Filters and Sequences	3-26
3.7.5.5 Comparison with the Direct Programming Method	3-26
3.8 REALIZATION OF SOME NON-RECURSIVE FILTERS AS RECURSIVE FILTERS	3-27
3.9 "LEGAL" FILTER FORMS	3-30
3.10 BANDWIDTH AND TRANSIENT RESPONSE TIME, OR BANDWIDTH OF A SAMPLED PULSE WAVEFORM	3-31
3.11 THE DISCRETE FOURIER TRANSFORM AS A DIGITAL FILTER	3-36
3.12 A USEFUL SUMMATION RELATION	3-37
4. NUMERICAL INTEGRATION OF LINEAR, CONSTANT COEFFICIENT, INTEGRO-DIFFERENTIAL EQUATIONS	4-1
4.1 INTRODUCTION TO ESSENTIALS OF DESCRIPTION OF CONSTANT COEFFICIENT SYSTEMS	4-1
4.2 NON-ZERO INITIAL CONDITIONS	4-4
4.3 CONVERSION OF A "HIGH PASS" TRANSFER FUNCTION TO A CONSTANT PLUS "LOW PASS" TRANSFER FUNCTION	4-6
4.4 FILTERING OF A BANDLIMITED FUNCTION REPRESENTED IN TERMS OF SAMPLED VALUES	4-8
4.4.1 Extension to High Pass Transfer Function	4-11
4.5 DEFINITION OF SIMULATION OF A LINEAR CONSTANT COEFFICIENT SYSTEM	4-13
4.5.1 Simulation and Numerical Integration	4-14
4.5.2 Exact Simulation is Impossible in Principle Except for Bandlimited Transfer Functions	4-14
4.6 RELATION BETWEEN ANALOG AND DIGITAL TRANSFER FUNCTIONS FOR FOURIER COEFFICIENT INTEGRATION FORMULA	4-17
5. NON-RECURSIVE FILTERS AND FILTER LIMITATIONS	5-1
5.1 REDUCTION OF THE GENERAL LINEAR DIGITAL FILTER FORM TO NON-RECURSIVE FORM	5-2
5.1.1 "Cosine" and "Sine" Central Filter Simplifications	5-6

TABLE OF CONTENTS (Continued)

	Page
5.2 FOURIER SERIES FUNDAMENTALS	5-7
5.2.1 Orthonormal Expansions and Their Relation to Fourier Series	5-13
5.2.2 Completeness of an Orthonormal Set	5-17
5.3 RELATION BETWEEN FOURIER SERIES EXPANSIONS AND LINEAR DIGITAL FILTER TRANSFER FUNCTIONS (INCLUDING RECURSIVE FILTERS)	5-18
5.3.1 The Transfer Function Approximation Problem and the Determination of τ	5-19
5.3.2 Fourier Least-Squares or "Dirichlet" Weights	5-20
5.4 CONVERGENCE RATES OF THE REAL AND IMAGINARY PARTS OF THE TRANSFER FUNCTION, OR MAGNITUDE OF FOURIER COEFFICIENTS IN GENERAL	5-23
5.5 FOURIER LEAST SQUARES APPROXIMATION USING DISCRETE FOURIER TRANSFORMS INCLUDING EXACT FITTING AT $N + 1$ EQUI-SPACED POINTS	5-25
5.6 SUPPRESSION OF THE "GIBBS PHENOMENA"	5-30
5.6.1 Separation of Discontinuities from Arbitrary Functions	5-31
5.6.2 Basic Approaches to Removal of the "Gibbs Phenomena"	5-33
5.6.3 "Filtering" The Transfer Function	5-35
5.6.4 "Filtering" of Non-Central Transfer Functions	5-43
5.7 PARTICULAR ("FILTER FILTER") METHODS	5-44
5.7.1 Fejer or Linear Weighting	5-44
5.7.2 Lanczos Weighting	5-46
5.7.3 Cosine Squared Weighting	5-47
5.7.4 Error Analysis of "Filtered" Filters	5-48
5.7.5 Ripple Errors of "Filtered" Filters	5-52
5.8 EXPANSION COEFFICIENTS AND WEIGHTS FOR THE DISCONTINUITY REMOVAL FUNCTIONS (OR "CENTRAL" BANDPASS AND DIFFERENTIATION FILTERS)	5-53
5.8.1 "Cosine" Low-Pass Filter	5-54
5.8.2 "Cosine" High-Pass Filter	5-55
5.8.3 "Sine" Low-Pass Filter	5-55
5.8.4 "Sine" High-Pass Filter	5-56
5.8.5 "Sine" Differentiating Filters	5-57
5.9 CERTAIN NON-RECURSIVE FILTER GENERALITIES	5-58
5.9.1 Cosine Filters (Low and High Pass)	5-58
5.9.2 Sine Filters (Low and High) Pass	5-60
5.9.3 Bandpass Filters (Sine and Cosine)	5-62
5.10 GRAPHICAL PRESENTATION OF NON-RECURSIVE FILTER DESIGNS	5-62
5.11 RELATIONSHIP OF AMPLITUDE AND PHASE ACCURACIES TO REAL AND IMAGINARY PART ACCURACIES	5-75
5.12 SYNTHESIS OF NON-RECURSIVE FILTERS BY POLYNOMIAL TECHNIQUES	5-78
5.12.1 Conversion of the Non-Recursive Filter Transfer Function to a Polynomial Form	5-78
5.12.2 Reconversion From Polynomial Form to Fourier Form by Determination of the Transforming Matrix Elements	5-83
5.12.3 Reconversion From Polynomial Form to Fourier Form in an Easily Machine Programmable Manner	5-85

TABLE OF CONTENTS (Continued)

	Page
5.12.4 Exact Transfer Function Fitting by Lagrange Interpolation	5-88
5.12.5 Convergence of Polynomial Interpolation	5-94
6. RECURSIVE FILTERING TECHNIQUES	6-1
6.1 DIGITAL FILTERS THROUGH APPROPRIATE z TRANSFORMATION OF ANALOG FILTERS	6-2
6.2 DIRECT z TRANSFORMATION	6-6
6.2.1 Mechanization of the Computation of the z Transform for m^{th} Order Poles	6-9
6.2.2 An Example of Digital Filter Design by the z Transform	6-15
6.3 THE DELAYED z TRANSFORMATION	6-20
6.4 MECHANIZATION OF INITIAL CONDITION OR "TRANSIENT" COMPUTATIONS	6-22
6.4.1 Example of Delayed z Transform Utilization to Compute Transient Responses	6-24
6.5 DIGITAL INPUT-OUTPUT RELATION FOR LINEAR, CONSTANT COEFFICIENT SYSTEMS WITH "IMPULSE SAMPLED" INPUTS	6-26
6.6 BI-LINEAR z TRANSFORMATION (DATA PROCESSING FORM)	6-29
6.6.1 General Low-Pass to Band-Pass, Band-Stop, and High-Pass Transformation of an Analog Filter Via the Bi-linear z Transform	6-34
6.6.2 General Transformation of Normalized Low-Pass Digital Filter Designs to Low/High-Pass, Band-Pass and Band-Stop Filters in the Digital Filter Domain	6-39
6.7 BI-LINEAR z TRANSFORM (SPECIAL FORM FOR NUMERICAL INTEGRATION)	6-46
6.8 EXACT FITTING OF A TRANSFER FUNCTION BY A RECURSIVE FILTER AT $Q + 1$ POINTS	6-48
6.8.1 Exact Fit at Equi-Spaced Points	6-51
6.8.2 Stability of Exact Fit Method	6-54
7. SYNTHESIZING OF RECURSIVE NUMERICAL INTEGRATION DIGITAL FILTERS WITH SMALL DELAY AND DATA RECONSTRUCTION	7-1
7.1 DATA RECONSTRUCTION TECHNIQUES	7-1
7.1.1 Time Invariant	7-2
7.2 SAMPLE AND HOLD RECONSTRUCTION	7-3
7.2.1 Zero Order Sample and Hold	7-4
7.2.2 First Order Sample and Hold (Delayed)	7-8
7.2.3 First Order Sampled and Hold (Extrapolation)	7-10
7.3 NUMERICAL INTEGRATION BY INPUT RECONSTRUCTION WITH POLYNOMIALS	7-13
7.3.1 Conversion of Hold Systems to Digital Recursive Filters	7-14
7.4 BANDPASS FILTER RECONSTRUCTION FUNCTIONS	7-17
7.4.1 Design of a Reconstruction Low-Pass Filter	7-18
7.4.2 Using an Analog Low-Pass Filter for Recursive Digital Filter Synthesis	7-26
7.4.3 Example of Digital Filter Design Using Butterworth Reconstruction Filter	7-28
7.5 ILLUSTRATIONS OF DESIGN TECHNIQUES	7-30

TABLE OF CONTENTS (Continued)

	Page
8. "COMPLEX" FILTERING	8-1
8.1 CERTAIN COMPLEX FILTER RELATIONS	8-1
8.2 FUNCTIONAL FORM OF "HIGH FREQUENCY", NARROW-BAND SEQUENCES AND IMPULSE RESPONSES	8-8
8.2.1 "Complex Modulation" of a Narrow Band, High Frequency Sequence	8-9
8.3 DIGITAL FILTERING OF A NARROW-BAND, HIGH FREQUENCY SIGNAL USING THE COMPLEX MODULATION	8-12
8.3.1 Impulse Response Form of Digital Complex Modulation Filtering	8-18
8.3.2 Symmetrical Filter Simplifications	8-19
9. DIGITAL FILTERING OF THE "COMPLEX MODULATION" OF ANALOG SIGNALS	9-1
9.1 SAMPLED REPRESENTATION OF ANALOG BANDLIMITED HIGH-FREQUENCY SIGNALS	9-2
9.2 NUMERICAL INTEGRATION WITH NARROW-BAND INPUT SPECTRA BY "COMPLEX" DIGITAL FILTERS	9-5
10. SYNTHESIS OF "COMPLEX" DIGITAL FILTERS FROM NARROW-BAND ANALOG FILTERS	10-1
10.1 SYNTHESIS OF NON-RECURSIVE "COMPLEX" FILTERS	10-1
10.2 SYNTHESIS OF RECURSIVE COMPLEX DIGITAL FILTERS FROM AN ANALOG BANDPASS PROTOTYPE IN RATIONAL FORM	10-3
11. STATISTICAL CONCEPTS	11-1
11.1 BASIC DESCRIPTION OF "RANDOM" PROCESSES	11-1
11.1.1 Probability Densities	11-1
11.1.2 Statistical Independence	11-3
11.1.3 Averages	11-3
11.1.4 Stationary and Ergodic Processes	11-7
11.1.5 Definition of the Fourier Transform and the Energy Spectrum For a Single Record	11-8
11.1.6 Definition of the Power Spectrum and Auto-Correlation Function For a Single Record	11-10
11.1.7 Power Spectrum and Auto-Correlation Functions of Random Processes	11-13
11.1.8 Cross-Power Spectra and Cross-Correlation Functions For Random Processes	11-16
11.1.9 Examples of Auto-Correlation Functions and Their Spectrums (and vice versa)	11-18
11.1.10 Relationship of Finite Interval Spectrum to Infinite Interval Spectrum	11-19
11.1.11 The Effect of a Linear Time Invariant System on the Input Power Spectrum	11-21
11.2 SAMPLED DATA SPECTRUMS AND AUTO-CORRELATION FUNCTIONS	11-22
11.2.1 Definition of the Sequence Energy Spectrum and Correlation Function	11-24
11.2.2 Definition of the Sequence Power Spectrum and Auto-Correlation Function	11-27
11.2.3 Sequence Power Spectrum and Auto-Correlation Function For Random Processes	11-29

TABLE OF CONTENTS (Continued)

	Page
11.2.4 Relation Between Analog and Sampled Spectra and Auto-Correlation Functions	11-31
11.2.5 Sequence Cross Spectrums and Cross-Correlation Functions	11-32
11.2.6 Modification of a Stationary Sampled Sequence Power Spectrum by a Digital Filter $H_z(\omega)$, and the Effect Upon "White" Noise as an Example	11-34
11.2.7 Modification of the Sampled Sequence Average by a Digital Filter $H_z(z)$	11-36
11.2.8 Use of the Discrete Fourier Transform to Evaluate the Sampled Spectrum For Deterministic Waveforms	11-38
12. DIGITAL SPECTRUM ESTIMATION	12-1
12.1 ESTIMATION THROUGH THE USE OF "LAGGED" PRODUCTS	12-1
12.1.1 Use of the Fast Fourier Transform Algorithm to Reduce Computation Time	12-24
12.2 ESTIMATION THROUGH THE USE OF DIGITAL FILTERS	12-27
12.3 COMPLEX DEMODULATE SPECTRUM ESTIMATION	12-36
12.4 LIMITATIONS OF A PERIODOGRAM ESTIMATE	12-41
12.5 INTERPRETATION OF THE RELATIVE STABILITY	12-43
13. LINEAR STATISTICAL FILTERS	13-1
13.1 ESTIMATION OF A FUNCTION OF KNOWN FORM OR REPRESENTATION	13-2
13.1.1 Estimation of the Function Directly	13-3
13.1.2 Estimation of the Coefficients of the Function Representation	13-9
13.1.3 Equivalence of Function and Coefficient Estimation	13-10
13.1.4 Generalization to Estimators of the Effect of an Arbitrary Linear Operator	13-11
13.1.5 Derivative Estimators	13-13
13.1.6 Explicit Expression of Certain Estimators and Their Square Errors	13-15
13.1.7 Specialization of the Estimators to an Important Case	13-15
13.1.8 Estimator Examples	13-21
13.1.9 Representation Error	13-21
13.1.10 Orthogonal Representations	13-23
13.1.11 Graphical Presentation of Polynomial Representation Estimator Properties With Stationary, Uncorrelated, Noise	13-24
13.2 WEINER FILTERS	13-30
13.2.1 Specialization of the Solution to an Important Case	13-33
14. THE FAST FOURIER TRANSFORM AND VARIOUS MECHANIZATIONS	14-1
14.1 BASIS OF THE FAST TRANSFORM ALGORITHM	14-2
14.2 THE BASE FACTOR 2 ALGORITHM	14-5
14.2.1 Time Decimation With a Bit Reversed Transform	14-6
14.2.2 Time Decimation With a Bit Reversed Input Sequence	14-12
14.2.3 Frequency Decimation With a Bit Reversed Transform	14-16
14.2.4 Frequency Decimation With a Bit Reversed Input Sequence	14-20
14.3 Multidimensional Discrete Fourier Transformation	14-24
15. APPLICATIONS OF DIGITAL FILTERS	15-1
15.1 TRANSLATION OF A SIGNAL SPECTRUM TO ANOTHER FREQUENCY "MIXING" OR "MODULATION"	15-2

TABLE OF CONTENTS (Continued)

	Page
15.1.1 Extraction of "Complex Modulation" by "Mixing"	15-4
15.2 DESIGN OF A "MOVING WINDOW" FILTER	15-6
15.2.1 Digital Considerations	15-9
15.3 A DIGITAL OSCILLATOR	15-9
15.4 DIGITAL COMPLEX DEMODULATE FM DETECTOR	15-12
 16. "PURE" DIFFERENTIATION AND INTEGRATION	 16-1
16.1 DIFFERENTIATION	16-1
16.1.1 Elementary Differentiation Considerations	16-1
16.1.2 Numerical Considerations for Differentiation	15-4
16.2 INTEGRATION	16-6
 APPENDIX A. CONVERSION OF A POLYNOMIAL TO AN EXPANSION IN TCHEBYCHEF POLYNOMIALS OF THE FIRST AND SECOND KINDS	 A-1
APPENDIX B. USEFUL TRIGONOMETRIC IDENTITIES	B-1
BIBLIOGRAPHY	

LIST OF ILLUSTRATIONS

Figure		Page
2.1.3-1	Subharmonic Sampling of a "High Frequency" Spectrum	2-9
2.1.4-1	Impulse Sampling Function	2-12
2.5.1-1	The "s" Plane	2-28
2.5.1-2	The "z" Plane	2-28
2.6.5-1	Range of Validity of Various Discrete Transforms	2-45
3.1-1	Input-Output Sequence Relation	3-2
3.11-1	Response of a Discrete Fourier Transform	3-37
4.3-1	Two Equivalent Forms of a Linear System for $M = N$	4-7
4.5-1	Simulation Illustrated	4-13
5.2-1	Gibbs Oscillations ($N = 12$)	5-12
5.6.1	The "Gibbs" Phenomena	5-31
5.6.1-1	Simple Rectangles	5-31
5.6.1-2	An Odd Function	5-32
5.6.2-1	Ormsby Replacement of Low-Pass Filters	5-33
5.6.2-2	Illustration of "Fall-Off"	5-34
5.9.1-1	An "Even" Ideal Low-Pass Filter	5-58
5.9.1-2	An "Image" Low-Pass Filter	5-59
5.9.2-1	An "Odd" Low-Pass Filter	5-61
5.10-1	Cosine Filter: Low Pass ($\omega_0 = \pi/12$)	5-63
5.10-2	Cosine Filter: Low Pass ($\omega_0 = \pi/2$) $N = 12$	5-65
5.10-3	Sine Filter: High Pass ($\omega_0 = \pi/2$) $N = 12$	5-67
5.10-4	Sine Filter: Differentiator ($\omega_0 = \pi/2$) $N = 12$	5-69
5.10-5	Sine Filter: Differentiator ($\omega_0 = 3\pi/4$) $N = 12$	5-71
5.10-6	Sine Filter: Differentiator ($\omega_0 = \pi$) $N = 12$	5-73
6.2.2-1	"Modified Impulse Invariant" Design of a Single Pole Bandpass Filter	6-16
6.6-1	Mapping of "s" Plane to "z" Plane	6-30
6.6-2	Frequency Relation Between ω and Ω	6-31
6.6-3	Bilinear z Transform Low-Pass Filter	6-33
6.6.1-1	Bilinear z Transform Bandpass Filter Design via Analog Low-Pass to Digital Bandpass Transformation	6-38
6.6.2-1	Low-Pass Digital Filter to Other Digital Filter Transformation Route	6-40
6.6.2-2	Direct Transformation of a 3 Pole "Normalized" Digital Low-Pass to Low-Pass, High Pass, and Band Stop Digital Filters	6-45
6.8-1	Transfer Function Sampling	6-49
6.8.1-1	Exact Fit of Amplitude Characteristic of the Saturn 1B203 SIB Stage Roll Attitude Rate Network	6-55
7.2.1-1	Zero Order Hold Output	7-5
7.2.1-2	Zero Order Sample and Hold Reconstruction Function	7-6
7.2.1-3	Amplitude (a) and Phase (b) of Zero Order Hold	7-7
7.2.2-1	First Order Sampled and Hold Reconstruction (Delayed)	7-9
7.2.2-2	First Order Sample and Hold (Delayed) Reconstruction Function	7-9

LIST OF ILLUSTRATIONS (Continued)

Figure		Page
7.2.3-1	First Order Sample and Hold (Extrapolated) Output	7-11
7.2.3-2	Amplitude (a) and Phase (b) Characteristics of a First Order Hold (Extrapolation)	7-12
7.2.3-3	First Order Sample and Hold (Extrapolation) Reconstruction Function	7-13
7.4.1-1	Reconstruction Filter	7-19
7.4.2-1	Impulse Sampling Spectrum	7-27
7.5.1	N = 3 Butterworth Low Pass Design via z Transformation with D. C. Level Set, No Reconstruction	7-32
7.5-2	Bandpass Filter Design via z Transformation with D. C. Level Set, No Reconstruction	7-34
7.5-3	Control Function via z Transformation, No Reconstruction	7-36
7.5-4	Control Function, z Transformation with Delayed 1st Order S&H	7-38
7.5-5	Control Function, z Transformation, Zero Order S&H	7-40
7.5-6	Control Function, z Transformation, 1st Order Extrapolation S&H	7-42
7.5-7	Control Function, z Transformation Butterworth Recons.; $\alpha = .1, \omega_a = 1/4 \omega_s$	7-44
7.5-8	Control Function, z Transformation Butterworth Recons.; $\alpha = .01, \omega_a = 1/4 \omega_s$	7-46
7.5-9	Derivative Control Function z Transformation, Compound Reconstruction Filters: Butterworth, $\alpha = .01, \omega_a = 1/3 \omega_s$; plus 1st Order Delayed S&H	7-48
7.5-10	Derivative Control Function, z Transformation, Zero Order S&H	
8.3-1	Implementation of a High Frequency Filter Using Only Low Frequency Digital Narrow-band Filters	8-16
11.1.6-4	Truncation to a Finite Time Process	11-11
12.1-1	Typical Frequency Window ($D(\omega)$) Shape	12-4
12.1-2	Relative Variation of Spectrum Estimates	12-19
12.2-1	Spectrum Measurement by Analog Filtering	12-28
12.3-1	Complex Demodulation Spectrum Estimation	12-36
13.1.1-1	Peaked Distribution	13-5
13.1.11-1	Normalized RMS Error vs. Normalized Prediction Time	13-26
13.1.11-2	Normalized RMS Error as a Function at the Number of Data Samples $\bar{N} + 1$	13-27
13.1.11-3	Normalized RMS Error versus Predictor Order	13-28
13.1.11-4	Frequency Response of Polynomial Estimators	13-29
14.2.1-1	Flow Graph of Time Decimation Method with Naturally Ordered Input Sequence	14-11
14.2.2-1	Flow Graph of Time Decimation Method with Bit Reversed Input Sequence	14-15
14.2.3-1	Flow Graph of Frequency Decimation Method with Naturally Ordered Input Sequence	14-19
14.2.4-1	Flow Graph of Frequency Decimation Method with a Bit Reversed Input Sequence	14-23

LIST OF ILLUSTRATIONS (Continued)

Figure		Page
15.1-1	(a) Spectrum Prior to Mixing	15-2
	(b) Spectrum After Mixing	15-3
15.1-2	Translation of D. C. Spectrum to ω_0	15-3
15.1-3	Mixing Block Diagram	15-4
15.1.1-1	Demodulation by Mixing	15-4
15.2-1	A "Moving" Filter	15-6
15.2.2	Moving Filter Implementation	15-7
15.5-1	Complex Demodulate FM Detector	
16.1.1-1	Linear Function "Turned-On"	16-4
16.1.2-1	Differentiation of a Function Generated by a Filter $H(s)$ by a Parallel Method	16-5
16.1.2-2	Differentiation by the "Black Box" Approach	16-6
16.2-1	Trapezoidal Rule Integration	16-7
16.2-2	Plots of Integration Procedure Comparisons With Ideal	16-10

LIST OF TABLES

Table		Page
3.7.5.3-1	N_{opt} as a Function of M (Assuming N_{opt} a Power of 2)	3-26
5.7.5-1	Low-Pass Filter Ripple Errors	5-53
6.8-1	Number of Simultaneous Equations for Various End Point Conditions	6-51
7.2.1-1	Zero Order Hold Values	7-8
7.2.2-1	First Order Hold (Delayed) Values	7-10
7.2.3-1	Amplitude and Phase Characteristics of a First Order Hold	7-12
7.4.1-1	Minimum Delay and Minimum Pole Reconstruction Butterworth Filter Design Parameters	7-23
12.5-1	Confidence Levels versus Normalized Relative Stability	12-46
16.2-1	Comparison Ratio, $D(\omega)$, for Various Integration Schemes	16-11

I. INTRODUCTION

In the past few decades, the electronic digital computer has allowed the use of many numerical techniques in the areas of analysis, problem solving, and data processing that would previously have been impossible by dint of the sheer volume of numbers.

A class of problems to which the computer has been applied from its earliest versions is the solution of linear systems of integro-differential equations which describe many physical systems. Emphasis previously had been upon analytic solutions to analytic formulations. Among the first uses of the computer was then, the computation of the values of the analytical solutions to analytically formulated input processes; simply a calculator for well known methods. Consideration was next given to solving a given analog system with the input an arbitrary function given only by sampled values, usually equally spaced. Simulation of the continuous system was essentially the aim.

Concurrently, physical system designers were using digital data transmission and reconstruction techniques along with analog and sometimes digital control systems.

As digital computers, both general and special purpose, have become more routinely used by programmers, numerical analysts, and engineers for analysis, simulation, and real-time control and processing, it has been recognized in wider circles that number sequences, as such, can be processed by linear (and non-linear) operations using frequency domain and transform techniques as can be done with analog functions. Transform techniques have been used with sequences by theorists in probability and statistics for quite some time however (see Grenander and Rosenblatt, *Statistical Analysis of Stationary Time Series*, 1967, Wiley). The ideas of filtering, bandwidth, spectral content, modulation and de-modulation, and quadrature processing, etc., that are commonplaces in the analog arena can be equally well applied to sampled sequences.

Not unexpectedly, the mathematical techniques for analyzing data sequences and analyzing and synthesizing discrete systems parallel quite closely those for continuous systems. z transformation with evaluation on the unit circle are the counterparts of the Laplace and Fourier transformations.

The purpose of this volume is to gather the generally applicable theory and techniques involved in discrete, or digital, data filtering and processing which can be treated by transform techniques into a single document for consumption by programmers, numerical analysts and engineers. Most material is specialized to linear, time invariant filters and processes which do however have a surprising range of applicability.

A general description of the contents by sections will probably impart most easily the aims and purposes of this book.

Not illogically the following section, the second, discusses the various transforms of functions found most useful; the Laplace, z , continuous and discrete Fourier transforms and interrelationships among them. Miscellaneous relations involving prospective transform and inverse transform usage are given. Section 3 is devoted to the various aspects of linear, time invariant digital filter definition, properties meaning, usage, and realization. Much of the material uses the z transform form of the input-output relationship which is central to most of the entire text.

An apparent digression is made in section 4 to present basic continuous integro-differential system properties. How these systems act upon a bandlimited function produces a numerical integration scheme with certain provisions. Fundamental problems in simulating a continuous system by a digital one are clarified. The need for this digression becomes more apparent when digital filter synthesis is undertaken in later sections often using analog system prototypes.

Sections 5, 6, and 7 primarily treat synthesis methods for digital filters. In Section 5 non-recursive filters, those without feedback, are discussed extensively. The general relationship between digital filters and one-sided Fourier series, with attendant limitations on synthesis of an arbitrary transfer function, is explored. Two basic transfer synthesis methods are given, least squares fitting and exact fitting at arbitrarily spaced discrete points by polynomial methods. Techniques to reduce the number of weights, or equivalently to hasten the convergence of Fourier series, are given. Actual designs of non-recursive bandpass and differentiating filters are included.

Section 6 is addressed to basic methods of generating recursive filters. Three general classes of techniques are given along with limitations; conversion of a general analog filter to a digital one, conversion of analog bandpass filters to digital bandpass filters with similar characteristics and fitting to an arbitrary transfer function. Parenthetically, general stable low-pass filter to high-pass, band-pass and band-reject filter transformations are given directly in the digital domain. Methods for overcoming some limitations of the basic z transformation method of generating recursive digital filters are given in Section 7. The bandwidth of the approximation region is extended considerably. Data reconstruction methods are found useful in these endeavors.

"Complex" digital filtering is the subject matter of the 8th, 9th and 10th sections. Digital filters with complex coefficients are examined in Section 8, and interpreted in terms of the filtering of "narrow-band" sequences which are shown to be capable of representation in terms of a two component complex data sequence. The ninth section is a counterpart, to some extent, of Section 4. Digital filtering and numerical integration of a narrow-band analog system, or simulation, in terms of the sampled analog signal components in complex form is developed, along with limitations in the digital synthesis problem. Section 10 develops synthesis procedures for complex filters analogous to the methods of Sections 5, 6, and 7 for "real" filters.

Statistical considerations for noise and signals imbedded in noise are focussed upon in Sections 11 through 13.

A brief survey of needed probabilistic concepts is given in Section 11. Spectrum and correlation definitions and relations for deterministic and statistical functions, and the effect of a linear, time invariant filter on a random function or sequence in terms of the input and output averages and spectra for stationary processes are presented. Both continuous and discrete processes are discussed along with relations between the stationary spectra when the discrete process is derived via sampling of a continuous one.

Measurement of the statistical spectrum is discussed in Section 12 by two basic methods, the use of "lagged products" and the use of digital filters. Methods of gauging the estimation accuracy in terms of resolution, statistical stability and data quantity are given. Use of the fast discrete Fourier transform algorithm to minimize computation time for the lagged product method is investigated.

Section 13 is devoted to relatively standard, but important, non-recursive filters for estimating a signal or some linear function of the signal, that has been perturbed by noise in terms of a past, present, or future value. Two basic representations of the signal are assumed. Most attention is given to the representation in which the signal is described by a linear combination of known functions with unknown coefficients. This form of representation has the power of allowing practical signals to be represented arbitrarily well if a sufficient number of functions are allowed in the basis set. The other form of representation assumes only that the signal is known statistically in terms of the mean, auto-correlation function and the cross-correlation with the noise. "Weiner" filters are obtained. Filters derived under either signal representation are optimized with respect to a simple minimum error variance criteria and no requirement is necessarily made as to stationary statistical processes.

A relatively new tool, at least on a popular scale, is the fast discrete Fourier transform algorithm described in Section 14. When the number of data points to be transformed is highly factorable into other integers a substantial reduction in computing time can be obtained in comparison to methods used classically. In particular, four versions of the algorithm are given when the number of data points is a power of 2.

Applications of the work of the preceding sections is contained within Section 15. Most are counterparts of devices and techniques found in work-a-day analog practice.

Finally, the processes of "pure" differentiation and integration are discussed in Section 16. Analyses of a number of common numerical procedures for performing these functions from a transfer function point of view are given.

2. SAMPLED OR "DIGITAL" DATA REPRESENTATION

Before discussing digital filters there should be an understanding of certain aspects of "representing" sampled data sequences or functions. Of primary interest are "transforms" of the data sequences which allow algebraic manipulation of sampled sequences and the filtering operations. These transforms also allow interpretation in terms of "frequency content" of the data and the "frequency response" of the filters, and assist in analyzing the errors introduced by sampling.

The transforms to be discussed subsequently are the Fourier, (analog and discrete) Laplace, and z transforms. Furthermore, the three transforms will be looked upon as being three versions of the same entity.

Certain other relations found useful, such as the "convolution," are presented. Also various parallels between the continuous and discrete Fourier transforms are included in tabular form.

A "sampled time function" is a function of time, $x(t)$, whose values are given only at discrete times t_k . That is, a series of values $x(t_k)$. In general, consider only the case where the times t_k are equally spaced; namely $t_k = kT$ where T is the interval between samples. Whenever unequal sampling is considered it will be explicitly stated. For brevity the time samples $x(kT)$ often will be denoted by

$$x_k = x(kT). \quad (2.1-1)$$

"Times series" or "sequence" means a sequence of values x_k which are obtained in any manner; not necessarily from sampling a time function. A "sampled time function" is a special case of a time series. Time series will be considered at equally spaced times unless otherwise stated.

To understand sampled data filtering a few basic relations about continuous functions or functions with a finite number of finite discontinuities must be summarized.

2.1 FOURIER TRANSFORMATION (ANALOG)

Define the Fourier transform of a real or complex continuous function, $x(t)$, by

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-i\omega t} dt, \quad i = \sqrt{-1} \quad (2.1-2)$$

If $x(t)$ possesses a finite number of finite discontinuities and it is magnitude integrable:

$$\int_{-\infty}^{\infty} |x(t)| dt < \infty \quad (2.1-3)$$

Now $x(t)$ is given in terms of its Fourier transform by

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega') e^{i\omega' t} d\omega' \quad (2.1-2)$$

or

$$x(t) = \frac{\omega_B}{\pi} \frac{1}{2\omega_B} \int_{-\omega_B}^{\omega_B} X(\omega') e^{i\omega' t} d\omega' \quad (2.1.1-3)$$

as $X(\omega) = 0, |\omega| > \omega_B$.

By comparing equation 2.1.1-2 with equation 2.1.1-3 it is found that

$$\alpha_n = \frac{\pi}{\omega_B} x\left(\frac{n\pi}{\omega_B}\right) \quad (2.1.1-4)$$

if $x(t)$ is continuous at $t = n/\omega_B$. The expansion coefficients are determined by data samples taken

$$T = \frac{\pi}{\omega_B} \quad (2.1.1-5)$$

seconds apart. But $T = \frac{2\pi}{\omega_s}$ where ω_s is the sampling frequency. Therefore ω_B must be less than or equal to $\omega_s/2$. Sampling must be at a rate at least twice the bandlimit of the signal. Hereafter it is assumed that $\omega_B = \omega_s/2$, though this is not necessary as long as ω_B is less than $\omega_s/2^*$. Therefore

$$\alpha_n = T x(nT) \quad (2.1.1-6)$$

and

$$X(\omega) = \begin{cases} T \sum_{n=-\infty}^{\infty} x(nT) e^{-in\omega T} & 0 \leq |\omega| \leq \omega_s/2, \\ 0, & \text{otherwise.} \end{cases} \quad (2.1.1-7)$$

*A. Papoulis, Error Analysis in Sampling Theory, Proc. IEEE, Vol. 54, No. 7 July 1966, pp 947-955

Using this form for $X(\omega)$, $x(t)$ can be found in terms of the data samples from the Fourier transform

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{i\omega t} d\omega \quad (2.1.1-8)$$

$$= \sum_{n=-\infty}^{\infty} x(nT) \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} e^{i\omega(t-nT)} d\omega$$

using equation 2.1.1-7.

Performing the integration it is found that

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT) \frac{\sin\left(\frac{\omega_s}{2}(t-nT)\right)}{\frac{\omega_s}{2}(t-nT)} \quad (2.1.1-9)$$

which is a result known as Nyquist's, or Shannon's Theorem among other names.

Examination of equation (2.1.1-9) shows that a bandlimited function $x(t)$ can be reconstructed in its entirety from samples taken $T = 2\pi/\omega_s$ apart where ω_s is at least twice the bandwidth ω_B of the function $x(t)$. Conversely, a bandlimited function can be constructed from a sequence of samples $x(nT)$ using formula (2.1.1-9). This remark will not be proven as it is an easy exercise.

Note that the reconstruction formula (2.1.1-9) requires an infinite delay to be applicable. All the data samples, past and future, must be known.

Unfortunately, most signals encountered in practice are not exactly band-limited. If they were, then the signals extend to $\pm$ infinity in time. To reconstruct signals from samples taken over a finite time span, even though band limited, it is usually necessary to sample a signal many times the rate $\omega_s = 2\pi/T$. In some applications it is necessary to sample hundreds of times that rate. Quite frequently, because of system complexity, analysis alone is not sufficient to tell when sampling is too slow, or too fast. Doubling (or halving) the rate and observing the effects is a practical measure of efficient sampling rates.

When $X(\omega)$ is not bandlimited between $\pm \omega_s/2$ then an effect known as "aliasing" occurs when sampling at $T = 2\pi/\omega_s$. Frequencies greater than $\omega_s/2$ appear as frequencies in the range 0 to $\omega_s/2$ when sampling $x(t)$. To avoid this effect $x(t)$ must be "prefiltered" to bandlimit it before sampling if possible, or raise the sampling frequency. If "aliasing" does occur then the original function cannot be reconstructed exactly. A discussion of "aliasing" follows in section 2.1.2.

In deriving the transform of a bandlimited function (2.1.1-7), in terms of the data samples it was assumed that $x(t)$ was continuous at the sampling points. If $x(t)$ is discontinuous at say $x(nT)$, then the transform must be modified as $x(t)$ is obtained from its transform $X(\omega)$ in deriving (2.1.1-4). At a discontinuity the inverse transform gives the value

$$x(t_0) = \lim_{\epsilon \rightarrow 0} \frac{x(t_0 + \epsilon) + x(t_0 - \epsilon)}{2}, \quad \epsilon > 0, \quad (2.1-5)$$

which, as mentioned previously (Section 2.1), is not necessarily the value of the function $x(t_0)$ prior to transformation.

An important discontinuity, is in a function $x(t)$ such that

$$x(t) = \begin{cases} x(t), & 0 < t < \infty \\ 0, & t < 0, \end{cases} \quad (2.1.1-10)$$

where $x(t)$ is continuous for $t > 0$. By repeating the derivation which led to (2.1.1-7) it is found that the modified transform is given by

$$X_M(\omega) = \begin{cases} T \sum_{n=0}^{\infty} x(nT) e^{-i\omega nT} - T \frac{x(0^+)}{2}, & 0 \leq |\omega| \leq \omega_s/2, \\ 0, & \text{otherwise.} \end{cases} \quad (2.1.1-11)$$

where $x(t^+)$ denotes $\lim_{\epsilon \rightarrow 0} x(t+\epsilon)$, ($\epsilon > 0$).

A contradiction may have been noted at this point; a discontinuity produces a spectrum that is not bandlimited. In practice, if $\omega_s/2$ is taken large enough then $X(\omega)$ will be effectively zero for $|\omega| > \omega_s/2$.

The inverse transform of $X_M(\omega)$ is

$$x(t) = \sum_{n=0}^{\infty} x(nT) \frac{\sin \frac{\omega_s}{2} (t-nT)}{\frac{\omega_s}{2} (t-nT)} - \frac{x(0^+)}{2} \frac{\sin \frac{\omega_s}{2} t}{\frac{\omega_s}{2} t} \quad (2.1.1-12)$$

as it should be.

Equations (2.1.1-11) and (2.1.1-12) above are the modified (because of the discontinuity) transform and function.

2.1.2 "Aliasing" Due to Sampling at a Rate Less Than Twice the "Bandlimit"

In the preceding work it was assumed that $x(t)$ was bandlimited below half the sampling frequency. If the sampling rate is such that the bandlimit is greater than half the sampling frequency then a phenomena known as "aliasing" occurs: frequencies higher than half the sampling rate appear as samples of frequencies in the range of frequencies in the range $0 \leq |\omega| \leq \omega_s/2$ ($\omega_s = 2\pi/T$, the sampling frequency.)

For example, consider the frequencies $\omega_1 = \omega_s - \Delta$ and $\omega_2 = \omega_s + \Delta$ where $0 \leq \Delta < \omega_s/2$. If sine waves are sampled at the frequencies ω_1 and ω_2 then

$$\sin \omega_1 nT = -\sin \Delta nT, \quad (2.1.2-1a)$$

and

$$\sin \omega_2 nT = \sin \Delta nT, \quad (2.1.2-b)$$

which verifies the remark; frequencies above $\omega_s/2$ appear as frequencies below $\omega_s/2$ when sampled.

A relation of great interest is that between the transform defined by equation (2.1.1-7) when the samples are taken from an arbitrary function (not necessarily bandlimited) and the transform of the continuous function. It will be found that unless the function $x(t)$ is bandlimited below half the sampling frequency ($\omega_s/2$) then the spectrum above $\omega_s/2$ gets "mixed in" with that below $\omega_s/2$. The function is "aliased" and cannot be reconstructed without prior knowledge!

If a time series, or sequence, is derived by sampling a Fourier transformable function $x(t)$ then the samples can be represented as

$$x(nT) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{i\omega nT} d\omega \quad (2.1.2-1)$$

By appropriately manipulating this expression a useful relation can be developed. First, write

$$x(nT) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega') e^{i\omega' nT} d\omega'. \quad (2.1.2-2)$$

changing the dummy variable of integration to ω' .

Next multiply both sides by $e^{-i\omega nT}$ and sum over n , interchanging the orders of summation and integration on the right:

$$\sum_{n=-\infty}^{\infty} x(nT) e^{-i\omega nT} = \frac{1}{T} \int_{-\infty}^{\infty} X(\omega') \left[\sum_{n=-\infty}^{\infty} \frac{T}{2\pi} e^{i(\omega' - \omega)nT} \right] d\omega' \quad (2.1.2-3)$$

The bracketed term is the translated periodic expansion of the "function"

$$\delta(\omega) = \lim_{\epsilon \rightarrow 0} \begin{cases} \frac{1}{\epsilon}, & |\omega| < \frac{\epsilon}{2}, \\ 0, & |\omega| > \frac{\epsilon}{2}, \end{cases} \quad (2.1.2-4)$$

with a period of $\omega_s = 2\pi/T$.

The "function" $\delta(\omega)$ possesses, more nearly, or is defined by, the useful property:

$$\int_{-a}^a f(\omega) \delta(\omega) d\omega = f(0), \quad a > 0, \quad (2.1.2-5)$$

It is called the Dirac "delta" function. Expanding it formally in a Fourier series of period $\omega_s = 2\pi/T$, the representation

$$\delta_{\omega_s}(\omega) = \sum_{n=-\infty}^{\infty} \frac{T}{2\pi} e^{i\omega nT} \quad (2.1.2-6)$$

as stated previously, is found. The subscripted notation $\delta_{\omega_s}(\omega)$ denotes the fact

that this representation of $\delta(\omega)$ repeats itself. Using this knowledge of the meaning of the series and the basic property (equation (2.1.2-5)) of $\delta(\omega)$ it is found that equation (2.1.2-3) can be written

$$\sum_{n=-\infty}^{\infty} x(nT) e^{-i\omega nT} = \frac{1}{T} \sum_{k=-\infty}^{\infty} X(\omega - k\omega_s), \quad (2.1.2-7)$$

assuming $x(t)$ and $X(\omega)$ continuous.

Examination of this expression shows that the left hand side will be exactly equal to $X(\omega)/T$ in the region $-\omega_s/2 \leq \omega \leq \omega_s/2$ if $X(\omega)$ is zero outside this region.

This result, appearing to be an abstract relation between two quantities, is in fact quite significant. For example, the transform of a bandlimited function represented in terms of its samples (equation (2.1.1-7)) is given by

$$\mathcal{F}\{x(t)\} = \begin{cases} T \sum_{n=-\infty}^{\infty} x(nT) e^{-in\omega T}, & 0 \leq \omega \leq \omega_s/2 \\ 0, & \text{otherwise} \end{cases} \quad (2.1.2-8a)$$

$$= X_D(\omega)$$

In general; the result shows that

$$\mathcal{F}\{x(t)\} = \sum_{k=-\infty}^{\infty} X(\omega - k\omega_s) \quad (2.1.2-8b)$$

If $x(t)$ is bandlimited by $\pm \omega_s/2$ then the reconstruction formula (2.1.1-9) will produce a correct result. On the other hand, when $x(t)$ is not bandlimited, the reconstruction formula will produce the aliased spectrum given by (2.1.2-8b).

Another form of the "aliased" transform is obtained if one of the sample points occurs at a discontinuity of $x(t)$, say at $t = 0$. Further, it is assumed that $x(t) = 0$ for $t < 0$. Then equation (2.1.2-3) becomes at $t = 0$:

$$\frac{x(0+)}{2} = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega') d\omega', \quad (2.1.2-9)$$

where $x(0+)$ is the limit as t approaches zero from positive t . See equation (2.1-5). Carrying out the rest of the development following equation (2.1.2-2):

$$\sum_{n=0}^{\infty} x(nT) e^{-i\omega nT} = \frac{x(0+)}{2} + \frac{1}{T} \sum_{k=-\infty}^{\infty} X(\omega - k\omega_s). \quad (2.1.2-10)$$

is obtained. A constant $x(0+)/2$ is added again because of the discontinuity.

2.1.3 Subharmonic Sampling

Usually in discussing sampling, the sampled function is assumed bandlimited below half the sampling frequency. An important generalization, however, is already contained within previous developments. Equation (2.1.2-8) asserts that the sampled spectrum may be derived from a continuous waveform at any frequency.

Furthermore, if the original continuous spectrum is localized such that

$$x(\omega) = \begin{cases} X(\omega), & m\omega_s \leq |\omega| \leq (m+1/2)\omega_s, \\ & (\text{or } (m+1/2)\omega_s \leq |\omega| \leq (m+1)\omega_s) \\ 0, & \text{otherwise,} \end{cases} \quad (2.1.3-1)$$

as pictorialized in figure (2.1.3-1a) then the relation

$$X_D(\omega) = \sum_{k=-\infty}^{\infty} X(\omega - k\omega_s) \quad (2.1.2-8)$$

illustrates that

$X_D(\omega) = X(\omega - m\omega_s) + X(\omega + m\omega_s)$, $|\omega| \leq \omega_s/2$, and repeats periodically outside the region defined by $|\omega| \leq \omega_s/2$ as depicted in figure (2.1.3-1b).

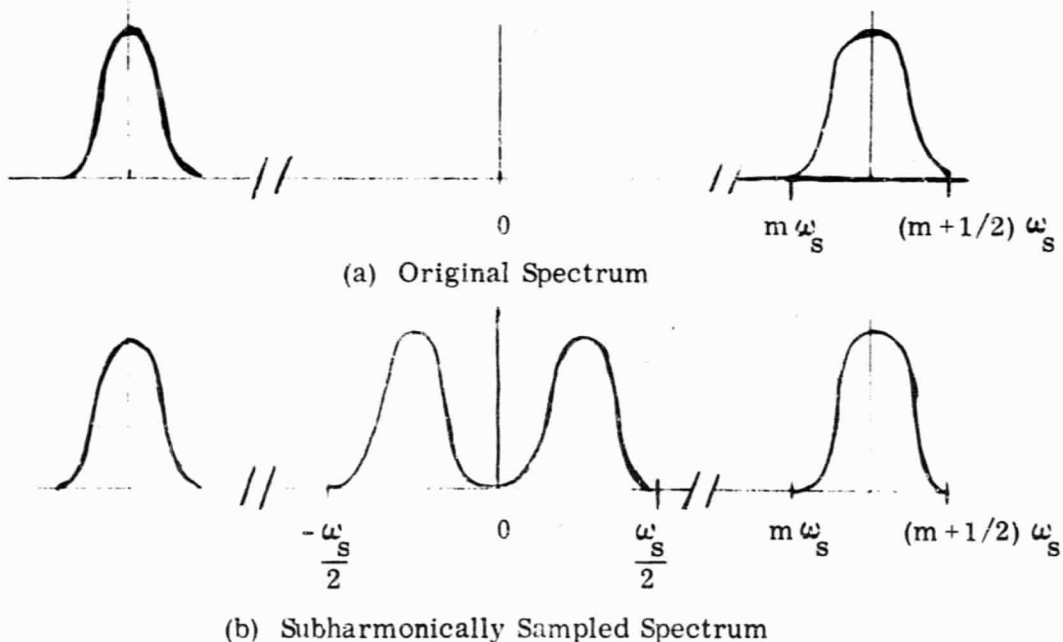


Figure 2.1.3-1 Subharmonic Sampling of a "High Frequency" Spectrum

This sampling process obviously aliases (actually translates) the "high frequency" spectrum down to the region between zero and half the sampling frequency where "demodulation" can be performed at low frequencies digitally. This downward frequency translation by sampling will be called "Subharmonic Sampling."

A constraint that should be considered in practice is the duration of the sampling "aperture." This "aperture" is the time which an analogue signal is observed when converting it to a sampled value. This aperture should be much less than the highest frequency being sampled. That is, if the highest frequency is 1 megahertz then the aperture should certainly be less than .1 microseconds.

If the original signal is localized in the region $(m+1/2)\omega_s \leq |\omega| \leq (m+1)\omega_s$ the high frequency spectrum is still translated to the region $|\omega| \leq \omega_s/2$; however, a peculiarity of sampling displays itself (See section 2.1.2). Frequencies of $(m+1/2)\omega_s + \Delta$ "map" into the frequencies $\omega_s/2 - \Delta$ in the region $|\omega| < \omega_s/2$ with a 180° phase reversal as may be seen upon careful scrutiny of equation (2.1.2-8). Frequencies are, therefore, mapped backwards; those extending from $(m+1/2)\omega_s$ to $(m+1)\omega_s$ become frequencies from $\frac{\omega_s}{2}$ to zero in that order.

An interesting application of "subharmonic sampling" would be the sampling of an intermediate frequency "IF" signal in a radio receiver with a relatively slow sampling frequency, 10-20 kilohertz, and then subsequent digital processing to further demodulate the signal and obtain the information carried.

In everyday life subharmonic sampling is found on movie and TV screens where automobile tires are seen to rotate backwards due to the frame rate.

Finally, it should be pointed out that it is possible in principle to reconstruct the high frequency signal from the "subharmonically sampled" sequence as long as $m\omega_s$ is known. The signal transform in terms of the samples was given by equation (2.1.2-8).

$$\begin{aligned} X_D(\omega) &= T \sum_{n=-\infty}^{\infty} x(nT) e^{-i\omega nT} \\ &= \sum_{k=-\infty}^{\infty} X(\omega - k\omega_s) \end{aligned} \quad (2.1.2-8)$$

Using $X(\omega)$ given by equation (2.1.3-1) it is possible to write, after considerable algebra, using the methods of section 2.1.2:

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT) \cos \left(m\omega_s (t-nT) + \frac{\omega_s}{4} (t-nT) \right) \left(\frac{\sin \frac{\omega_s}{4} (t-nT)}{\frac{\omega_s}{4} (t-nT)} \right) \quad (2.1.3-2)$$

2.1.4 Bandlimited Functions and "Impulse Sampling"

Another form of sampling that is found useful is "impulse sampling." This form of sampling is a mathematical idealization of "pulse sampling" that appears often in physical electronic systems. As interest here centers primarily upon the handling of numbers, or data sequences, it might seem that "impulse sampling" is of little value for the work in this volume. Quite to the contrary, impulse sampling with its Fourier transform (and the subsequent z transform) is a most valuable tool for analysis of sampled data where only numbers, or data sequences, are processed.

"Impulse sampling" is the idealized "modulation" of a periodic stream of narrow pulses by a time function $x(t)$. That is, assume a stream of narrow pulses given by

$$p(t) = \sum_{n=-\infty}^{\infty} v(t-nT) \quad (2.1.4-1)$$

where

$$v(t) = \begin{cases} \frac{1}{\epsilon}, & t < \frac{\epsilon}{2}, (\epsilon > 0) \\ 0, & \text{otherwise,} \end{cases} \quad (2.1.4-2)$$

and T is the period of the pulses. Figure 2.1.4-1 illustrates the shape of the pulse stream.

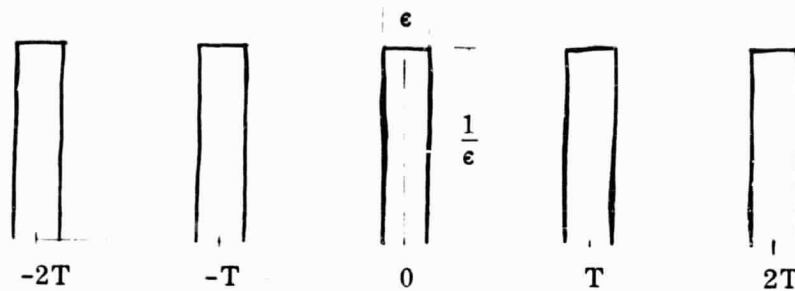


Figure 2.1.4-1 Impulse Sampling Function

"Modulation" of this pulse stream by $x(t)$ is impulse sampling, namely the function

$$x_s(t) = x(t) p(t)$$

(2.1.4-3)

$$= x(t) \sum_{n=-\infty}^{\infty} v(t-nT)$$

is the "impulse sampled" function, if the width of the pulses (ϵ) is small compared to variations of $x(t)$. In the limit $v(t)$, a typical impulse, approaches the delta function (2.1.2-5) discussed previously. While the impulses approach a width of zero and a height of infinity as ϵ approaches zero, which is an unrealistic situation, this approximation is highly useful as impulse sampled functions are invariably integrated in applications and, as will be seen, simplify analysis. In fact, it will be seen that no approximation is involved when discussing number sequences.

Taking the Fourier transform of $x_s(t)$, assuming $v(t)$ has approached the "delta function" $\delta(t)$, it is found that

$$X_s(\omega) = \sum_{n=-\infty}^{\infty} x(nT) e^{-i\omega nT}, \quad (2.1.4-4)$$

where $x(t)$ has been taken to be continuous at nT , which, is remarkably similar to the transform of a bandlimited function (section 2.1.1). A significant difference does exist however! $X_s(\omega)$ is never identically zero (except for $x(t)$ identically zero) for $|\omega| > \omega_s/2$; in fact $X_s(\omega)$ is periodic with period ω_s . However, if $x(t)$ is bandlimited between $\pm \omega_s/2$, then $X_s(\omega)$ will be equal to $X(\omega)/T$ over $-\omega_s/2 < |\omega| < \omega_s/2$ and be its "periodic extension" outside this region. This remark is proved in a manner similar to techniques used when discussing "aliasing."

First, $p(t)$ must be converted to another form; expansion in a Fourier series. Expansion for the limit as $\epsilon \rightarrow 0$ is of chief interest. While this series is a formal one, it does not converge in a strict mathematical sense, it can be used if it is considered as the limiting form of the expansion after the expansion is used in subsequent integration. That is, assume that limiting and subsequent integration are interchangeable in order. Usually, this is a safe assumption in practice.

Expansion of $p(t)$ results in

$$p(t) = \frac{1}{T} \sum_{n=-\infty}^{\infty} e^{-i\omega_s kt}, \quad \omega_s = \frac{2\pi}{T}, \quad (2.1.4-5)$$

allowing

$$x_s(t) = \frac{1}{T} \sum_{k=-\infty}^{\infty} x(t) e^{-i\omega_s kt} \quad (2.1.4-6)$$

to be written.

Again, taking the Fourier transform produces

$$X_s(\omega) = \frac{1}{T} \sum_{k=-\infty}^{\infty} X(\omega - k\omega_s) \quad (2.1.4-7)$$

where $X(\omega)$ is the transform of $x(t)$ which has again been assumed continuous at the sampling points. Examination of this expression shows us that successive terms of the right side do not "overlap" if $X(\omega)$ is bandlimited at $\pm \omega_s/2$. Both equations (2.1.4-4) and (2.1.4-7) demonstrate the periodicity of $X_s(\omega)$ with a period ω_s .

It can be seen, then, that both simple sampling of a function $x(t)$ as in section 2.1.1 and "impulse sampling" produce the same transform within a constant $-\omega_s/2 < \omega < \omega_s/2$. If $x(t)$ is bandlimited by $\omega_s/2$ then both sampling methods have spectrums equal to that of $x(t)$ over $-\omega_s/2 < \omega < \omega_s/2$, however the impulse sampling spectrum continues periodically outside this interval producing higher sampling harmonics. "Low-pass" filtering of an impulse sampled function can reconstruct the function if the transfer function of the filter is approximately

$$H(\omega) = \begin{cases} 1, & |\omega| < \omega_s/2. \\ 0, & \text{otherwise.} \end{cases} \quad (2.1.4-8)$$

This filter is unrealizable but can be approximated if sufficient time delay is allowed.

2.2 LAPLACE TRANSFORMATION

The Fourier transform is often apparently limited by the condition that the time function $x(t)$ involved is not magnitude integrable (eq. 2.1-3). For example, a simple time function that is zero prior to, say, $t = 0$ and constant thereafter is not magnitude integrable. The counterpart of this function in the physical world is the "switching-on" of a battery. Obviously, this is a meaningful function that must be handled if the transform is to have general applicability. Fortunately, the Fourier transform can be "patched" to handle this situation.

The function to be transformed is simply multiplied by $e^{-\gamma t}$ and only situations for $t \geq 0$ are considered. If functions that are non-zero for $t < 0$ are to be considered then that region must be handled by a factor $e^{\gamma_1 t}$. Obviously, considering only functions for $t \geq 0$ is simpler so attention here is confined to this case. The Fourier transform is, then

$$F(\gamma + i\omega) = \int_0^{\infty} x(t) e^{-\gamma t} e^{-i\omega t} d\omega, \quad t \geq 0; \quad (2.2-1)$$

where γ must be such that

$$\int_0^{\infty} |x(t)| e^{-\gamma t} dt < \infty \quad (2.2-2)$$

and $x(t)$ has been assumed equal to zero for $t < 0$. The modification of the Fourier transformation is simply the introduction of a convergence factor to dominate $x(t)$ so that (2.1-3) holds. If a change of variables is introduced

$$s = \gamma + i\omega$$

then equation (2.2-1) is called the Laplace transform ("one-sided") of $x(t)$. If $t < 0$ is of interest the corresponding modified transform is called the "two-sided" Laplace transform which is not discussed in this volume.

Equation (2.2-1) may be inverted to find $x(t)$ by taking the inverse Fourier transform

$$x(t) e^{-\gamma t} = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\gamma + i\omega) e^{i\omega t} d\omega \quad (2.2-3a)$$

or

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\gamma + i\omega) e^{i(\gamma + i\omega)t} d\omega \quad (2.2-3b)$$

Substituting in the integral the change of variable $s = \gamma + i\omega$:

$$x(t) = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} F(s) e^{st} ds \quad (2.2-4)$$

which is the Laplace inversion integral. Further discussion of these relations is not warranted here. Many excellent texts on the subject are available*.

With this introduction of the Laplace transform as a modified Fourier transform, subsequent discussion in this section will assume the Laplace transform as a separate entity.

As was done with the Fourier transform earlier, the Laplace transform and its inverse is stated. The Laplace transform $F(s)$ of a function $x(t)$ is defined by

$$F(s) = \int_0^{\infty} x(t) e^{-st} dt, \quad (x(t) = 0, t < 0); \quad (2.2-5)$$

where s is a complex number. Existence of the Laplace transform is guaranteed if a positive real constant γ exists such that

$$\int_0^{\infty} |x(t)| e^{-\gamma t} dt < \infty \quad (2.2-6)$$

and $x(t)$ has a finite number of finite discontinuities. These conditions are sufficient, but not necessary. Inverse transformation is obtained through the complex integral

$$x(t) = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} F(s) e^{st} ds \quad (2.2-7)$$

where γ is chosen so that $F(s)$ is analytic to the right of the line $u = \gamma$ in the plane ($s = u + i\omega$). Through this expression appears complicated, knowledge of complex variable theory and the fact that many transforms appear as rational functions of s considerably ease use of the inverse transform. In fact, many transform pairs ($x(t)$, $F(s)$) can be found in books of tables; e.g. Bateman Manuscript Project, Integral Transforms, McGraw-Hill.

* A. Papoulis, The Fourier Integral and Its Application, McGraw-Hill, 1962

At a point of discontinuity, t_0 , of $x(t)$ the inverse transform gives the value

$$x(t_0) = \lim_{\epsilon \rightarrow 0} \frac{x(t_0 + \epsilon) + x(t_0 - \epsilon)}{2}, \quad \epsilon > 0, \quad (2.2-8)$$

as does the Fourier transform.

The sampled values of $x(t)$ can be written, then, in terms of their Laplace transform:

$$x(nT) = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} F(s') e^{s'nT} ds'; \quad n = 0, 1, \dots \quad (2.2-9)$$

As before multiply both sides by e^{-snT} and sum over n , and interchange orders of integration and summation on the right side:

$$\begin{aligned} \sum_{n=0}^{\infty} x(nT) e^{-snT} &= \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} F(s') \sum_{n=0}^{\infty} e^{-(s-s')nT} ds' \quad (2.2-10) \\ &= \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} F(s') G(s-s') ds' \end{aligned}$$

where

$$G(s) = \frac{1}{1 - e^{-sT}} \quad (2.2-11)$$

Now $G(s)$ has an infinite number of simple poles along the imaginary axis; namely, where

$$e^{-s_k T} = 1 \quad (2.2-12)$$

or

$$s_k = ik\omega_s, \quad k = 0, \pm 1, \pm 2, \dots; \quad \text{and } \omega_s = 2\pi/T.$$

$G(s)$ is a meromorphic function as it has a finite number of poles within a finite portion of the complex plane allowing expansion into partial fractions by Mittag-Leffler's theorem.

If the complex integral is converted to a contour integral over the right hand plane, starting at $s = \gamma$ which is allowable as both $F(s)$ and $G(s)$ approach zero for $s \rightarrow \infty$, then poles within the contour occur at

$$s - s' = ik\omega_s \quad (2.2-13)$$

As $F(s)$ will have its poles to the left of $u = \gamma$, the integration over the right hand plane will be sum of the residues at the poles s_k . It is found that

$$\sum_{n=0}^{\infty} x(nT) e^{-snT} = \frac{1}{T} \sum_{k=-\infty}^{\infty} F(s - ik\omega_s) \quad (2.2-14)$$

Also, if $x(t)$ has a discontinuity at $t = 0$ the transform relation is modified as before with Fourier transforms:

$$\sum_{n=0}^{\infty} x(nT) e^{-snT} = \frac{x(0^+)}{2} + \frac{1}{T} \sum_{k=-\infty}^{\infty} F(s - ik\omega_s). \quad (2.2-15)$$

The relation (2.2-15) can either be viewed as the Fourier series expansion (with a convergence factor $e^{-\gamma t}$) of $F(s)$ over $-\omega_s/2 \leq \omega \leq \omega_s/2$ or as the Laplace transform of an impulse sampled function. In one case interest centers on the bandlimited spectrum and reconstruction of a function, and in the other case on the spectrum of an impulse sampled function. See sections 2.1.1 through 2.1.4 as the interpretation of the Laplace transform results are identical to those for Fourier transforms. Bandlimiting, with Laplace transforms, means that $F(s)$ is essentially zero outside the strip in the complex s plane bounded by $s = \pm \omega_s/2$ (see section 2.2.1)

2.2.1 Generalization of Bandlimited Function Representation to Functions which are not Fourier Transformable.

In Section 2.1.1 bandlimited functions which are Fourier transformable, namely magnitude integrable, are considered when their Fourier transforms are zero, or approximately so, for frequencies greater than $\omega_s/2$ ($\omega_s = 2\pi/T$). However, when the function under consideration is not magnitude integrable then slight modifications must be made; the Laplace transform must be used.

Taking the Laplace transform of $x(t)$, Section 2.2:

$$F(s) = \int_0^{\infty} x(t) e^{-st} dt, \quad s = \gamma + i\omega \quad (2.2-5)$$

and defining bandlimiting of $x(t)$ by the vanishing of $F(s)$ outside of the strip in the complex plane bounded by $\omega = \pm \omega_s/2$, then $F(s)$ can be expanded in a Fourier series along the line $s = \gamma$

$$F(s) = \begin{cases} \sum_{k=-\infty}^{\infty} a_k e^{-i\omega kT}, & |\omega| < \omega_s/2; \\ 0, & \text{otherwise.} \end{cases} \quad (2.2.1-1)$$

Now the coefficients are given by

$$a_k = \frac{T}{2\pi} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} F(s) e^{i\omega kT} d\omega, \quad s = \gamma + i\omega \quad (2.2.1-2)$$

From the inversion integral, equation (2.2-7), the relation;

$$e^{\gamma kT} a_k = T x(kT) \quad (2.2.1-3)$$

is recognized. Therefore, equation (2.2.1-1), can be written as;

$$F(s) = \begin{cases} T \sum_{k=0}^{\infty} x(kT) e^{-skT}, & |\omega| \leq \omega_s/2; \\ 0, & \text{otherwise} \end{cases} \quad (2.2.1-4)$$

with $s = \gamma + i\omega$.

Taking the inverse transform:

$$\begin{aligned} x(t) &= \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} F(s) e^{st} ds \\ &= \frac{T}{2\pi i} \sum_{k=0}^{\infty} x(kT) \int_{\gamma - i\frac{\omega_s}{2}}^{\gamma + i\frac{\omega_s}{2}} e^{s(t-kT)} ds \end{aligned} \quad (2.2.1-5)$$

which results in the generalized formula for reconstruction of a bandlimited function:

$$x(t) = \sum_{k=0}^{\infty} x(kT) e^{\gamma(t-kT)} \frac{\sin\left(\frac{\omega_s}{2}(t-kT)\right)}{\frac{\omega_s}{2}(t-kT)}, \quad t \geq 0. \quad (2.2.1-6)$$

A striking note is heard here. The reconstruction formula apparently depends on γ !! It is well known that the Laplace inversion formula is independent of γ so long as it is greater than the minimum needed to obtain magnitude integrability.

A simple proof demonstrates that $x(t)$ given by the reconstruction formula does not depend on γ as long as the Laplace transform vanishes also on the lines $s = \pm i\omega_s/2$, namely, vanishes at half the sampling frequency.

By taking the partial derivative of $x(t)$, given by formula (2.2.1-6), with respect to γ , the variation of the reconstruction with γ can be found:

$$\frac{d x(t)}{d \gamma} = \frac{2}{\omega_s} \sum_{k=0}^{\infty} x(kT) e^{\gamma(t-kT)} \sin \left[\frac{\omega_s}{2}(t-kT) \right] \quad (2.2.1-7)$$

$$= \frac{1}{i\omega_s} \left[e^{(\gamma + i\frac{\omega_s}{2})t} \sum_{k=0}^{\infty} x(kT) e^{-(\gamma + i\frac{\omega_s}{2})kT} - e^{(\gamma - i\frac{\omega_s}{2})t} \sum_{k=0}^{\infty} x(kT) e^{-(\gamma - i\frac{\omega_s}{2})kT} \right]$$

The two summations may be seen to be the Laplace transform of $x(t)$, evaluated at $s = \gamma \pm i\frac{\omega_s}{2}$. If the Laplace transform is zero on the lines

$s = \pm i\omega_s/2$, then both summations will be zero and the variation with γ of the reconstructed $x(t)$ will be zero, as stated. The signal must be bandlimited by, but not including $\omega_s/2$ for reconstruction.

The choice of γ in a practical situation is dictated by the variation of $x(t)$ with γ as practical spectrums are not actually bandlimited. As equation (2.2.1-7) can be written

$$\frac{d x(t)}{d \gamma} = \frac{2}{\omega_s} \int_0^{\infty} e^{(\gamma + i \frac{\omega_s}{2} t) F_1 / + \frac{i \omega_s}{2} s} dt \quad (2.2.1-8)$$

it may be seen that γ should be kept to its minimum allowed value, to insure magnitude integrability, as $F(s)$ usually falls off inversely with γ (or a power of γ), but $e^{\gamma t}$ is exponential in γ .

2.3 THE z TRANSFORM

The rudiments of another useful data representation, the z transform, which is merely the previous Fourier bandlimited sequence transformation in disguise ($z = e^{i \omega T}$) is developed. In fact, subsequent developments will show that it is probably the most important sample data representation.

Given a sequence of real or complex numbers, x_n , a "time series," or a sampled time function, $x(nT)$; the "z transform," $\mathcal{Z}(x_n)$ or $X(z)$, of the sequence is defined as

$$\begin{aligned} \mathcal{Z}(x_n) &= X(z), \\ &= \sum_{n=0}^{\infty} x_n z^{-n} \end{aligned} \quad (2.3-1)$$

where it is assumed that the series converges except possibly at distinct points called "poles," and z is a complex number. It is also assumed that $x_n = 0$ when $n < 0$ for convenience. For many familiar sequences the z transform can be expressed in closed form.

For example, consider the important time function e^{-at} , where "a" may be complex, sampled at intervals $t = nT$; $n = 0, 1, \dots$. The sequence in this case is

$$x_n = e^{-anT} \quad (2.3-2)$$

and the transform

$$\begin{aligned} \mathcal{Z}(x_n) &= \sum_{n=0}^{\infty} e^{-anT} z^{-n}, \\ &= \sum_{n=0}^{\infty} (e^{-aT} z^{-1})^n \end{aligned} \quad (2.3-3)$$

Recognizing this sum as an infinite geometric series

$$\mathcal{Z}(x_n) = \frac{1}{1 - e^{-aT} z^{-1}} \quad (2.3-4)$$

which is a closed form.

It will be seen in later sections that the sampled impulse response of a lumped - constant, linear, continuous network can also be expressed in a closed form, as it is composed of additive terms of the general form:

$$t^m e^{-at}$$

Some useful and basic theorems on z transforms follow:

a. Linearity.

$$\mathcal{Z}(ax_n + by_n) = \sum_{n=0}^{\infty} (ax_n + by_n) z^{-n} \quad (2.3-5)$$

$$= a \sum_{n=0}^{\infty} x_n z^{-n} + b \sum_{n=0}^{\infty} y_n z^{-n}$$

$$= a \mathcal{Z}(x_n) + b \mathcal{Z}(y_n)$$

$$= a X(z) + b Y(z)$$

b. Time Translation

$$\mathcal{Z}(x_{n-k}) = \sum_{n=0}^{\infty} x_{n-k} z^{-n}$$

$$= \sum_{n=-k}^{\infty} x_n z^{-(k+n)} \quad (2.3-6)$$

$$= z^{-k} \sum_{n=0}^{\infty} x_n z^{-n} \text{ as } x_n = 0, n < 0$$

$$= z^{-k} \mathcal{Z}(x_n)$$

$$= z^{-k} X(z)$$

A delay of a sequence by k units corresponds to multiplication of its transform by z^{-k} . Similarly, the z transform of x_{n+k} is given by

$$\mathcal{Z}(x_{n+k}) = z^k X(z) \quad (2.3-7)$$

if we assume in this case that $x_n = 0$ for $0 < n < k$. An advance of k units corresponds to multiplication of its transform by z^k .

c. Scale Change in the z Domain.

Multiply the time series members x_n by another, particular, time series $e^{\pm \alpha n}$, then the transform of the product time series is

$$\begin{aligned} \mathcal{Z}(e^{\pm \alpha n} x_n) &= \sum_{n=0}^{\infty} x_n e^{\pm \alpha n} z^{-n} \\ &= X(e^{\pm \alpha} z) \end{aligned} \quad (2.3-8)$$

The argument of the transform of $X(z)$ has been modified to $e^{\pm \alpha} z$.

2.3.1 Initial and Final Value Theorems

Two theorems of much use in the analysis of sampled data systems are the initial and final theorems. They express the initial and limiting values of a time sequence in terms of operations on the z transform of the sequence.

The initial value theorem is quite simple:

$$x_0 = \lim_{z \rightarrow \infty} X(z) \quad (2.3.1-1)$$

Proof of this remark easily follows from the definition of the transform

$$X(z) = x_0 z^0 + x_1 z^{-1} + x_2 z^{-2} + \dots \quad (2.3.1-2)$$

The final value theorem is a trifle more complicated:

$$x_{\infty} = \lim_{z \rightarrow 1} (1 - z^{-1}) X(z) \quad (2.3.1-3)$$

Proof of this remark may be seen readily by examination of the obvious identity.

$$x_N = \left[\sum_{n=0}^N x_n z^{-n} - z^{-1} \sum_{n=0}^{N-1} x_n z^{-n} \right]_{z=1} \quad (2.3.1-4)$$

This theorem is extensively used in theoretical analyses of the steady state performance of sampled data feedback systems.

2.4 THE INVERSE z TRANSFORM

A number of approaches to determining the time series which generated the transform have been developed. Only the simplest methods will be stated. Its relationship to the Laplace transform will not be explored here. (See Section 2.5).

2.4.1 Direct Inverse Transformation

The z transform was defined (equation 2.3-1) for a time series

$$\mathcal{Z}(x_n) = X(z) = \sum_{n=0}^{\infty} x_n z^{-n} \quad (2.3-1)$$

where z is a complex variable and the series is assumed to converge except at distinct points called poles. This can be looked upon as a Taylor-Laurent series. If X(z) is given then multiply both sides by z^{n-1} and integrate over a contour which surrounds all poles of X(z) and the origin and

$$x_n = \frac{1}{2\pi i} \oint X(z) z^{n-1} dz; \quad n = 0, 1, \dots \quad (2.4.1-1)$$

is obtained.

Subsequently it will be found that if $X(z)$ is a z transform transfer function of a stable linear analog or digital filter, then its poles appear in the interior of the unit circle or on it. See Section 3.4. Therefore, for a stable linear filter integration about the unit circle will do. For this case, of all poles of $X(z)$ within or on the unit circle, upon setting $z = e^{i\theta}$, it is found that

$$x_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{i\theta}) e^{in\theta} d\theta \quad (2.4.1-2)$$

which are simply the Fourier coefficients of the expansion of $X(e^{i\theta})$. The inverse transform is also a linear operation.

2.4.2 Inversion by Long Division

Another technique for inverting a z transform is by synthetic division. That is, given a z transform $F(z)$ in a form frequently found in practice

$$F(z) = \frac{a_0 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_M z^{-M}}{1 + b_1 z^{-1} + b_2 z^{-2} + \dots + b_N z^{-N}} \quad (2.4.2-1)$$

and $F(z)$ is expressible as a power series in negative powers of z plus a constant then divide the numerator by the denominator and obtain a term of the form $f_n z^{-n}$ and a remainder. Successive divisions of the remainder by the denominator give terms for larger n .

An example best describes the technique. The z transform of the sequence e^{-anT} is

$$F(z) = \frac{1}{1 - e^{-aT} z^{-1}} \quad (2.4.2-2)$$

The division is set up as for ordinary numbers and carried out exactly as done with them:

$$\begin{array}{r} 1 - e^{-aT} z^{-1} \overline{) 1} \\ \underline{1} \phantom{+ e^{-aT} z^{-1} + e^{-2aT} z^{-2} + \dots} \\ 0 \phantom{+ e^{-aT} z^{-1} + e^{-2aT} z^{-2} + \dots} \\ \underline{0 + e^{-aT} z^{-1}} \phantom{+ e^{-2aT} z^{-2} + \dots} \\ 0 - e^{-aT} z^{-1} \phantom{+ e^{-2aT} z^{-2} + \dots} \\ \underline{0 - e^{-aT} z^{-1} + e^{-2aT} z^{-2}} \\ 0 + e^{-2aT} z^{-2} \phantom{+ e^{-3aT} z^{-3} + \dots} \\ \underline{0 + e^{-2aT} z^{-2} - e^{-3aT} z^{-3}} \\ 0 - e^{-3aT} z^{-3} \\ \underline{0 - e^{-3aT} z^{-3} + e^{-4aT} z^{-4}} \\ \vdots \end{array} \quad (2.4.2-3)$$

etc.

Obviously, the original sequence has been obtained as the coefficients by successive divisions.

More complicated numerators and denominators are handled in the same manner; but computations can become complicated. Mechanization of the procedure utilizing a digital computer is recommended. Care must be taken for large n as successive round-offs can effect sequence values considerably.

2.4.3 Inverse z Transform as an "Impulse Response"

In Section 3.4 it is found that the inverse z transform of a transfer function $H_z(z)$ in terms of its series expansion: namely,

$$H_z(z) = \sum_{n=0}^{\infty} w_n z^{-n} \quad (3.4-2)$$

can be found by applying the sequence

$$x_n = \begin{cases} 1, & n = 0 \\ 0, & \text{otherwise} \end{cases}$$

to the difference equation which is represented by $H_z(z)$ (Section 3.4).

If the function $F(z)$ is considered as a transfer function then the inverse transform sequence can be found by using a difference equation. Of course, $F(z)$ must be a rational function to use this procedure though it need not have poles only within the unit circle. An example is the best means of describing the procedure. Again, take the sequence e^{-anT} whose transform is

$$\frac{1}{1 - e^{-aT} z^{-1}}$$

The equivalent difference equation is (Section 3.2).

$$y_n = e^{-aT} y_{n-1} + x_n \quad (2.4.3-1)$$

Letting $y_n = 0, n \leq 0$, gives

$$y_1 = 1,$$

$$y_2 = e^{-aT}, \quad (2.4.3-2)$$

$$y_3 = e^{-2aT},$$

etc.

For complicated functions the procedure should be mechanized on a digital computer.

A bit of thought will show that this procedure is identical to the "Long Division" method as far as computations are concerned.

2.5 GEOMETRICAL INTERPRETATION OF CONNECTION AMONG FOURIER, LAPLACE, AND z TRANSFORMS.

Each transform introduced can be looked upon as an entity unto itself; that is, the individual transforms may be defined independently and conditions for existence and formulas for inversion investigated. This would be the "pure" mathematical approach of history though it is to be doubted if a modern mathematician would claim or would want to claim lack of knowledge of the interrelationships of the various transforms. While the z transform will usually be the preferred approach in subsequent analysis, the fact that it can be looked upon as a simple substitution ($z = e^{sT}$) of variables in the Laplace (or Fourier) transforms of bandlimited functions, generalizes its utility infinitely beyond the simple transformation of sequences.

The transforms used to this point have shown that a sampled continuous function can be represented by its samples if it is "bandlimited." Conversely, a sequence represents a unique "bandlimited" function (section 2.1.1).

A sequence can, of course, represent an infinity of "aliased" functions where these functions have frequency content above $\omega_s = 2\pi/T$. An aliased function which gives no unique knowledge is almost useless unless a prior knowledge tells how the particular function has been, or will be, aliased. For example, knowledge that a function has a basic frequency of ω_0 much higher than the sampling frequency but has a spectrum (transform) whose width about ω_0 is much less than half the sampling frequency still allows unique reconstruction. (See Section 2.1.3).

When a sequence is considered as representing a unique bandlimited function then numerical operations (rather than analog filtering) can be considered as affecting the frequency content of the bandlimited function and, therefore, the resultant sequence obtained. Useful insight into the computing process and its effect on simple number sequences is obtained through the corresponding bandlimited function. For example, the simple squaring ($y_k = x_k^2$) of a number sequence produces a spectrum of the corresponding squared bandlimited function up to twice that of the initial function. Immediately it is seen that simple squaring requires a sampling rate twice that of the sequence before squaring.

The sequence prior to squaring can either be generated with half the sampling interval (more frequent sampling) or reconstruction of values between samples is performed so that the squared sequence is not "aliased;" namely, can define infinitely many functions.

While the z transform was defined as a somewhat abstract transformation of a time series, it is obviously connected to the Laplace transform and to the Fourier transform.

For functions that are zero for $t < 0$ the z transform can be looked upon as simply a bandlimited Laplace transform where $z = e^{sT}$. The Fourier transform is related to both by letting $s = i\omega$ ($z = e^{i\omega T}$) and assuming $x(t) = 0$ for $t < 0$ if the transform exists. Notice that the sampled Laplace transform involves only the variable e^{sT} ; not simply s . Therefore, the effect of sampling can be looked upon as a mapping from the s domain to another, the z domain, via the transformation $z = e^{sT}$. The line $s = i\omega$ and its map in the z plane informs us about the effect of sampling upon the various frequencies.

2.5.1 The "s" To "z" Plane Mapping Induced by Sampling

The mapping of the s plane into the z plane which is induced by sampling, is of fundamental importance for proper understanding of analog and digital filter relations (along with the "bilinear" z transformation to be introduced later).

It will be found again that simple sequences do not have the same frequency content over $0 \leq \omega \leq \omega_s/2$ as the signal from which they are derived, in general.

To understand these remarks consider the mapping in detail. The sampling operation produces Laplace transforms and, as shall be seen, transfer functions which are functions of the complex number $z = e^{sT}$ rather than simply s . The s plane is depicted in Figure 2.5.1-1 with $s = \gamma + i\omega$ and the z plane with $z = u + iv$ in Figure 2.5.1-2. When

$$s = \gamma + i\omega \quad (2.5.1-1)$$

then

$$z = e^{\gamma T} e^{i\omega T} \quad (2.5.1-2)$$

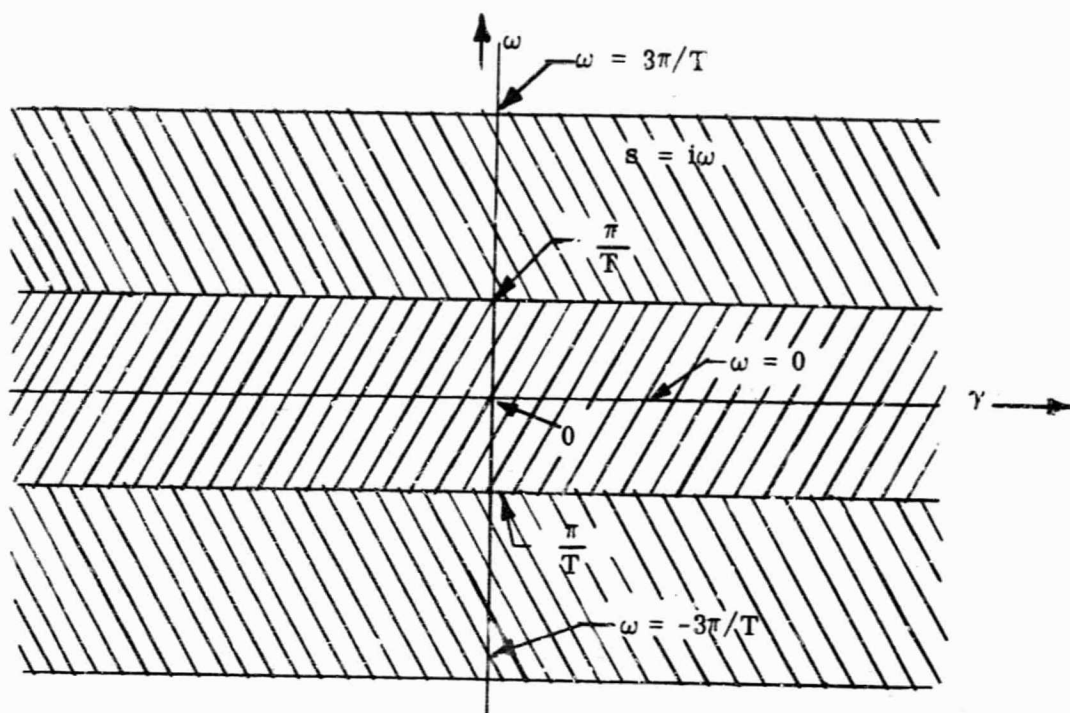


Figure 2.5.1-1. The " s " Plane

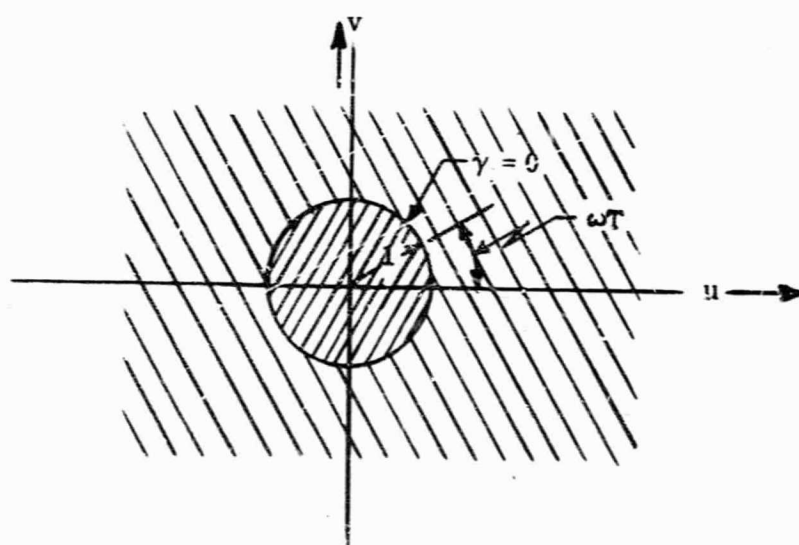


Figure 2.5.1-2. The " z " Plane

It is immediately seen that for $\gamma T < 0$, the left half of the s plane maps into the interior of the unit circle as $e^{-|\gamma T|} < 1$ and $e^{i\omega T}$ describes a unit circle as ω travels from 0 to $2\pi/T$. The right half of the s plane is mapped outside of the unit circle as $e^{\gamma T}$ is greater than unity for $\gamma > 0$.

It is also seen that the mapping $z = e^{sT}$ is not unique, that is, many points in the s plane are mapped into the same point in the z plane. Consider the points $s = \gamma + i\omega_n$; where $\omega_n = \omega_0 + 2\pi n/T$, $-\frac{\pi}{T} < \omega_0 < \frac{\pi}{T}$, $n = 0, \pm 1, \pm 2, \dots$

It is found that

$$\begin{aligned} z_n &= e^{\gamma T} e^{i\omega_n T} \\ &= e^{\gamma T} e^{i\omega_0 T} \\ &= z_0 \end{aligned} \quad (2.5.1-3)$$

All s plane points for fixed γ and values of ω differing by a multiple of $2\pi/T$ are mapped into the same point in the z plane. In other words strips of width $2\pi/T$ in ω in the s plane are mapped into each other in the z plane. For the particular line $\gamma = 0$, the ω axis, it is seen that it is "rolled-up" on the unit circle in the z plane, modulo $2\pi/T$.

This many-to-one mapping has a destructive effect on the function that is sampled. Many different functions when sampled produce the same sequence of samples. For example, the function $\sin \omega t$, $0 \leq \omega \leq \frac{\pi}{T}$, produces the same sequence of sampled values as $\sin (\frac{2\pi}{T} - \omega) t$. In general, frequencies contained in a continuous function higher than $1/2$ the sampling frequency ($1/T$) are mapped into angular frequencies between 0 and π/T . Adding a function that is periodic and zero at the sampling points obviously allows many functions to have the same z transform. Only the values of the sampling points are of interest.

When the function that is sampled is bandlimited (occupies only the strip $-\frac{\pi}{T} < \omega < \frac{\pi}{T}$ in the s plane) then the mapping from the s plane to the z plane may be considered as unique.

2.6 THE DISCRETE FOURIER TRANSFORM

Another transform of considerable utility is the discrete Fourier transformation. Discrete Fourier transformation (DFT) is a transformation of a number sequence, or time series, which is a close analogue of the continuous Fourier transform. It has been in use over the last half century, at least. Presently it has become quite prominent due to a number of reasons; the increased interest in sampled data and spectrum analysis which has been allowed and prompted by modern high speed digital computers, and the popularization of an algorithm for computing the transform with significantly less operations, the so-called "fast Fourier transform."

The fast Fourier transform does not alter the mathematics of the discrete Fourier transformation, it is simply a computation method with greater speed and accuracy than methods used by most people until the recent past. This increased speed and accuracy brings a host of mathematical operations into the realm of practical tools for everyday data processing, primarily spectrum analysis, auto-correlation and cross-correlation calculations, and convolution computations. Historical development* of the fast Fourier transform algorithm is highly interesting as it was present in its embryonic form in the works of very prominent mathematicians and not widely used for over a half century.

This section will be devoted to the aspects of the basic technique, the discrete Fourier transform. Section 14 will be devoted to the fast Fourier transform mechanization.

It should be emphasized that the discrete Fourier transfer with the fast Fourier transform algorithm does not provide any panacea to the problems of digital filtering. Before it can be used properly and above all, appropriately, the subject of digital filtering must be understood.

* J. Cooley, P.A.W. Lewis, and P.D. Welch, Historical Notes on the Fast Fourier Transform, IEEE Trans. on Audio and Acoustics, Vol. AU-15, No. 2, June 1967.

2.6.1 Definition of the Discrete Fourier Transform

The definition of the discrete Fourier transform poses a number of problems. It can be defined in numerous ways, each definition having its particular virtues, though they are mathematically equivalent.

Technical literature of today is dominated by a group primarily interested in spectrum estimation in which case the most convenient definition for the programming and computing of spectra is used; this definition produces some difficulties in conceptual utility when a spectrum, or transform, is given and the time waveform is desired. This definition apparently removes the problem encountered with another definition when an even or odd number of data points is used, however care must still be taken when using an even or an odd number of points in sampling various frequency waveforms. Two forms of the transform will be presented.

First Form (DFT1)

The defining relation for this transform version is given by:

DFT 1:

$$X_n = \sum_{j=0}^{N-1} x_j W^{-jn}, \quad W = e^{i \frac{2\pi}{N}}, \quad i = \sqrt{-1} \quad (2.6.1-1)$$

where:

DFT1 = First version of discrete Fourier transform

x_j = Sequence of values (real or complex) in the "time" domain

j = "Time" index

X_n = Sequence of values (real or complex) in the "frequency" domain

n = "Frequency" index

N = Number of data points

This transform is a "mapping" operation where a sequence of time values is converted to a set of frequency values. As in the case of the continuous Fourier transform the discrete Fourier transform also possesses an inverse; given the frequency value sequence, X_n , it is possible to recover the time sequence, x_j . The inverse (IDFT1) is given by

$$\text{IDFT1:} \quad x_j = \frac{1}{N} \sum_{n=0}^{N-1} X_n W^{jn} \quad (2.6.1-2)$$

As in the case of the continuous transform the discrete Fourier transform and its inverse are identical except for the sign of the exponent of W . Some writers use different values for the constant $1/N$ (with a corresponding change in the DFT).

It is a simple matter to show that the linear DFT1 does indeed have the IDFT1 given by (2.6.1-2). Insertion of X_n given by (2.6.1-1) into (2.6.1-2) gives

$$x_j \stackrel{?}{=} \frac{1}{N} \sum_{k=0}^{N-1} x_k \sum_{n=0}^{N-1} e^{i(j-k)n \frac{2\pi}{N}} \quad (2.6.1-3)$$

The right hand summation can be summed in closed form when it is recognized as a geometric series:

$$\frac{1}{N} \sum_{n=0}^{N-1} e^{i(j-k)n \frac{2\pi}{N}} = \frac{1}{N} \frac{1 - e^{i 2\pi (j-k)}}{1 - e^{i \frac{2\pi}{N} (j-k)}} \quad (2.6.1-4)$$

$$= \begin{cases} 1, & j = k + m N, \quad m = 0, \pm 1, \pm 2, \dots \\ 0, & \text{otherwise} \end{cases}$$

$$= \delta_N(j-k),$$

where $\delta_N(j-k)$ is the Kronecker delta function, modulo N . This equation is often called the "orthogonality" relationship for the discrete functions $W^{\pm jn}$. It follows that

$$x_j = x_j$$

as was to be shown.

Examination of the discrete Fourier transform will show that it is periodic in frequency. That is, as the running index n is increased (or decreased) outside of the range $0 \leq n \leq N-1$.

It is found that

$$X_n = X_{N+n} = X_{2N+n}, \text{ etc.} \quad (2.6.1-5)$$

and also

$$X_{-n} = X_{N-n} = X_{2N-n}, \text{ etc.}$$

Furthermore, even though the original time sequence may be taken from a non-periodic sequence, once it is used to compute the frequency sequence then the inverse transform (IDFT1) produces a sequence such that

$$x_j = x_{N+j} = x_{2N+j}, \text{ etc.} \quad (2.6.1-6)$$

or

$$x_{-j} = x_{N-j} = x_{2N-j}, \text{ etc.}$$

Transformation followed by inverse transformation produces a time sequence which is periodic outside of the original range of definition, $0 \leq j \leq N-1$. In many applications this periodicity is undesirable and must be carefully taken into account. Computation of correlations and convolutions are processes which must be adjusted (Section).

Second Form of the Discrete Fourier Transform (DFT2)

Another form of the discrete Fourier transform commonly used is more closely analogous to the continuous Fourier transform in that positive and negative frequencies are used. It is also nearly symmetrical about time, $t = 0$, and frequency, $\omega = 0$. Parenthetically, the data sequence is also translated in time, but this may be taken into account by a simple translation of the origin.

The transform pair is given by

DFT2:

$$X_n = \sum_{j=-N'_0}^{N_0} x_j w^{-jn}, \quad w = e^{i \frac{2\pi}{N}} \quad (2.6.1-7)$$

and the inverse transform.

IDFT2:

$$x_j = \frac{1}{N} \sum_{n=-N'_0}^{N_0} X_n w^{jn} \quad (2.6.1-8)$$

where:

DFT2 = Second version of discrete Fourier transform

x_j = Sequence of values (real or complex) in the "time" domain

j = "Time" index

X_n = Sequence of values (real or complex) in the "frequency" domain

n = "Frequency" index

N = number of data points

N_0 is defined by: $N = 2N_0$, even number of data points,
 $2N_0 + 1$, odd number of data points.

$$N'_0 = \begin{cases} N_0 - 1, & N \text{ even,} \\ N_0, & N \text{ odd} \end{cases}$$

The validity of the IDFT2 is proved in a manner identical to that for the first version of the transform.

Periodicity of X_n and x_j are obtained also in an identical fashion. However, the running indices in this version are such that

$$-N'_0 \leq \begin{pmatrix} j \\ n \end{pmatrix} \leq N_0$$

completely symmetrical about zero in the odd N case and almost so for the even N case.

2.6.2 Interpretation of the "Time" and "Frequency" Indices "j" and "n".

As might be expected the seemingly abstract discrete Fourier transform with its "time" and "frequency" indices actually is related to sampled time and frequency with which people are normally familiar.

Consider the continuous Fourier transform

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t} dt \quad (2.6.2-1)$$

where the computation of $X(\omega)$ is to be performed with paper and pencil or a computer. Then, the limits of integration must be finite and $x(t)$ and $e^{-i\omega t}$ must be evaluated at discrete time. If a simple rectangular approximation is used over $0 \leq t \leq (N-1)T$ with $x(t)e^{-i\omega t}$ sampled at intervals $t = jT$, and N samples are taken, then an approximation to $X(\omega)$ is

$$X^A(\omega) = T \sum_{j=0}^{N-1} x(jT)e^{-i\omega jT} \quad (2.6.2-2)$$

If the particular values of frequency at which $X^A(\omega)$ is evaluated are $\omega = n\Delta\omega$, and $\Delta\omega$ is chosen as an appropriate fraction of the sampling frequency so that successive values of n , $0 \leq n \leq N-1$, allow evaluation of $X^A(\omega)$ throughout the useful range, namely, $\Delta\omega = \omega_s/N = 2\pi/NT$, then

$$\begin{aligned} X^A(n\Delta\omega) &= T \sum_{j=0}^{N-1} x(jT)e^{-in\Delta\omega jT}, \\ &= T \sum_{j=0}^{N-1} x(jT)e^{-\frac{i2\pi}{N} jn} \end{aligned} \quad (2.6.2-3)$$

Comparison with the DFT1 shows that this approximation is identical within a constant factor T when

$$X_n = X^A(n\Delta\omega) \quad (2.6.2-4)$$

and

$$x_j = x(jT)$$

are equated.

Therefore the indices of the discrete Fourier transform are easily interpreted: j corresponds to the multiple of the sampling time T and n corresponds to a fraction of the sampling frequency: $\frac{n}{N} \omega_s$.

Of interest in practical computing situations is the fact that odd N will not allow sampling in the frequency domain evenly from 0 to the point $\frac{\omega_s}{2}$, half the sampling frequency. The nearest point to half the sampling frequency for odd N , occurs at $\frac{N_0}{N} \omega_s$, where $N = 2N_0 + 1$. Even N , however, allows sampling at $\omega_s/2$.

A more exact description of the relation between the continuous function and the continuous transform, and the discrete samples and the discrete transform are given in Section 2.6.5.

2.6.3 Relationships Between Various Forms of the Discrete Fourier Transform.

As will be seen the various versions of the discrete Fourier transform do not produce identical output sequences. In many applications this is unimportant as long as the same discrete Fourier transform and inverse transform pair are used. However, it often occurs that a particular version is most easily applicable to a given problem statement or data processing function. When the opposite version has been implemented it is useful to be able to convert between versions.

Two methods are available in converting from one form to another. The first method takes the input sequence and transforms it by the method available. Phase shifting must be performed on the output sequence to put it in the form of the other version. Also if the indices are constrained within the original defining range then the periodicity of the output sequence must be used to reorder it into the other discrete Fourier transform form. The second method requires a re-ordering of the input sequence, but the phase modification of the output sequence is not necessary, though reordering is also necessary if the output indices are constrained to the original defining range.

It should become apparent from subsequent pages that "simple" interchange of the two versions of the transform can only be performed knowledgeably.

Phase Modification to Use Various Forms.

One method of using the DFT1 to compute the DFT2 is to carry the computation through using the DFT1 and then modify the resulting frequency sequence by a complex phase factor. If the procedure for computing the DFT1 does not allow indices outside of the range $0 \leq n \leq N - 1$ then the periodicity relationship must be used to rearrange the DFT1 output sequence.

Identical considerations obtain for the reverse situation. Also, the IDFT's can be handled in the same manner to produce identical time sequences from a given frequency sequence with a reversed sign in the phase factor.

Given the sequence x_j ; $j = 0, 1, \dots, N-1$, the DFT1 is given by the definition

$$X_n^{(1)} = \sum_{j=0}^{N-1} x_j W^{-jn}, \quad n = 0, 1, \dots, N-1, \quad (2.6.1-1)$$

and the DFT2 by (assuming the same sequence is transformed)

$$X_n^{(2)} = \sum_{j=-N'_0}^{N_0} x_{j+N'_0} W^{-jn}, \quad n = -N'_0, -N'_0 + 1, \dots, N_0. \quad (2.6.1-7)$$

Shifting indices on the DFT2

$$X_n^{(2)} = e^{i \frac{2\pi}{N} n N'_0} \sum_{j=0}^{N-1} x_j W^{-jn}, \quad (2.6.3-1)$$

and multiplying this equation by $\frac{1}{N} W^{nj}$ and summing over n from $n = 0, \dots, N-1$ (summing over this range is allowable as $X_n^{(2)}$ is defined by the given x_j^s):

$$x_j = \frac{1}{N} \sum_{n=0}^{N-1} X_n^{(2)} e^{-i \frac{2\pi}{N} n N'_0} W^{jn} \quad (2.6.3-2)$$

Comparing this result with the IDFT1 it is found that the two frequency transforms are related by

$$X_n^{(1)} = e^{-i \frac{2\pi}{N} n N'_0} X_n^{(2)} \quad (2.6.3-3a)$$

As both transforms are determined from the given x_j the range of n in this relation is unimportant as $X_n^{(1)}$, $X_n^{(2)}$, and $e^{-i\frac{2\pi}{N}nN'_0}$ are periodic in N . If the DFT1 is used then

$$X_n^{(2)} = e^{i\frac{2\pi}{N}nN'_0} X_n^{(1)} \quad (2.6.3-3b)$$

is used.

When the procedure for computing, say, the DFT1 is used with indices restricted to $0 \leq n \leq N-1$ then the periodicity relationship

$$X_{-n}^{(1)} = X_{N-n}^{(1)} \quad (2.6.1-5)$$

can be used to compute $X_{-n}^{(2)}$. Similarly, when $X_n^{(1)}$ is to be computed using the DFT2 when the procedure restricts $-N'_0 \leq n \leq N'_0$, then periodicity may again be invoked and

$$X_{N-n}^{(2)} = X_{-n}^{(2)} \quad (2.6.1-5)$$

Conversely, if the two frequency transforms are defined as being identical then the IDFT1 produces:

$$x_j^{(1)} = \frac{1}{N} \sum_{n=0}^{N-1} X_n w^{jn}, \quad j = 0, 1, \dots, N-1. \quad (2.6.1-2)$$

and the IDFT2:

$$x_j^{(2)} = \frac{1}{N} \sum_{n=-N'_0}^{N'_0} X_{n+N'_0} w^{jn}, \quad j = -N'_0, \dots, 0, \dots, N'_0 \quad (2.6.1-8)$$

By reasoning identical to that for the frequency transforms the time transforms are related by

$$x_j^{(1)} = e^{-i\frac{2\pi}{N}jN'_0} x_j^{(2)} \quad (2.6.3-4)$$

If procedures are available for calculating only one IDFT then this relation may be used. Input-Output Rearrangement to Use Various Forms of the Discrete Fourier Transform.

Obviously, by virtue of equation (2.6.3-4) it is not possible simply to use the different IDFT's with the same frequency "input" albeit translated, and obtain the same time sequence with the appropriate index range, without using the phase factor W^{nN} . However, a simple dodge allows use of the IDFT1 when the IDFT2 would be more appropriate, if input and output sequence rearrangement is allowable.

Suppose that the frequency sequence for input to the IDFT1 is defined in terms of the input and the IDFT2 by

$$X_n^{(1)} = X_n^{(2)}; n = 0, 1, \dots, N_0 \quad (2.6.3-5)$$

and

$$X_n^{(1)} = X_{-N + n}^{(2)}; n = N_0 + 1, \dots, N,$$

where

$$N = 2N_0, \text{ even}$$

$$N = 2N_0 + 1, \text{ odd.}$$

Then, the IDFT1 gives

$$x_j = \frac{1}{N} \sum_{n=0}^{N_0} X_n^{(2)} W^{jn} + \frac{1}{N} \sum_{n=N_0+1}^N X_{-N+n}^{(2)} W^{jn}; j=0, 1, \dots, N-1 \quad (2.6.3-6)$$

Translating the indices in the second summation by $-N$, and using the fact that $W^{nN} = 1$:

$$x_j = \frac{1}{N} \sum_{n=-N_0}^{N_0} X_n^{(2)} W^{jn}, \quad j = 0, 1, \dots, N-1, \quad (2.6.3-7)$$

which appears to be the same as the IDFT2 which was desired. But the index range is different, which is not bothersome if the program implemented for the IDFT1 allows negative indices. If negative indices are not allowed then use of the periodicity relationship

$$x_{-j} = x_{N-j} \quad j = 1, 2, \dots, N'_0$$

allows rearranging of the IDFT1 output into the same form as would be obtained by the IDFT2.

Conversely, if the IDFT2 is available, and the output of the IDFT1 is desired, given the frequency input, then the equivalence needed is

$$X_n^{(2)} = X_n^{(1)} \quad n = 0, 1, \dots, N'_0, \quad (2.6.3-8)$$

$$X_n^{(2)} = X_{n+N}^{(1)} \quad n = -N'_0, \dots, -1$$

Taking the IDFT2:

$$x_j = \frac{1}{N} \sum_{n=0}^{N'_0} X_n^{(1)} w^{jn} + \frac{1}{N} \sum_{n=-N'_0}^{-1} X_{n+N}^{(1)} w^{jn}, \quad j = -N'_0, \dots, 0, \dots, N'_0. \quad (2.6.3-9)$$

Again, adjust indices, and use the fact that $w^{nN} = 1$:

$$x_j = \frac{1}{N} \sum_{n=0}^N X_n^{(1)} w^{jn}, \quad j = -N'_0, \dots, 0, \dots, N'_0. \quad (2.6.3-10)$$

The IDFT2 produces the IDFT1 result upon juggling the input and also requires that the program for the IDFT2 allow indices outside of its range of original definition if x_j for $N'_0 + 1 \leq j \leq N-1$ are desired. Use of the periodicity relationship here also allows rearranging if the program will not allow this extended index range:

$$x_{N-j} = x_{-j}$$

Obviously, if the time waveforms are given and identical frequency outputs from the two versions of the transform are desired then the previous technique will work by simply interchanging the roles of the time and frequency sequences.

2.6.4 Effect on the Transform When an Output Sequence is Real.

When choosing among versions of the transform it is often true that the transform must be real. This places a certain constraint on the input sequence. A pertinent example is the synthesis of a non-recursive digital filter weighting sequence (Section) where the weighting sequence is derived by taking the inverse discrete Fourier transform of the desired (often analog) transfer function that possesses the property $H(-\omega) = H^*(\omega)$. It is most natural to use the IDFT2 as negative frequencies are allowed, however the IDFT1 may be used as the results of Section 2.6.3 demonstrate.

First, suppose that $x_j^{(1)}$ is to be real under the IDFT1, then

$$x_j^{(1)} = x_j^{(1)*}$$

$$x_j^{(1)*} = \frac{1}{N} \sum_{n=0}^{N-1} X_n^{(1)*} W^{-jn}, \quad j = 0, 1, \dots, N-1;$$

and also

$$x_j^{(1)} = \frac{1}{N} \sum_{n=0}^{N-1} X_n^{(1)} W^{jn};$$

where $X_n^{(1)}$ is the frequency transform under the DFT1. Using the orthogonality relationship (2.6.1-4) within the allowed range of indices, it is found that

$$X_n^{(1)*} = X_{N-n}^{(1)}, \quad n = 0, 1, \dots, N-1. \quad (2.6.4-2)$$

Therefore, $X_0^{(1)} = X_N^{(1)}$ must be real (periodicity), and if N is even, $N = 2N_0$, then $X_{N_0}^{(1)}$ also must be real. (Notice that DFT1/IDFT1 does not eliminate consideration of the evenness or oddness of N .)

Next, consider the reality of $x_j^{(2)}$ under the IDFT2, then

$$x_j^{(2)} = x_j^{(2)*}$$

and

$$x_j^{(2)*} = \frac{1}{N} \sum_{n=-N'_0}^{N_0} X_n^{(2)*} W^{-jn}; \quad j = -N'_0, \dots, 0, \dots, N_0; \quad (2.6.4-3)$$

$$x_j^{(2)} = \frac{1}{N} \sum_{n=-N'_0}^{N_0} X_n^{(2)} W^{-jn};$$

where $X_n^{(2)}$ is the frequency transform under the DFT2, implying

$$X_n^{(2)*} = X_{-n}^{(2)}; \quad n = -N'_0, \dots, 0, \dots, N_0; \quad (2.6.4-4)$$

using the orthogonality relationship. As was true for the IDFT1, it can be seen that $X_0^{(2)}$ must be real and that $X_{-N_0}^{(2)} = X_{N_0}^{(2)}$ must be real if $N = 2N_0$ is even (periodicity).

2.6.5 Relation Between Continuous and Discrete Fourier Transforms

As the discrete Fourier transform is the means by which computation of continuous transforms is usually performed the relation between the two should be carefully analyzed. Also, the discrete Fourier transform is quite often used to process data which was derived from an analog signal which also makes its relation to the continuous process pertinent.

It is relatively easy to show, with the machinery developed in the preceding pages, that the "aliased" time and frequency waveforms are discrete Fourier transforms of each other, "a DFT pair." By appropriate use of these relations it is possible to have separate control of frequency aliasing via the time sampling interval and the frequency resolution by the data span.

From section 2.1.1 the spectrum of a bandlimited waveform was given as

$$X_D(\omega) = \sum_{k=-\infty}^{\infty} X(\omega - k\omega_s) \quad (2.1.2-8b)$$

or also as

$$X_D(\omega) = T \sum_{K=-\infty}^{\infty} x(kT) e^{-j\omega kT} \quad (2.1.2-8a)$$

Examination of this equation will reveal that $X_D(\omega)$ is periodic with period $\omega_s = 2\pi/T$.

By taking equi-spaced samples in the range $0 \leq \omega < \omega_s$ and using the IDFT! an interesting relation may be found.

Let

$$\begin{aligned} \omega &= \omega_n \\ &= n \frac{\omega_s}{N} \\ &= \frac{n}{N} \frac{2\pi}{T} \end{aligned} \quad (2.6.5-1)$$

and take the IDFT1:

$$\begin{aligned}
 \frac{1}{N} \sum_{n=0}^{N-1} X_D \left(n \frac{\omega_s}{N} \right) W^{nj} &= T \sum_{k=-\infty}^{\infty} x(kT) \frac{1}{N} \sum_{n=0}^{N-1} W^{n(j-k)}, \\
 &= T \sum_{k=-\infty}^{\infty} x(kT) \delta_N(k-j) \quad (2.6.5-2) \\
 &= T \sum_{M=-\infty}^{\infty} x(jT - mNT),
 \end{aligned}$$

where $\delta_N(j)$ is the Kronecker delta, modulo N . Defining

$$\begin{aligned}
 x_D(t) &= \sum_{m=-\infty}^{\infty} x(t - m\bar{T}), \quad (2.6.5-3) \\
 \bar{T} &= NT,
 \end{aligned}$$

then it can be seen that $X_D(\omega)$ and $T x_D(t)$ are a discrete Fourier transform pair. It may be noted that $x_D(t)$, the "aliased" time function is periodic with a period $\bar{T} = NT$.

This DFT relation between the aliased frequency and time functions sheds considerable light on the computation of transforms digitally. But first, a few bookkeeping items should be spelled out.

Not only are $X_D(\omega)$ and $T x_D(t)$ DFT1 pairs but also, as can be proven in a manner identical to the one above, they are DFT2 pairs by virtue of their periodicity.

Procedurally, the actual sampling times and frequencies do not appear explicitly in the various versions of the DFT, only implicitly in the indices j and n . A time sequence is discretely transformed and the resultant frequency spectrum is labelled by the integers 0 through $N-1$. To know the frequencies with respect to the actual time scale it is only necessary to replace the integer labelling, n , by $n\omega_s/N$.

Returning to a discussion of the significance of the DFT between the aliased time and frequency waveforms certain relations are seen immediately.

This discussion assumes that the frequency transform is being computed from the time samples. A parallel discussion would allow the reverse computation. Suppose that the sampling interval, T , is chosen so that the aliasing of $X_D(\omega)$ is small. That is

$X(\omega) \approx 0$ for $|\omega| > \omega_s/2$. Then, as $X_D(\omega)$ is periodic with period ω_s

$$X_D(\omega) \approx X(\omega), \quad |\omega| \leq \frac{\omega_s}{2}. \quad (2.6.5-4)$$

Unfortunately, the DFT1 evaluates $X_D(\omega)$ for the range $0 \leq \omega \leq \omega_s$ which can be compensated by noticing that

$$X_D(\omega + \omega_s) \approx X(\omega), \quad -\frac{\omega_s}{2} \leq \omega \leq 0, \quad (2.6.5-5)$$

or

$$X_D^{(\omega)} \approx X(\omega - \omega_s), \quad \frac{\omega_s}{2} \leq \omega \leq \omega_s$$

Therefore, the region $\frac{\omega_s}{2} \leq \omega \leq \omega_s$, when using the DFT1,

approximates the analog spectrum in the region $-\omega_s/2 \leq \omega \leq 0$. If the DFT2 is used the problem does not appear as this version of the transform evaluates for the proper region: $|\omega| < \omega_s/2$.

Figure 2.6.5-1 illustrates the relationships.

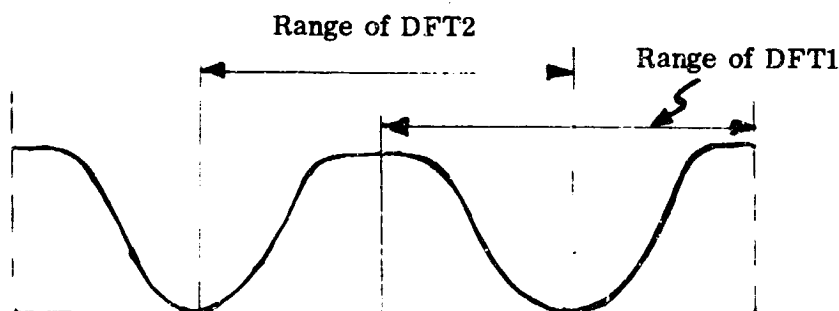


Figure 2.6.5-1. Range of Validity of Various Discrete Transforms

Assuming that T has been chosen small, so that the aliasing frequency harmonics $X(\omega - k\omega_s)$, $k \neq 0$, produce a tolerable error, then the remaining parameter to choose is N which determines the sampling increment between successive frequencies. This sampling increment is again

$$\begin{aligned}\Delta\omega &= \frac{\omega_s}{N} \\ &= \frac{1}{N} \frac{2\pi}{T}\end{aligned}\tag{2.6.5-6}$$

When using the fast Fourier transform it will be seen that N should be a highly composite integer and, with present binary techniques, preferably a power of two.

To compute the frequency transform with the desired spacing the quantity

$$x_D(jT) = \sum_{m=-\infty}^{\infty} x(jT - mNT)\tag{2.6.5-7}$$

must be formed and then the discrete Fourier transform must be taken. Notice that N need only be chosen to give the frequency resolution desired, the amount of time aliasing is unimportant.

If it should happen that $x(jT)$ is zero outside of the range $0 \leq j \leq N-1$ then $x_D(jT)$ is simply given by

$$x_D(jT) = x(jT), \quad 0 \leq j \leq N-1,\tag{2.6.5-8}$$

and the frequency and time sequences are direct discrete transforms of each other. However, it should be observed that frequency-time transforms in the continuous domain (and discrete domain also, see section 3.10) possess the property that localize time waveforms have Fourier transforms that have infinite extent. Conversely, localized frequency transforms have time waveforms infinite in extent. Fortunately, most frequency spectra found in practice can be assumed to vanish at a sufficiently high frequency for finite time waveforms.

If frequency resolution is desired only to a certain accuracy, and N is significantly larger than the minimum amount needed, then the time aliasing should be performed first to save on computations in the taking of the transform.

One assumption made to this point has been that $x(t)$ actually extended to plus and minus infinity (though probably small at large distances from the origin). In practice $x(t)$ is usually truncated in some finite interval. This implies that the actual spectrum is convolved with the frequency function:

$$\frac{T'}{2\pi} \frac{\sin(\omega T'/2)}{(\omega T'/2)} \quad (2.6.5-9)$$

where T' is the truncation interval (Section 11.1.10). This function allows a resolution in frequency of approximately $\Delta\omega = 2\pi/T'$ with, however, large leakage "sidelobes" either side of the main peak at $\omega = 0$. The time weighting functions of section 5.6 can be used to reduce these "leakage" lobes at the expense of about a factor of two in resolution.

2.6.6 Fourier Series and the Discrete Fourier Transform

Suppose a periodic function $x(t)$ has a Fourier series expansion

$$x(t) = \sum_{m=-\infty}^{\infty} c_m e^{i 2\pi m \frac{t}{T}} \quad \bar{T} = \text{period}; \quad (2.6.6-1)$$

with coefficients given in the least squares sense by

$$c_m = \frac{1}{\bar{T}} \int_0^{\bar{T}} x(t) e^{-i 2\pi m \frac{t}{\bar{T}}} dt \quad (2.6.6-2)$$

If the discrete Fourier transform (DFT) is taken at equispaced intervals:

$$t = 0, T, \dots, (N-1) T; T = \frac{\bar{T}}{N} \quad (2.6.6-3)$$

then

$$C_n = \sum_{m=-\infty}^{\infty} c_m \sum_{j=0}^{N-1} W^{j(m-n)} \quad (2.6.6-4)$$

$$= N \sum_{m=-\infty}^{\infty} c_m \delta_{N(m-n)}$$

$$= N \sum_{r=-\infty}^{\infty} c_{n-rN}$$

where C_n is the DFT1 transform. Therefore, the discrete Fourier transform

$$C_n = \sum_{j=0}^{N-1} x(jT) W^{-jn} \quad (2.6.6-5)$$

is related to the continuous Fourier series expansion coefficients by

$$C_n = N \sum_{r=-\infty}^{\infty} c_{n-rN} \quad (2.6.6-6)$$

To effectively compute the c_n by use of the DFT1, N should be chosen so that

$$C_n \approx N c_n \quad (2.6.6-7)$$

with an allowable error. Then $T = \bar{T}/N$ is used to sample $x(t)$ and the discrete Fourier transform taken.

As with the Fourier integral transform use of the DFT1 maps the region below $n = 0$ of the c_n into the region above $N/2$ for the C_n . That is, if N is chosen sufficiently large, then

$$C_n \approx c_n, \quad n = 0, 1, \dots, N/2,$$

and

(2.6.6-8)

$$C_{N-n} \approx c_{-n}, \quad n = 1, 2, \dots, N/2.$$

2.6.7 Use of the Discrete Fourier Transform for the Evaluation of Fourier Series

As Fourier series occur quite frequently in practice it is quite useful to be able to evaluate them by the use of the discrete Fourier transform as the fast algorithm (Section 14) may be applied. However the discrete transform only evaluates at certain distinct frequencies. By a slight manipulation the Fourier series may be evaluated at any set of equi-spaced time values.

The form (2.6.1-7), the DFT2 will be used here though with trivial modification the DFT1 (2.6.1-1) may be used.

By definition a continuous Fourier series is (truncated)

$$f(t) = \sum_{k=-N'_0}^{N_0} c_k e^{-i\omega_k t} \quad (2.6.7-1)$$

with N_0 and N'_0 defined by

$$N = \begin{cases} 2N_0 + 1, & \text{odd number of terms,} \\ 2N_0, & \text{even number of terms;} \end{cases}$$

and

$$N_0 + N'_0 + 1 = N.$$

The other quantities are the given c_k and ω_k the harmonics of the basic frequency $\omega_1 = 2\pi/\bar{T}$ with $\bar{T}$ the expansion interval:

$$\omega_k = k \frac{2\pi}{\bar{T}}; \quad -N'_0 \leq k \leq N_0 \quad (2.6.7-2)$$

To evaluate $f(t)$ at N equi-spaced points over the interval $0 \leq t < T$ set

$$t = j \frac{\bar{T}}{N}; \quad -N_0' \leq j \leq N_0, \quad (2.6.7-3)$$

with j an integer; then

$$f\left(j \frac{\bar{T}}{N}\right) = \sum_{k=-N_0'}^{N_0} c_k e^{-i \frac{2\pi}{N} jk} \quad (2.6.7-4)$$

which is just the discrete Fourier transform of Section 2.6.1 (DFT2). To evaluate at other equi-spaced time points between these N points use the following artifice. N points equi-spaced, but falling between the previously computed N points, may be defined by

$$t = (j + \tau) \frac{\bar{T}}{N}; \quad \begin{array}{l} -N_0' \leq j \leq N_0, \\ |\tau| < 1. \end{array} \quad (2.6.7-5)$$

Insertion of these values into (2.6.7-1) gives

$$f\left((j + \tau) \frac{\bar{T}}{N}\right) = \sum_{k=-N_0'}^{N_0} \left[c_k e^{-i \frac{2\pi}{N} \tau k} \right] e^{-i \frac{2\pi}{N} jk} \quad (2.6.7-6)$$

which is merely the discrete Fourier transform (DFT2) of the phase shifted sequence c_k .

As τ varies between plus and minus one a new set of values offset from the original ($\tau = 0$) are formed. For example, if $2N$ values equi-spaced from 0 to $\bar{T}$ are desired then the original set of N values for $\tau = 0$ should be computed and then another set of N values for $\tau = \frac{1}{2}$.

2.7 THE CONVOLUTION THEOREM: CONTINUOUS AND DISCRETE WAVEFORMS

A relation of considerable practical interest is the "convolution" theorem. This theorem allows the description of a linear filter in terms of the "impulse response" and the input waveform which is quite useful in practice. Both continuous and discrete systems have a particular version of the relation (See section 3.4.1 for an application of discrete convolution) and, as might be expected, the discrete version can be used to approximate the continuous one.

The two versions of the theorem will be derived here.

2.7.1 Continuous Convolution Theorem

Given the time functions $x(t)$ and $w(t)$, with Fourier transforms $X(\omega)$ and $H(\omega)$ respectively, the inverse Fourier transform of the product of the transforms is given by

$$y(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\omega) X(\omega) e^{i\omega t} d\omega \quad (2.7.1-1a)$$

$$= \int_{-\infty}^{\infty} x(\tau) w(t-\tau) d\tau$$

and also

$$y(t) = \int_{-\infty}^{\infty} w(\tau) x(t-\tau) d\tau \quad (2.7.1-1b)$$

Proof:

The defining relation of the inverse transform is (2.1-4). Inserting the definitions of $H(\omega)$ and $X(\omega)$ into this relation:

$$y(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} dt' \int_{-\infty}^{\infty} dt'' w(t') x(t'') e^{i\omega(t-t'-t'')} \quad (2.7.1-2)$$

Inverting the order of integration, and using the fact that

$$\delta(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega t} d\omega \quad (2.7.1-3)$$

is the Dirac delta function:

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} dt' \int_{-\infty}^{\infty} dt'' w(t') x(t'') \delta(t-t'-t'') \\ &= \int_{-\infty}^{\infty} w(\tau) x(t-\tau) d\tau. \end{aligned} \quad (2.7.1-4)$$

or reversing orders of integration,

$$y(t) = \int_{-\infty}^{\infty} x(\tau) w(t-\tau) d\tau,$$

where the dummy integration variables, t' or t'' , have been replaced by τ .

2.7.2 Discrete Convolution Theorem

Given the time sequences x_j and w_j with discrete Fourier transforms (DFT1) X_n and H_n , respectively, the inverse discrete Fourier transform (IDFT1) y_j , of the product sequence of the transforms is given by

$$y_j = \frac{1}{N} \sum_{n=0}^{N-1} H_n X_n W^{jn}, \quad j = 0, 1, \dots, N-1, \quad (2.7.2-1a)$$

$$= \sum_{k=0}^{N-1} x_k w_{j-k}$$

and also

$$y_j = \sum_{k=0}^{N-1} w_k x_{j-k} \quad (2.7.2-1b)$$

where the indices of the sequences x_j and w_j are considered circularly, periodically, or modulo N ; (all these terms are equivalent).

Proof:

The defining relation of the inverse discrete transform is (2.6.1-2). Inserting the definitions of H_n and X_n into this relation:

$$y_j = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} \sum_{m=0}^{N-1} x_k w_m W^{(j-k-m)n} \quad (2.7.2-2)$$

Reversing the order of summation, and using the orthogonality of the functions W^{jn}/N ,

$$y_j = \sum_{k=0}^{N-1} \sum_{m=0}^{N-1} x_k w_m \delta_N(j-k-m) \quad (2.7.2-3)$$

where $\delta_N(k)$ is the Kronecker delta, modulo N . Therefore, if x_k and w_k are considered as repeating periodically outside of the region $0 \leq k \leq N-1$, which is caused by the inverse transformation relation, then

$$y_j = \sum_{k=0}^{N-1} x_k w_{j-k}$$

or

(2.7.2-4)

$$y_j = \sum_{k=0}^{N-1} w_k x_{j-k}$$

It should be noted that circular usage of x_k and w_k is not generally desirable in approximating the Fourier integral, or in some cases where the quantity

$$y_j = \sum_{n=0}^{N-1} x_n w_{j-n}$$

is desired with neither x_k or w_k circular. This can be circumvented by appropriately defining the sequences x_k and w_k with enough zeroes appended in the range $0 \leq k \leq N-1$.

2.8 TABULAR COMPARISON OF VARIOUS CONTINUOUS AND DISCRETE FOURIER TRANSFORM RELATIONS

It is highly informative to write the "equivalent" continuous and "discrete" Fourier transform relations in a comparative, tabular form. Though care must be exercised in replacing one form of a relationship with another the correspondences are valid if the appropriate constraints are used. The work of this section is devoted to a large extent, to the connection between the two sets of transforms, though often knowledgeable use requires an understanding of much of this book.

Some of these relationships, which are relatively obvious, are not proven elsewhere.

Continuous Transform Relations	Discrete Transform Relations
<u>Definition</u>	
$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t} dt$	$X_n = \sum_{j=0}^{N-1} x_j W^{-jn},$
	where
	$W = e^{-i \frac{2\pi}{N}}$
<u>Linearity</u>	
$\int_{-\infty}^{\infty} [a y(t) + b x(t)] e^{-i\omega t} dt$ $= aY(\omega) + bX(\omega)$	$\sum_{j=0}^{N-1} [a y_j + b x_j] W^{-jn}$ $= a Y_n + b X_n$
<u>Inverse</u>	
$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{i\omega t} d\omega$	$x_j = \frac{1}{N} \sum_{n=0}^{N-1} X_n W^{jn}$
<u>Orthogonality of Basis Functions</u>	
$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{\pm i\omega(t-t')} d\omega = \delta(t-t')$	$\frac{1}{N} \sum_{n=0}^{N-1} W^{\pm n(k-k')} = \delta_N(k-k')$

Where $\delta_N(k-k')$ is the Kronecker delta considered modulo N

Convolution Theorem Time Domain

$$\begin{aligned} & \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\omega) X(\omega) e^{i\omega t} d\omega \\ &= \int_{-\infty}^{\infty} w(\tau) x(t-\tau) d\tau \\ &\text{or} \\ &= \int_{-\infty}^{\infty} x(\tau) w(t-\tau) d\tau \end{aligned}$$

where

$$\begin{aligned} H(\omega) &= \int_{-\infty}^{\infty} w(t) e^{-i\omega t} dt \\ X(\omega) &= \int_{-\infty}^{\infty} x(t) e^{-i\omega t} dt \end{aligned}$$

Correlation Functions

$$\begin{aligned} & \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) Y^*(\omega) e^{i\omega t} d\omega \\ &= \int_{-\infty}^{\infty} x(t+\tau) y^*(\tau) d\tau \\ &\text{or} \\ &= \int_{-\infty}^{\infty} x(\tau) y^*(-(t-\tau)) d\tau \end{aligned}$$

$$\begin{aligned} & \frac{1}{N} \sum_{n=0}^{N-1} H_n X_n w^{jn} \\ &= \sum_{n=0}^{N-1} x_k w_{j-k} \\ &\text{or} \\ &= \sum_{k=0}^{N-1} w_k x_{j-k} \end{aligned}$$

where

$$\begin{aligned} H_n &= \sum_{j=0}^{N-1} w_j w^{-jn} \\ X_n &= \sum_{j=0}^{N-1} x_j w^{-jn} \end{aligned}$$

and the indices on w_j and x_j are considered modulo N

$$\begin{aligned} & \frac{1}{N} \sum_{n=0}^{N-1} X_n Y_n^* w^{jn} \\ &= \sum_{k=0}^{N-1} x_{j+k} y_k^* \\ &\text{or} \\ &= \sum_{k=0}^{N-1} x_k y_{j-k}^* \end{aligned}$$

and the indices on x_j and y_j are considered modulo N .

where

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-i\omega t} dt$$

$$Y(\omega) = \int_{-\infty}^{\infty} y(t) e^{-i\omega t} dt$$

Convolution Theorem (Frequency Domain)

$$\begin{aligned} & \int_{-\infty}^{\infty} x(t) y(t) e^{-i\omega t} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega') Y(\omega - \omega') d\omega' \end{aligned}$$

or

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega - \omega') Y(\omega') d\omega'$$

where

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-i\omega t} dt$$

$$Y(\omega) = \int_{-\infty}^{\infty} y(t) e^{-i\omega t} dt$$

where

$$X_n = \sum_{j=0}^{N-1} x_j W^{-jn}$$

$$Y_n = \sum_{j=0}^{N-1} y_j W^{-jn}$$

$$\begin{aligned} & \sum_{k=0}^{N-1} x_k y_k W^{-nk} \\ &= \frac{1}{N} \sum_{n'=0}^{N-1} X_{n'} Y_{n-n'} \end{aligned}$$

or

$$= \frac{1}{N} \sum_{n'=0}^{N-1} X_{n-n'} Y_{n'}$$

where

$$X_n = \sum_{j=0}^{N-1} x_j W^{-jn}$$

$$Y_n = \sum_{j=0}^{N-1} y_j W^{-jn}$$

and the indices on X_n and Y_n are considered modulo N .

3. DIGITAL FILTER CONCEPTS

This section introduces the concept of a digital filter and develops the general properties of these filters and how they behave in terms of "frequency response" and "impulse response". Methods of implementing these filters are also discussed.

Specialization for a particular, important kind of digital filter, the "non-recursive" filter, is introduced. Also mentioned is the relation between "transient" response and "bandwidth" of a digital filter.

While the work of this section does not preclude digital filters that consist of complex numbers and operate on complex data the general discussion of "complex" filters and their interpretation is deferred until section 8.

Finally only certain forms of digital filters are found allowable on stability and physical grounds.

3.1 DEFINITION OF A LINEAR TIME INVARIANT, SAMPLED DATA OR "DIGITAL" FILTER

A linear, time-invariant sampled data, or "digital", filter is defined by a simple formula, namely,

$$b_0 y_j = - \sum_{n=1}^N b_n y_{j-n} + \sum_{m=0}^M a_m x_{j-m} \quad (3.1-1)$$

where:

x_n = n^{th} member of the input sequence, real or complex,

y_n = n^{th} member of the output sequence, real or complex,

and a_m , and b_n , are constant coefficients, or "weights", that are moved across the sequence when filtering. A "recursive" filter means a filter which involves past outputs, which implies two or more non-zero b 's, as well as past inputs. If all the b_n are zero but b_0 the recursive filter reduces to an important special case known as a "non-recursive" filter. Equations of the form of 3.1-1 are also known as linear, constant coefficient difference equations and have properties, solutions, and solution techniques analogous to linear, constant coefficient differential equations. Their properties were studied extensively by George Boole in his "Calculus of Finite Differences" during the 19th century. While his treatise was relatively complete, it was the advent of complex transforms, such as the Fourier-Laplace transform and its offspring, the z transform, that has simplified the study of difference equations and unified them into a common basis with differential equations. Transforms also assist in the description of difference equations as being defined by their frequency response.

In the subsequent presentation a mixture of conventional time domain analysis and transform analysis will be used wherever they seem most appropriate.

A pictorial description of how the "digital" filter is applied to sampled data, independent of its description, synthesis, and properties, which are the main subjects of this manuscript, aids greatly in understanding future discussion. Two horizontal lines are drawn, along which is placed the sequence members at their time of occurrence. Positions on the two lines that occur at the same point occur at the same times.

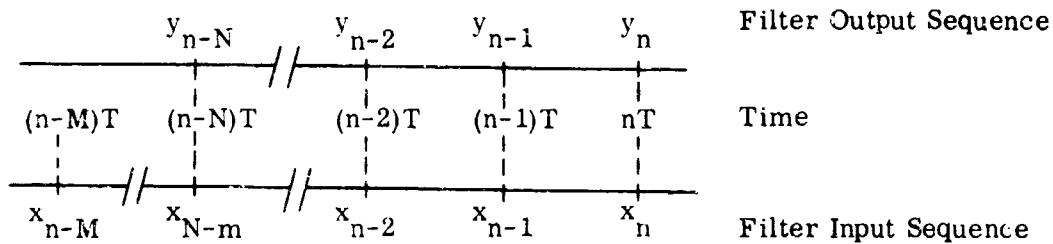


Figure 3.1-1. Input-Output Sequence Relation

The output sequence member y_n is produced by taking the past N output sequence members back to y_{n-N} and multiplying them by the b 's, summing them and subtracting them from the similarly multiplied (by the a 's) and summed present input sequence member x_n and the past M input sequence members. The a 's and b 's are multiplied using the ones with smallest indices on the most recent data. That is $a_0 \cdot x_n, a_1 \cdot x_{n-1}$, etc. and $b_0 \cdot y_n, b_1 \cdot y_{n-1}$ etc. The complete form (equation 3.1-1) will be discussed with the a 's and b 's determined by the synthesis techniques of other sections or the problem at hand.

The linearity of the filter is apparent. If the input sequence is the sum of a number of sequences then the output is the sum of the responses to the individual input sequences. This follows directly from the filter definition.

A "time invariant" filter means simply that the coefficients a_m and b_n do not depend upon the time at which filtering is occurring. Mathematically, $\frac{d}{dt}$ means that they do not depend upon the value j of the input and output sequences.

3.2 DESCRIPTION OF DIGITAL FILTERS BY THE z TRANSFORM

If the z transform of the digital filter, defined by equation 3.1-1, is taken

$$\mathcal{Z} \left(b_0 y_j + \sum_{n=1}^N b_n y_{j-n} - \sum_{m=0}^M a_m x_{j-m} \right) \quad (3.2-1)$$

then,

$$\sum_{n=0}^N b_n \sum_{j=0}^{\infty} y_{j-n} z^{-j} = \sum_{m=0}^M a_m \sum_{j=0}^{\infty} x_{j-m} z^{-j}, \quad (3.2-2)$$

where the linearity property of the z transform (equation 2.3-5) has been used. Changing indices by the substitutions, $j' = j-n$ and $j'' = j-m$, gives:

$$\left[\sum_{n=0}^N b_n z^{-n} \right] \left[\sum_{j=0}^{\infty} y_j z^{-j} \right] = \left[\sum_{m=0}^M a_m z^{-m} \right] \left[\sum_{j=0}^{\infty} x_j z^{-j} \right] \quad (3.2-3)$$

where the primes on the new indices have been removed as they are only dummy indices of summation. It is also assumed that x_j and y_j are identically zero for $j < 0$. Assuming the input and output sequences are zero for $j < 0$ is not a restrictive assumption as physical systems cannot be observed in the infinite past, and as there is liberty to define the time origin.

Now

$$X(z) = \sum_{j=0}^{\infty} x_j z^{-j}, \quad (3.2-4a)$$

and

$$Y(z) = \sum_{j=0}^{\infty} y_j z^{-j}, \quad (3.2-4b)$$

are the z transforms of the input and output sequences, respectively. The relationship between these sequences is seen to be:

$$\frac{Y(z)}{X(z)} = \frac{\sum_{m=0}^M a_m z^{-m}}{\sum_{n=0}^N b_n z^{-n}} \quad (3.2-5)$$

Therefore the input-output relation of a digital filter may be described by a "z transform transfer function":

$$H_z(z) = \frac{\sum_{m=0}^M a_m z^{-m}}{\sum_{n=0}^N b_n z^{-n}}, \quad (3.2-6)$$

$$= \frac{Y(Z)}{X(Z)}$$

which is a major result.

3.3 FREQUENCY BEHAVIOR OF THE z TRANSFER FUNCTION

3.3.1 Complex Frequency Response

To investigate the behavior of the difference equation on a sinusoidal input sequence let

$$x_n = e^{i\omega nT} \quad (3.3.1-1)$$

and assume y_n is given by

$$y_n = G(\omega) e^{i\omega nT} \quad (3.3.1-2)$$

Inserting these expressions into equation (3.1-1) gives

$$G(\omega) \sum_{n=0}^N b_n e^{i\omega jT} e^{-i\omega nT} = \sum_{m=0}^M a_m e^{i\omega jT} e^{-i\omega mT},$$

or

$$G(\omega) = \frac{\sum_{m=0}^M a_m e^{-i\omega mT}}{\sum_{n=0}^N b_n e^{-i\omega nT}} \quad (3.3.1-3)$$

It is seen that the sinusoidal sequence is modified in amplitude and phase by $G(\omega)$ which is simply $H_z(z)$ above (3.2-6) with $z = e^{j\omega T}$; namely,

$$G(\omega) = H_z(e^{j\omega T}). \quad (3.3.1-4)$$

Hereafter, the notation for the frequency response, setting $H_z(\omega) = H_z(e^{j\omega T})$, is abridged in an obvious notational simplification.

Given this frequency response interpretation of the transfer function it is remarkably simple to express the input-output sequence relation in terms of frequency transforms exactly as for analog systems.

The input-output z transform relation is written

$$Y(z) = H_z(z) X(z) \quad (3.2-5)$$

Taking the inverse z transforms (eq. 2.4.1-1)

$$y_n = \oint H_z(z) X(z) z^n \frac{dz}{z}$$

Now, if all the poles of $Y(z)$ are inside the unit circle then an integration along this circle with

produces

$$y_n = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H_z(e^{j\omega T}) X(e^{j\omega T}) e^{j\omega n T} d\omega \quad (3.3.1-6)$$

which, with $\omega_s = 2\pi/T$, looks almost exactly like an inverse analog system Fourier transform.

3.3.2 Transfer Function Definition Without Recourse to Complex Numbers

It is possible to define the transfer function of a digital (or analog) filter without using complex numbers, through the resulting formulae are much more cumbersome and make transform methods much more obscure as will be seen.

Take as the input the real function

$$x_n = \sin(\omega n T - \epsilon) \quad (3.3.2-1)$$

where ϵ is an arbitrary phase angle. If $\epsilon = \pi/2 + \epsilon'$, ϵ' arbitrary, then a cosinusoidal sequence would be the input.

Allowing the filter to operate upon this input sequence, and assuming an output sequence

$$y_n = T(\omega) \sin(\omega nT + \epsilon + \theta(\omega)) \quad (3.3.2-2)$$

it is found, after considerable algebra that:

$$\begin{aligned} & T(\omega) \cos \theta(\omega) \sum_{n=0}^N b_n (\sin(\omega jT + \epsilon) \cos \omega nT - \cos(\omega jT + \epsilon) \sin \omega nT) \\ & + T(\omega) \sin \theta(\omega) \sum_{n=0}^N b_n (\cos(\omega jT + \epsilon) \cos \omega nT + \sin(\omega jT + \epsilon) \sin \omega nT) \\ & = \sum_{m=0}^M a_m (\sin(\omega jT + \epsilon) \cos \omega mT - \cos(\omega jT + \epsilon) \sin \omega mT) \end{aligned} \quad (3.3.2-3)$$

As time jT is arbitrary, equate coefficients of $\sin(\omega jT + \epsilon)$ and $\cos(\omega jT + \epsilon)$ on both sides of the equation and obtain:

$$\begin{aligned} & T(\omega) \cos \theta(\omega) \sum_{n=0}^N b_n \cos \omega nT + T(\omega) \sin \theta(\omega) \sum_{n=1}^N b_n \sin \omega nT \\ & = \sum_{m=0}^M a_m \cos \omega mT \end{aligned} \quad (3.3.2-4a)$$

and

$$\begin{aligned} & T(\omega) \cos \theta(\omega) \sum_{n=1}^N b_n \sin \omega nT - T(\omega) \sin \theta(\omega) \sum_{n=0}^N b_n \cos \omega nT \\ & = \sum_{m=1}^M a_m \sin \omega mT. \end{aligned} \quad (3.3.2-4b)$$

These equations can be solved simultaneously for $T(\omega) \cos \theta(\omega)$ and $T(\omega) \sin \theta(\omega)$ as they are linear in these quantities. The result:

$$\begin{aligned} T(\omega) \cos \theta(\omega) &= \frac{AC + BD}{A^2 + B^2}, \\ T(\omega) \sin \theta(\omega) &= \frac{AD - BC}{A^2 + B^2}, \end{aligned} \quad (3.3.2-5)$$

where

$$A = \sum_{n=0}^N b_n \cos \omega nT, \quad B = \sum_{n=1}^N b_n \sin \omega nT,$$

$$C = \sum_{m=0}^M a_m \cos \omega mT, \quad D = \sum_{m=1}^M a_m \sin \omega mT.$$

Squaring and adding the equations it is found, after more algebra, that:

$$T^2(\omega) = \frac{C^2 + D^2}{A^2 + B^2}.$$

or

$$T(\omega) = \sqrt{\frac{\left(\sum_{m=1}^M a_m \sin \omega mT\right)^2 + \left(\sum_{m=0}^M a_m \cos \omega mT\right)^2}{\left(\sum_{n=1}^N b_n \sin \omega nT\right)^2 + \left(\sum_{n=0}^N b_n \cos \omega nT\right)^2}}$$

Similarly, division of the two equations gives

$$\Theta(\omega) = \tan^{-1} \left[\frac{BC - AD}{AC + BD} \right]$$

$$= \tan^{-1} \left\{ \frac{-\sum_{n=1}^N b_n \sin \omega nT \sum_{m=0}^M a_m \cos \omega mT + \sum_{n=0}^N b_n \cos \omega nT \sum_{m=1}^M a_m \sin \omega mT}{\sum_{n=1}^N b_n \sin \omega nT \sum_{m=1}^M a_m \sin \omega mT + \sum_{n=0}^N b_n \cos \omega nT \sum_{m=0}^M a_m \cos \omega mT} \right\}$$

After a moments comparison with the complex form of the transfer function in the preceding section, it is seen that

$$H_z(\omega) = T(\omega) e^{i\Theta(\omega)} \quad (3.3.2-8)$$

Obviously, complex numbers are much easier to work with, though this exercise shows clearly that the amplitude modification of a sine (cosine) sequence is $T(\omega) = |H(\omega)|$ and the phase shift $\Theta(\omega)$ is given by the inverse tangent of the ratio of the imaginary to real components of $H_z(\omega)$.

3.3.3 Transfer Function Time (Or Phase) Reference

A great deal of care must be taken when defining a transfer function and in interpreting it. While the magnitude of the transfer function is quite easy to understand, it is the amplitude ratio of the output to input at a particular frequency, the phase characteristic can lead to a great deal of ambiguity in questions of filter synthesis and in applications of the filters.

The filter is defined such that the phase reference is with respect to simultaneous inputs and outputs; that is, the phase is defined between members y_j of the output sequence and x_j of the input sequence for the same j .

This definition of phase will be seen to imply a delay in the output sequence at a given frequency with respect to the input sequence when the phase is negative; and an advance of the output sequence at a given frequency when the phase is positive.

Further, a linear phase shift as a function of frequency $\Theta(\omega) = \pm k\omega T$, $k = 0, 1, 2, \dots$ will correspond to an advanced (+) or delay (-) of " k " units in the output sequence "in toto", not just at a particular frequency.

First, the remark about the sign of the phase at a given frequency determining the delay or advance is proven. From Section 3.3.1 it can be seen that the output complex sinusoid sequence is given as

$$y_n = H_z(\omega) e^{i\omega n T},$$

with

$$x_n = e^{i\omega n T}$$

the input sequence. Writing

$$H_z(\omega) = |H_z(\omega)| e^{i\Theta(\omega)} \quad (3.3.3-1)$$

where $\Theta(\omega)$ is the "phase" of the complex number $H_z(\omega)$ gives

$$y_n = H_z(\omega) e^{i(\omega nT + \Theta(\omega))}$$

or

(3.3.3-2)

$$y_n = H_z(\omega) e^{i\omega(nT + \frac{\Theta(\omega)}{\omega})}$$

As stated, it can be seen that $|H_z(\omega)|$ modifies the amplitude at a given frequency ω and there is a time shift of $\Theta(\omega)/\omega$. If $\Theta(\omega)$ is positive there is an "advance", and when negative there is a "delay" in the output sinusoid.

Next, it is seen that the time shift, " τ ", is

$$\tau = \frac{\Theta(\omega)}{\omega} = \pm kT \quad (3.3.3-3)$$

for $\Theta(\omega) = \pm k\omega T$, a linear phase shift. As a general bandlimited input sequence can be written as the superposition (sum) of sinusoids (equation (3.3.1-6)), then the entire sequence will have a time shift $\tau = \pm kT$.

3.4 DEFINITION OF THE "IMPULSE RESPONSE" (AND FILTER DEFINITION FROM THE "IMPULSE RESPONSE")

To this point digital filters have been defined by their transfer functions which were functions of coefficients a_m and b_n . A property of digital filters, their "impulse response" can be defined, which in turn can be used to define the transfer function and, concomitantly, the a_m 's and b_n 's. The impulse response, w_n , is defined by the reaction of a filter to the sequence (the impulse)

$$x_n = \begin{cases} 1, & n = 0, \\ 0, & \text{otherwise.} \end{cases} \quad (3.4-1)$$

Obviously, a physically meaningful impulse response can only have values for output times of 0, T, 2T, ..., etc.

Relation (3.4-1) and the definition of the z transform gives the output sequence, the impulse response, in terms of its z transform, $W(z)$:

$$\begin{aligned} W(z) &= \sum_{n=0}^{\infty} w_n z^{-n}, \\ &= H_z(z)X(z), \\ &= H_z(z), \end{aligned} \quad (3.4-2)$$

as $X(z) = 1$. Taking the inverse transform gives

$$w_n = \frac{1}{2\pi i} \oint H_z(z) z^n \frac{dz}{z}; \quad n = 0, 1, 2, \dots \quad (3.4-3)$$

An obvious physical restriction (causality) that no future inputs are required before the present output is computed, disallows a "double-sided" z transform for $H_z(z)$. That is, there cannot occur

$$H_z(z) = \sum_{n=0}^{\infty} w_n z^{-n} + \sum_{n=1}^{\infty} w_n z^n \quad (3.4-4)$$

where positive powers of z are included. If positive powers of z are allowed then there could be outputs prior to the impulse. Therefore, it must be true that $w_n = 0$, for $n < 0$.

Interestingly, and usefully, the integral (3.4-3) can be computed for successive values of n by "impulsing" the digital filter equivalent to $H_z(z)$. $H_z(z)$ must, of course, be expressible as a rational function of z^{-1} .

Stable filters which are of the form (3.1-1) guarantee that $w_n = 0$ for $n < 0$ as expression (3.4-3) is zero for $n < 0$ in this case.

The simplest method for computing the impulse response is applying x_n above to the filter and observing the output.

3.4.1 Filter Input-Output Relation in Terms of the Impulse Response (Convolution)

The output of the digital filter can be expressed directly in terms of the "impulse response" w_n and a general input sequence x_n .

$H_z(z)$ is defined as the ratio of input and output transforms:

$$Y(z) = H_z(z) X(z); \quad (3.2-6)$$

and using the expansions of the various transforms produces

$$\begin{aligned} \sum_{k=0}^{\infty} y_k z^{-k} &= \sum_{n=0}^{\infty} w_n z^{-n} \sum_{j=0}^{\infty} x_j z^{-j} \\ &= \sum_{j=0}^{\infty} \left[\sum_{n=-j}^{\infty} w_n x_j z^{-n} z^{-j} \right] \end{aligned} \quad (3.4.1-1)$$

where the index n has been extended below zero, as $w_{-n} = 0$. The purpose of this index extension is to allow a change of variables of summation: namely, let $k = n + j$ in the inside summation:

$$\begin{aligned} \sum_{k=0}^{\infty} y_k z^{-k} &= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} w_{k-j} x_j z^{-k}, \\ &= \sum_{k=0}^{\infty} \left[\sum_{j=0}^{\infty} w_{k-j} x_j \right] z^{-k}. \end{aligned} \quad (3.4.1-2)$$

Comparing terms on both sides (or taking inverse transforms) gives:

$$y_k = \sum_{j=0}^{\infty} w_{k-j} x_j, \quad (3.4.1-3)$$

which is the "convolution" of the input and impulse response sequence, and is a basic result. In fact, this equation is often taken as the starting point of linear digital, time invariant filter theory and defines these filters. Either the "impulse response" or the "transfer function" are equally good starting points. In statistical analysis, the convolution relation between input and output is frequently used to extract averages.

As the initial time from which transforms are taken is arbitrary (even though zero was used) the relation (3.4.1-3) holds for all k ($w_j = 0, j < 0$).

The input-output relation can also be written in another form. Note that the argument leading to the present format, is symmetric in the variables w_n and x_j , n and j both being dummy variables of summation, with the assumption that $x_j = 0, j < 0$. Therefore, it can be immediately stated that:

$$y_k = \sum_{j=0}^{\infty} w_j x_{k-j}. \quad (3.4.1-4)$$

The statement can also be written for all k as the time origin on the sequences y_k and x_k is arbitrary.

3.4.2 Filter Definition and Stability from the Impulse Response

If equation (3.4.1-4) is accepted as the definition of a linear time invariant digital filter, it is of interest to determine the conditions on the w_j , the impulse response, and the transfer function $H_z(z)$ defined in terms of the w_j .

First, if the filter is "causal" there can be no future inputs required to compute the present output which confines the impulse response:

$$w_j = 0, \quad j < 0, \quad (3.4.2-1)$$

as before.

Stability is the next fundamental requirement usually imposed on filters. A stable, linear, digital filter is usually defined by the statement.

A linear, digital filter is "stable" if the output sequence in response to a bounded input sequence is bounded.

To see what this definition implies take the magnitude of eq. (3.4.1-4)

$$|y_k| = \left| \sum_{j=0}^{\infty} w_j x_{k-j} \right| \quad (3.4.2-2)$$

which obviously satisfies the inequality

$$|y_k| \leq \sum_{j=0}^{\infty} |w_j| |x_{k-j}| \quad (3.4.2-3)$$

By a bounded input sequence the following condition is meant:

$$|x_k| \leq R, \quad \text{for all } k, \quad (3.4.2-4)$$

where R is some finite number. Therefore,

$$|y_k| \leq R \sum_{j=0}^{\infty} |w_j|, \quad (3.4.2-5)$$

and if y_k , the output sequence, is to be bounded then a sufficient condition on the impulse response is

$$\sum_{j=0}^{\infty} |w_j| < \infty; \quad (3.4.2-6)$$

the sum of the impulse response magnitudes is bounded.

Necessity of this condition can be determined by finding a bounded input sequence which needs this condition. An input sequence which is bounded and has signs such that the terms are all positive makes this condition necessary.

Stability is obtained, then, if the filter impulse response satisfies the necessary and sufficient condition

$$\sum_{j=0}^{\infty} |w_j| < \infty$$

namely, bounded.

This condition, though mathematically implied by the definition of stability, is not as directly useful as a condition on the transfer function

$$H_z(z) = \sum_{k=0}^{\infty} w_k z^{-k}. \quad (3.4-2)$$

To find the conditions on $H_z(z)$, the obvious inequality

$$\sum_{k=0}^{\infty} w_k z^{-k} \leq \sum_{k=0}^{\infty} |w_k| \cdot |z^{-k}| \quad (3.4.2-7)$$

is noted. For $|z|^{-1} \leq 1$, or $|z| > 1$, the condition (3.4.2-7) is certainly satisfied. Therefore, a condition on $H_z(z)$ for a stable system is that it be an analytic function for $|z| > 1$.

A system with a transfer function having one or more singularities outside of the unit circle satisfies, mathematically,

$$\sum_{k=0}^{\infty} w_k z^{-k} \Rightarrow \infty, \quad (3.4.2-8)$$

for some z such that $|z| > 1$.

As

$$\sum_{k=0}^{\infty} |w_k| > \sum_{k=0}^{\infty} w_k z^{-k}, \quad |z| > 1; \quad (3.4.2-9)$$

then

$$\sum_{k=0}^{\infty} |w_k| \Rightarrow \infty, \quad (3.4.2-10)$$

if a singularity exists outside the unit circle.

Summarizing:

1. A stable linear digital filter transfer function, $H_z(z)$, has no singularities or poles outside of the unit circle. Some singularities must exist within, or on, the unit circle if all the w_j are to be non-zero.

2. An unstable linear digital filter transfer function, $H_z(z)$, has singularities or poles outside the unit circle.

3.4.3 Filter Stability Via the Difference Equation

The criteria for stability of the output sequence of the difference equation 3.1-1 from the basic definition is developed. If the input is zero the output should remain bounded is the criterion used.

For zero input the equation is

$$\sum_{n=0}^N b_n y_{j-n} = 0, \quad b_0 = 1. \quad (3.4.3-1)$$

Assuming a solution,

$$y_j = \beta^j; \beta, \text{ a constant}, \quad (3.4.3-2)$$

it is found that β must satisfy a "characteristic equation"

$$\sum_{n=0}^N b_n \beta^{-n} = 0; \quad (3.4.3-3)$$

namely, a solution of a polynomial of the N^{th} degree. In general, there are N roots which may or may not be distinct. If they are distinct, then the most general solution to (3.4.3-1) is

$$y_j = \sum_{k=1}^N B_k \beta_k^j \quad (3.4.3-4)$$

where the B_k are determined by the initial conditions. Furthermore, if the b_n are real, the roots β_k when complex must occur in conjugate pairs.

For the y_j 's to remain bounded, the magnitude of the β 's must be less than unity. This is the stability criteria. If the roots are not distinct, say, there is an "s fold" root, then the part of the solution, $y_j^{(s)}$, corresponding to this root is

$$y_j^{(s)} = \sum_{k=1}^s C_{k-1} j^{k-1} \cdot \beta_s^j; \quad C'_k s = \text{constants}. \quad (3.4.3-5)$$

Again, it is seen that β must have a magnitude less than unity for a bounded solution. Therefore, in general, stability is determined by the roots of the characteristic equation:

$$\sum_{n=0}^N b_n z^{-n} = 0. \quad (3.4.3-6)$$

lying within the unit circle in the complex z plane. In other words, the poles of the z transfer function must lie within the unit z -plane circle for stability.

Non-recursive filters with a finite number of coefficients are automatically stable as no characteristic equation exists.

A more general discussion of filter stability was considered in Section 3.4.2 where the "impulse response" of a digital filter was discussed.

3.5 SOME PROPERTIES OF DIGITAL FILTERS

From the form (3.3.1-3) it is seen that digital filters possess certain properties.

The transfer function is periodic in the sampling frequency;

$$\begin{aligned} H_z(\omega + \frac{2\pi}{T}) &= H_z(\omega + \omega_s) \\ &= H_z(\omega), \end{aligned} \quad (3.5-1)$$

If it also true that the filter coefficients a_n and b_n are real numbers then the filter properties for negative frequencies are known from the behavior for positive frequencies as

$$H_z(-\omega) = H_z^*(\omega), \quad * = \text{complex conjugation.} \quad (3.5-2)$$

For real filters properties (3.5-1) and (3.5-2) imply that

$$H_z(0) = H_z^*(0), \quad (3.5-3a)$$

and

$$\begin{aligned} H_z(-\frac{\pi}{T}) &= H_z(\frac{\pi}{T}), \\ &= H_z(\omega_s/2), \end{aligned} \quad (3.5-3b)$$

namely; digital transfer functions with real coefficients are real (possess only zero and 180° phases) at $\omega = 0$ and $\omega = \omega_s/2$.

Now, property (3.5-2) doesn't produce any difficulties in approximating continuous data filters as they also possess the same property; but property (3.5-1),

the periodicity and consequent reality at $\omega = \frac{\omega_s}{2}$, is not usually found in continuous filters and produces approximation difficulties.

3.6 SERIES FILTERS

As with analog filters, the transfer function of successive digital filtering operations is of considerable interest. The transform methods used here are equally applicable to recursive and non-recursive filters.

The situation is as follows. There is an input sequence, x_n , which is operated upon to produce an output sequence, y_n , namely

$$b_o^{(1)} y_j = - \sum_{n=1}^{N_1} b_n^{(1)} y_{j-n} + \sum_{m=0}^{M_1} a_m^{(1)} x_{j-m}; \quad b_o^{(1)} = 1, \quad (3.6-1)$$

Taking z transforms; as before, produces

$$Y(z) = H_z^{(1)}(z) X(z) \quad (3.6-2)$$

where $Y(z)$ and $X(z)$ are the transforms of the output and input, respectively; and

$$H_z^{(1)}(z) = \frac{\sum_{m=0}^{M_1} a_m^{(1)} z^{-m}}{\sum_{n=0}^{N_1} b_n^{(1)} z^{-n}} \quad (3.6-3)$$

as before.

Next, pass the output sequence, y_j , through another digital filter to produce an output sequence v_j . This filter is defined as

$$b_o^{(2)} v_j = - \sum_{q=1}^{N_2} b_q^{(2)} v_{j-q} + \sum_{k=0}^{M_2} a_k^{(2)} y_{j-k}; \quad b_o^{(2)} = 1. \quad (3.6-4)$$

The input-output z transform relationship is

$$V(z) = H_z^{(2)}(z) Y(z) \quad (3.6-5)$$

with $V(z)$ the output sequence transform, and

$$H_z^{(2)}(z) = \frac{\sum_{k=0}^{M_2} a_k^{(2)} z^{-k}}{\sum_{q=0}^{N_2} b_q^{(2)} z^{-q}} \quad (3.6-6)$$

The relation between the input sequence to the first filter, x_j , and the output of the second filter v_j , is given obviously, in terms of z transforms as

$$\begin{aligned} \frac{V(z)}{X(z)} &= H_z^{(2)}(z) H_z^{(1)}(z), \\ &= H_z^{(1)}(z) H_z^{(2)}(z), \text{ (the filtering is commutative),} \end{aligned} \quad (3.6-7)$$

$$= \frac{\sum_{m=0}^{M_1} a_m^{(1)} z^{-m} \sum_{k=0}^{M_2} a_k^{(2)} z^{-k}}{\sum_{n=0}^{N_1} b_n^{(1)} z^{-n} \sum_{q=0}^{N_2} b_q^{(2)} z^{-q}}$$

$$= H_z(z),$$

with $H_z(z)$ being the product filter. Simply multiply the transforms to find the composite filter that directly produces the sequence v_j from x_j . The equivalent digital filter relation is given by the coefficients of the multiplication. Note that $H_z(z)$ has the form

$$H_z(z) = \frac{\sum_{m=0}^{M_1 + M_2} a_m z^{-m}}{\sum_{n=0}^{N_1 + N_2} b_n z^{-n}} \quad (3.6-8)$$

where a_m and b_n are the coefficients of the product coefficients of the numerator and denominator of equation (3.6-7), respectively. The equivalent difference equation is,

$$v_j = - \sum_{n=1}^{N_1+N_2} b_n v_{j-n} + \sum_{m=0}^{M_1+M_2} a_m x_{j-m} \quad (3.6-9)$$

3.7 REALIZATION METHODS

When a digital filter has been synthesized, as far as determining the coefficients of the numerator and denominator of the z transfer function, it may be useful to synthesize the actual numerical filter in other ways. Reasons for desiring other "realization" methods include the desires for faster computation times, less numerical operations, and less storage in the computer.

Five basic "realization" methods are:

1. Direct Programming
2. Factorization
3. Partial Fraction
4. Simple Product
5. Discrete Fourier Transformation Using the Fast Fourier Transformation Algorithm

3.7.1 Direct Programming

"Direct programming" means just that. Take $H_z(z)$ and use its equivalent difference equation as the filter realization. Namely, when $H_z(z)$ is given in the form

$$H_z(z) = \frac{\sum_{m=0}^M a_m z^{-m}}{\sum_{n=0}^N b_n z^{-n}} \quad b_0 = 1; \quad (3.2-6)$$

use the formula

$$y_j = - \sum_{n=1}^N b_n v_{j-n} + \sum_{m=0}^M a_m x_{j-m} \quad (3.1-1)$$

for filtering. In fact, $H_z(z)$ was considered as derived from this basic filter definition in the initial investigations of digital filters.

3.7.2 Factorization

This method of realization assumes that $H_z(z)$ is factored from the form

$$H_z(z) = \frac{\sum_{m=0}^M a_m z^{-m}}{\sum_{n=0}^N b_n z^{-n}} \quad (3.2-6)$$

into

$$H_z(z) = \frac{a_0 (1 - \alpha_1 z^{-1}) (1 - \alpha_2 z^{-1}) \dots (1 - \alpha_M z^{-1})}{(1 - \beta_1 z^{-1}) (1 - \beta_2 z^{-1}) \dots (1 - \beta_N z^{-1})} \quad (3.7.2-1)$$

The α_k and β_k are the roots of the numerator and denominator polynomials which, if they are complex, must exist in complex conjugate pairs when the a_m and b_n are all real.

Then, look upon $H_z(z)$ as a series of simple filters (See Section 3.6)

$$H_z(z) = H_z^{(1)}(z) H_z^{(2)}(z) \dots H_z^{(N)}(z), \quad (3.7.2-2)$$

where

$$H_z^{(k)}(z) = \begin{cases} \frac{(1 - \alpha_k z^{-1})}{(1 - \beta_k z^{-1})}, & 1 \leq k \leq M, \\ \frac{1}{(1 - \beta_k z^{-1})}, & M + 1 \leq k \leq N - 1, \\ \frac{a_0}{1 - \beta_N z^{-1}}, & k = N. \end{cases}$$

if it is assumed that $M \leq N$ (the numerator of (3.2-6) can always be factored into two parts such that $M \leq N$ for one factor, the other factor becomes a non-recursive filter).

The implementation of one of the filters, say, $H_z^{(k)}(z)$, is

$$x_j^{(k)} = \beta_k x_{j-1}^{(k)} + x_j^{(k-1)} - \alpha_k x_{j-1}^{(k-1)} \quad (3.7.2-3)$$

where

$$x_j^{(k)} = \text{output sequence of } k^{\text{th}} \text{ series filter.}$$

Each simple series filter operates on the output of the simple filter to its left. $H_z^{(1)}(z)$, of course, operates on the actual input sequence x_n and the output of $H_z^{(N)}(z)$ after the operation of all the others is y_n , which would have been obtained if $H_z(z)$ had operated directly on x_n . The order of the usage of the filters $H_z^{(k)}(z)$ is unimportant, but they must all be used in the total filtering operation.

The primary advantage of this type of realization is the ease of mechanization of the simple formula (3.7.2-3) which is used N times while filtering, merely changing the constants each time.

This method of realization can be considered an "iterative" method, in programming terminology, as the same formula is applied again and again, N times.

3.7.3 Partial Fraction

The partial fraction method of realization is just what is meant by its name. Expand $H_z(z)$ into its partial fraction form:

$$H_z(z) = \sum_{k=1}^N \frac{A_k}{1 - \beta_k z^{-1}} + \sum_{r=0}^M B_r z^{-r} \quad (3.7.3-1)$$

where A_k and B_r are the expansion coefficients and β_k are the denominator roots. This method is obviously not applicable to non-recursive filters as there is no denominator in $H_z(z)$ to expand. Assume the β_k 's distinct for simplicity. An obvious modification results if a root has a multiplicity greater than unity.

An individual partial fraction term can be realized as a filter operating on the input sequence x_j and gives a partial output $x_j^{(k)}$, namely the term

$$\frac{A_k}{1 - \beta_k z^{-1}} \quad (3.7.3-2)$$

has the realization

$$x_j^{(k)} = \beta_k x_{j-1}^{(k)} + A_k x_j \quad (3.7.3-3)$$

where $x_j^{(k)}$ is the partial output sequence due to this term in the partial fraction expansion. The total output of $H_z(z)$ is the sum of the partial outputs

$$y_j = \sum_{k=1}^N x_j^{(k)} + \sum_{r=1}^M B_r x_{j-r} \quad (3.7.3-4)$$

This approach has the advantage of possessing the capability of being the fastest, when implemented with a special purpose machine, as all of the partial sequences may be computed simultaneously (in parallel), rather than sequentially, as with the direct and factorization methods. Storage is also reduced as only the immediate past $x_i^{(k)}$'s, of which there are N for distinct roots need be stored, excluding the non-recursive storage defined by the B_r 's

3.7.4 Simple Product

Another realization method that has the simplicity of the "Direct Programming" method and the minimized storage of the "Partial Fraction" method is a simple product factoring of the expression for $H_z(z)$. Namely, write

$$H_z = \frac{\sum_{m=0}^M a_m z^{-m}}{\sum_{n=0}^N b_n z^{-n}}$$

as

$$H_z = A(z) B(z) \quad (3.7.4-1)$$

with

$$A(z) = \sum_{m=0}^M a_m z^{-m} \quad (3.7.4-2)$$

and

$$B(z) = \frac{1}{\sum_{n=0}^N b_n z^{-n}} \quad (3.7.4-3)$$

Operate on the input sequence first with $B(z)$ and on this sequence with $A(z)$ to form the output (see Section 3.6). If $Y(z)$ is the output transform and $X(z)$ the input transform

$$\begin{aligned} Y(z) &= H_z(z) X(z) \\ &= A(z) [B(z) X(z)], \\ &= A(z) V(z) \end{aligned} \quad (3.7.4-4)$$

with

$$V(z) = B(z) X(z). \quad (3.7.4-5)$$

The implementations of (3.7.4-4) and (3.7.4-5) are

$$b_0 v_j = - \sum_{n=1}^N b_n v_{j-n} + x_j, \quad (3.7.4-6)$$

and

$$y_j = \sum_{m=0}^M a_m v_{j-m}; \quad (3.7.4-7)$$

respectively.

Only the N or M past v_j 's need be stored.

3.7.5 Discrete Fourier Transform Realization

The fast Fourier transform algorithm (section 14), used to compute the discrete Fourier transform, allows the non-recursive filtering operation

$$y_j = \sum_{m=0}^{M-1} a_m x_{j-m}, \quad j = 0, 1, \dots, R-1 \quad (3.7.5-1)$$

to be performed "indirectly" with less computation, though with more delay, for large M (Section 3.7.5.5). Indirect computation means that the above formula, though defining the filtering operation is computed by another method which is completely equivalent, neglecting round-off. The numbers y_n , a_m , and x_m are allowed to be complex though certain computational improvements will be discussed when they are real. This method uses the discrete convolution relationship (Section 2.7.2) with appropriate modification to account for the periodic, cyclic, or modulo N nature of the discrete Fourier transform.

Repeating here the convolution relation in terms of the DFT:

$$\begin{aligned} \overline{y_j} &= \sum_{m=0}^{N-1} \overline{a_m} \overline{x_{j-m}} \\ &= \frac{1}{N} \sum_{n=0}^{N-1} \overline{A_n} \overline{X_n} W^{jn}, \quad W = e^{i \frac{2\pi}{N}} \end{aligned} \quad (2.7.2-1)$$

where

$$\overline{A}_n = \sum_{k=0}^{N-1} \overline{a}_k W^{-kn}$$

and

$$\overline{X}_n = \sum_{k=0}^{N-1} \overline{x}_k W^{-kn}$$

Use of this relation implies that the quantities $\overline{a}_m$, $\overline{x}_m$, $\overline{A}_n$, and $\overline{X}_n$ repeat periodically outside of the region: $0 \leq m, n \leq N-1$. By appropriately appending zeros to the original sequences, a_m and x_m , sequencing them properly, and choosing N appropriately the discrete convolution can be constrained so that $\overline{y}_j = y_j$.

First, note that (2.7.2-1) can be written

$$\overline{y}_j = \sum_{m=0}^j \overline{a}_m \overline{x}_{j-m} + \sum_{m=j+1}^{N-1} \overline{a}_m \overline{x}_{N+j-m} \quad (3.7.5-2)$$

where the periodicity of $\overline{x}_m$ has been used. Two equivalent methods are available for computing the desired filtering operation via (2.7.2-1) with zeroes appended to avoid the periodicity. Both methods repeatedly compute (2.7.2-1) via the discrete Fourier transform with a value $N=N_{\text{opt}}$ until R data points x_m have been filtered.

The methods described subsequently follow directly the work of Helms, Sande, and Stockham as described by Helms*. It should be mentioned that Helms method of generating the non-recursive digital filter coefficients from recursive filters or transfer function samples are somewhat naive and should be replaced by the methods of Section 5.7 which still uses the discrete Fourier transform for generation but also reduces the number of filter coefficients and uses convergence methods to reduce the "Gibbs phenomena".

* Fast Fourier Method of Computing Difference Equations and Simulating Filters, H. D. Helms, IEEE Trans on A & E. Vol. Au-15, NO. 2, June 1967.

Terminology also follows that of Helms.

3.7.5.1 The "Select-Saving" Method

The technique is best described by giving the computational procedure:

- 1) A number N , whose value will be discussed subsequently in section 3.7.5.3, is chosen such that $N \geq M$.
- 2) $N-M$ zeroes are appended to the sequence a_m , producing the desired sequence

$$\bar{a}_m = \begin{cases} a_m, & m = 0, 1, \dots, M-1; \\ 0, & m = M, M+1, \dots, N-1. \end{cases}$$

- 3) The discrete Fourier transform (say, DFT1) of the sequences $\bar{a}_m$ and $\bar{x}_m$, where $\bar{x}_m$ is chosen in this step as equal to the first N values of x_m , is taken. The resultant frequency sequences are directly multiplied at corresponding frequencies. It should be noted that the discrete Fourier transform of the filter weighting sequence, $\bar{a}_m$, need only be computed once.
- 4) The inverse discrete Fourier transform (IDFT1) of the product sequence formed in (3) is taken.
- 5) The resultant time sequence $\bar{y}_j$ is the desired output y_j for values of j :

$$j = M-1, M, \dots, N-1.$$

Values of $\bar{y}_j$ for

$$j = 0, 1, \dots, M-2$$

are considered as invalid and are discarded. "Select-saving" is regarded by Helms as being defined by this operation.

- 6) The sequence x_m is then redefined for the next filtering operation. Discard the present values of $\bar{x}_m$ and replace them by N values of x_m starting with the $N-L+2^{\text{th}}$ member of the present set of values.
- 7) Repeat steps (3) through (6) until $R+M-1$ values of x_m have been used.
- 8) The values of y_j obtained by step (5) are placed end to end to form the filtered sequence.

This procedure works because the second term on the right side of 3.7.5-2 vanishes if j is restricted to $M-1 \leq j \leq N-1$ as $m > j$ and $a_m = 0$ for $m \geq M$ in that term.

3.7.5.2 The "Overlap-Adding" Method

This method is a slight variation of the previous method.

Steps (1) and (2) are identical to that of the previous method.

- 3) The discrete Fourier transform (DFT1) of the sequences $\bar{a}_m$ and $\bar{x}_m$ is taken. $\bar{x}_m$ is defined this time as the first $N-M+1$ values of x_m with $M-1$ zeros appended. The product frequency sequence is formed.
- 4) The inverse discrete Fourier transform (IDFT1) of the product sequence is taken.
- 5) The N output values y_j 's are shifted $N-M+1$ units towards larger input indices and added to the previous y_j 's. Obviously, the first time addition is not necessary. This shifting and adding begets the name, "overlap-adding", for the technique. Only the first $M-1$ values are actually added to the previous values.

- 6) The sequence $\bar{x}_m$ is redefined by replacing the previously used $N-M+1$ values of x_m by the next $N-M+1$ values of x_m with zeros appended
- 7) Steps (3) through (6) are repeated until $R+M-1$ values of x_m are filtered.

Veracity of this process is best shown by appeal to equation 3.7.5-2 plus a few examples of the "end-effect" summing worked out by hand by the reader.

3.7.5.3 Determination of N to Minimize the Number of Computations

While it should, at this point, be quite obvious that this method of realizing a digital filter is only useful for mechanizing a non-recursive filter with an additional delay determined by $N-M$, it is also true that this method is the fastest of all methods for implementing them when M is greater than a "break-even" point.

To compute the value of $N = N_{opt}$ that minimizes the number of computations a few simple relations can be used.

Computation of the discrete Fourier transform using the "fast" algorithm (Section 14) requires $2 N \log_2 N$ real multiplications when N is taken as a power of 2. Real additions and subtractions require $3 N \log_2 N$ operations.

With these parameters the computation time for the two methods can be estimated though the methods differ in details. Two transforms, the transform of $\bar{x}_m$ and the inverse transform of $A_n X_n$, must be taken for each sub-sequence of $N - M + 1$ points. Therefore, the required number of transform applications is approximately.

$$\frac{2 R}{N - M + 1} \quad (3.7.5.3-1)$$

and the total computing time is

$$TC = \frac{2 (2 \alpha + 3 \beta) R N \log_2 N}{N - M + 1} \text{ seconds,} \quad (3.7.5.3-2)$$

where

R = total number of data points filtered
 α = time per multiplication in seconds,
 β = time per addition/subtraction in seconds.

Helms determined the number $N = N_{opt}$ that minimized the computing time TC . It is found from table 3.7.5.3-1 that the minimum value is approximately given by

$$N_{opt} \approx M \log_2 M. \quad (3.7.5.3-3)$$

The minimum computing time is also found approximately from the table to be

$$TC_{opt} \approx (6 \alpha + 9 \beta) R \log_2 M \quad (3.7.5.3-4)$$

M	N_{opt}	$\log_2 N_{opt}$	$\frac{2N_{opt} \log_2 N_{opt}}{N_{opt} - M + 1}$
< 11	32	5	< 13.9
11-17	64	6	14.22-16
18-29	128	7	16-18
30-52	256	8	18-20
53-94	512	9	20-22
95-171	1024	10	22-24
172-310	2048	11	24-26
310-575	4096	12	26-28
575-1050	8192	13	28-30
1050-2000	16384	14	30-32
2000-3800	32768	15	32-34
3800-7400	65536	16	34-36
> 7400	131072	17	≥ 36

Table 3.7.5.3-1 N_{opt} as a Function of M (assuming N_{opt} a power of 2)

From the table it is also obvious that TC_{opt} is not a sensitive function of N.

3. 5.4 Reduction of Operations for Real Filters and Sequences

When the non-recursive filter and the sequence being filtered are real than it is quite easy to reduce the number of computations by a factor of two by either of the methods of the previous section.

In both methods the sequence $\bar{x}_m$ is progressively replaced with $N-M+1$ values of the sequence being filtered, x_m . If x_m is real successive sequences can be equated to the real and imaginary parts of x_m and as long as the $\bar{x}_m$ are real the result of the convolution will be the two filtered sequences y_j as the real and imaginary parts. Linearity of a complex number under multiplication by a real constant allows the result.

Therefore, both the "select-saving" and "overlap-adding" methods may be used with two successive filtering operations of $N - M + 1$ points performed with the single application of the DFTI and the IDFTI.

3.7.5.5 Comparison with the Direct Programming Method

It is of considerable interest to compare the number of computations using the fast Fourier transform realization with that of the direct programming method of Section 3.7.1. For the direct method each filter output requires $4M$ multiplications and $4M$ addition/subtractions for complex weights and data. Real weights and data require one fourth as much computation, but as the fast Fourier transform operates on two sets of data points simultaneously (Section 3.7.5.4) for real quantities a comparison with complex quantities is valid, identically.

For R input data points the total computing time, TC_D , for filtering by direct programming is approximately

$$TC_D = 4 (\alpha + \beta) MR \quad (3.7.5.5-1)$$

with quantities as defined previously. The computational efficiency ratio of the fast Fourier transform algorithm to the direct programming method is

$$E = \frac{TC_D}{TC_{opt}} \quad (3.7.5.5-2)$$

$$= \frac{4 (\alpha + \beta) M (N_{opt} - M + 1)}{2 (2\alpha + 3\beta) N_{opt} \log_2 N_{opt}}$$

from (3.7.5.3-2) for optimum N. Using the approximation (3.7.5.3-3)

$$E \simeq \frac{4 (\alpha + \beta) M}{3 (2\alpha + 3\beta) \log_2 M} \quad (3.7.5.5-3)$$

If the multiplication and addition/subtraction times are assumed approximately equal ($\alpha = \beta$), for comparison purposes,

$$E \simeq \frac{8}{15} \frac{M}{\log_2 M} \quad (3.7.5.5-4)$$

This approximation is useful, however, only when M is large as examination of table 3.7.5.3-1 will indicate. Careful examination will show that the "break-even" point for using the fast Fourier transform is between M = 18 to 29 when $\alpha = \beta$. Thereafter rapid gains may be expected. By varying the multiplication time slightly with respect to the addition/subtraction time the break-even time may be varied greatly. M = 20 is the approximate break-even point for equality.

3.8 REALIZATION OF SOME NON-RECURSIVE FILTERS AS RECURSIVE FILTERS

Oddly enough, certain non-recursive filters can be realized as recursive filters. For example, the filter:

$$y_k = \sum_{j=0}^N \frac{1}{N+1} x_{k-j} \quad (3.8-1)$$

which is recognized as a simple averaging over a span of $N + 1$ points from the most recent data point back, has the weighting sequence

$$w_j = \begin{cases} \frac{1}{N+1}, & 0 \leq j \leq N, \\ 0, & j > N. \end{cases} \quad (3.8-2)$$

The z transfer function is, therefore:

$$\begin{aligned} H_z(z) &= \sum_{j=0}^N \frac{z^{-j}}{N+1}, \\ &= \frac{\frac{1}{N+1} - \frac{1}{N+1} z^{-(N+1)}}{1 - z^{-1}} \end{aligned} \quad (3.8-3)$$

The equivalent difference equation may take the form (Section 3.1)

$$y_k = y_{k-1} + \frac{1}{N+1} x_k - \frac{1}{N+1} x_{k-(N+1)} \quad (3.8-4a)$$

Note that $N + 2$ values of past history must be kept in storage as successive y_k 's are computed.

Actually, this result is obvious without recourse to z transforms and associated paraphernalia. It simply states that the most recent running average can be formed by taking the last running average (y_{k-1}), subtracting the oldest point divided by the number of points being averaged, and adding the newest, similarly weighted. Computing the running average by hand, or desk calculator, shows this result rather quickly; it would be simpler to perform, 1 addition; 1 subtraction and two divisions (multiplications), rather than $N + 1$ additions and 1 division (multiplication), when N is large. It is necessary to make sure that $y_0 = 0$; or any non-zero value will be carried as an error when computing further y_k .

Other non-recursive filters of $N + 1$ weights can be realized as recursive filters as long as

$$\sum_{k=0}^N w_k z^{-k}$$

is expressible as the ratio of two polynomials in z^{-1} . Care must be taken with initial values in recursive realizations as initial and/or round-off errors can propagate and/or grow with time. This occurs because the denominator of the recursive realization often has a zero on the unit circle which can cause instability. Infinite precision arithmetic and care in initial values can avoid these problems as the exact mathematical formulation has no poles. As there is seldom infinite precision arithmetic, error effects in recursive realizations of basically non-recursive filters (finite number of members in the impulse response sequence) should always be analyzed.

A technique to assist in minimizing finite precision arithmetic instability is the moving of "apparent" poles, of non-recursive filter realization by recursive filters, inside the unit circle. The adjective "apparent" is used because the recursive realization has no poles (or the impulse response sequence would have an infinite number of members). Denominator terms of the form; for example, $1 - z^{-1}$ as in (3.8-3), actually can be divided into the numerator, giving the original non-recursive filter expression, with a finite number of terms.

The instability is caused by the homogeneous solution to the difference equation with which the filter is implemented. For example, the running average recursive realization can be written

$$y_k - y_{k-1} = \frac{1}{N+1} x_k - \frac{1}{N+1} x_{k-(N+1)}. \quad (3.8-4b)$$

Regardless of the input sequence, x_k , a solution for y_k is the solution to

$$y_k - y_{k-1} = 0.$$

As with differential equations there are "homogeneous" solutions which, due to the linearity of the difference equation can be added to the solution for a particular input sequence x_k . These homogeneous solutions are "transients". For the example above, a solution $y_k = C\beta^k$ is assumed and it is found that $\beta = 1$. Therefore, the homogeneous solution is

$$\begin{aligned} y_k &= C\beta^k \\ &= C \end{aligned} \quad (3.8-5)$$

If the process (filter) is started with an error C it is retained. Quantization errors and round-off errors can contribute to C , as if the filter were being re-started at some time other than zero. By making β less than 1, the error will tend to diminish with time. Setting

$$\beta = 1 - 2^{-n},$$

where 2^{-n} is a few more bits significant than the precision of the arithmetic used, will cause the transient error to decay faster than it is generated. For example, if the machine has 32 bits of significance, then $n = 22$ to 28 will help keep errors under control.

The modified recursive filter is; then,

$$y_k = (1 - 2^{-n}) y_{k-1} + \frac{1}{N+1} x_k - \frac{1}{N+1} x_{k-(N+1)}. \quad (3.8-6)$$

For modest running averages ($N < 5000$) this modification hardly affects the frequency response.

3.9 "LEGAL" FILTER FORMS

From Section 3.2 the general filter expression was determined, and Section 3.4 where the impulse response was defined and discussed, it was found that only certain forms for digital filters are legitimate.

Further discussion is warranted to clarify the restrictions on the filter expression.

Writing the transfer function in terms of positive powers of z produces

$$H_z(z) = \frac{\sum_{m=0}^M a_m z^{N-m}}{\sum_{n=0}^N b_n z^{N-n}} \quad (3.9-1)$$

To ensure that only negative (and zero) powers of z are involved in the expansion of $H_z(z)$, as physically required (Section 3.4), b_0 cannot be equated to zero if a_0 is not zero. Having both a_0 and b_0 zero is legitimate as this would be equivalent to having a digital filter with a numerator of the order $M-1$ and a denominator of order $N-1$. Therefore assuming $b_0 = 1$ is not restrictive. Finally, any of the a_m 's may be zero, though there should be at least one non-zero a_m so that the filter transfer function is not trivially zero.

An example of a legitimate, or "legal", non-recursive transfer function is

$$H_z(z) = \sum_{m=0}^M a_m z^{-m}, \quad (3.9-2)$$

with the corresponding filter

$$y_j = \sum_{m=0}^M a_m x_{j-m}. \quad (3.9-3)$$

An "illegal" non-recursive filter transfer function is

$$H_z(z) = z \sum_{m=0}^M a_m z^{-m}, \quad (3.9-4)$$

as $b_0 = 0$, $b_1 = 1$. The equivalent "illegal" filter is

$$y_{j-1} = \sum_{m=0}^M a_m x_{j-m}. \quad (3.9-5)$$

While equation (3.9-5) is a perfectly "computable" expression within itself, it implies a one unit time delay as it stands, contrary to the definition of a transfer function expressing the relation between input and output sequences at the same instant (or index).

Recursive filters, by the above arguments, must have the form

$$y_j + \sum_{n=1}^N b_n y_{j-n} = \sum_{m=0}^M a_m x_{j-m} \quad (3.9-6)$$

where any a_m or b_n ($b_0 = 1$) can be zero as long as stability is preserved.

3.10 BANDWIDTH AND TRANSIENT RESPONSE TIME, OR BANDWIDTH OF A SAMPLED PULSE WAVEFORM

A general measure of impulse decay time can be obtained. Equivalently, a relation between a digital waveshape pulse width and its spectrum is obtained. A general measure is needed to handle the somewhat arbitrarily shaped filters and pulses which occur in practice.

Assume that a low-pass filter, with a bandwidth relatively small compared to half the sampling frequency is impulsed. Equivalently, assume that a digital pulse waveform is z transformed.

In either case, a filter, or spectrum, effective bandwidth, α is defined

$$\left[\frac{\alpha}{2} \right]^2 = \frac{1}{\omega_s} \int_{-\omega_s/2}^{\omega_s/2} \omega^2 |H_z(\omega)|^2 d\omega \quad (3.10-1)$$

where

$H_z(\omega)$ = the digital filter transfer function or
the spectrum of the sampled pulse waveform.

The filter or spectrum is assumed to be normalized, for convenience, such that:

$$\frac{1}{\omega_s} \int_{-\omega_s/2}^{\omega_s/2} |H_z(\omega)|^2 d\omega = 1 \quad (3.10-2)$$

This normalization can be performed for any filter or spectrum.

Conversely, this normalization is equivalent to the time domain relationship, or normalization;

$$\sum_{k=0}^{\infty} w_k^2 = 1. \quad (3.10-3)$$

The impulse response, or pulse waveform is such that the sum of the square of the sample values is unity. To prove (3.10-3) note that

$$H_z(\omega) = \sum_{k=0}^{\infty} w_k e^{-i\omega kT}, \quad \omega_s = \frac{2\pi}{T}; \quad (3.10-4)$$

for stable filters and

$$\frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} e^{-i\omega(k-j)T} d\omega = \begin{cases} 1, & k=j \\ 0, & \text{otherwise.} \end{cases} \quad (3.10-5)$$

Next define an effective time waveform width, β ,

$$\left[\frac{\beta^2}{2} \right] = \sum_{k=0}^{\infty} (kT)^2 w_k^2 \quad (3.10-6)$$

It can be seen that both the effective bandwidth and effective waveform width are r.m.s. values defined with respect to positive weight functions, $\frac{1}{\omega_s} |H_z(\omega)|^2$ and w_k^2 , consecutively, which integrate or sum to unity as do probability density functions.

The relationship to be proven is that,

$$\alpha\beta \leq 4. \quad (3.10-7)$$

Wide filter bandwidths or spectra, correspond to short time waveforms, or conversely.

First, the expression for β^2 must be converted to one involving $H_z(\omega)$. As

$$H_z(\omega) = \sum_{k=0}^{\infty} w_k e^{-i\omega kT}$$

$$\frac{\partial H_z(\omega)}{\partial \omega} = -i \sum_{k=0}^{\infty} (kT) w_k e^{-i\omega kT} \quad (3.10-8a)$$

and,

$$\frac{\partial H_z^*(\omega)}{\partial \omega} = i \sum_{j=0}^{\infty} (jT) w_j e^{i\omega jT} \quad (3.10-8b)$$

Therefore, it is found that

$$\frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} \left| \frac{\partial H_z(\omega)}{\partial \omega} \right|^2 d\omega = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} kTjTw_k w_j \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} e^{i\omega(k-j)T} d\omega \quad (3.10-9)$$

$$= \sum_{k=0}^{\infty} (kT)^2 w_k^2$$

The product $\alpha^2 \beta^2$ can then be written:

$$\frac{\alpha^2 \beta^2}{16} = \frac{1}{\omega_s} \int_{-\omega_s/2}^{\omega_s/2} \omega^2 |H_z(\omega)|^2 d\omega \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} \left| \frac{\partial H_z(\omega)}{\partial \omega} \right|^2 d\omega \quad (3.10-10)$$

From Schwarz' inequality

$$\left[\int_a^b |f|^2 dx \int_a^b |g|^2 dx \right]^{1/2} \geq 2 \left| \operatorname{Re} \left[\int_a^b f g^* dx \right] \right| \quad (3.10-11)$$

the relation

$$\frac{\alpha \beta}{4} \geq \left| \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} \left[\omega H_z(\omega) \frac{\partial H_z^*(\omega)}{\partial \omega} + H_z^*(\omega) \frac{\partial H_z(\omega)}{\partial \omega} \right] d\omega \right| \quad (3.10-12)$$

can be written.

The integral on the right may be recognized as

$$\frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} \omega \frac{\partial}{\partial \omega} |H_z(\omega)|^2 d\omega \quad (3.10-13)$$

which can be integrated by parts:

$$\frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} \omega \frac{\partial}{\partial \omega} |H_z(\omega)|^2 d\omega = -1 \quad (3.10-14)$$

from equation (3.10-8) and assuming that $H_z(\omega_s/2) \approx 0$. As was to be shown:

$$\alpha\beta \geq 4.$$

The relation can be generalized to include filters, or spectra, that are peaked about some average frequency, $\bar{\omega}$, and time waveforms that have an average time, τ , of occurrence where $\bar{\omega}$, and τ are defined by usual methods:

$$\bar{\omega} = \frac{1}{\omega_s/2} \int_0^{\omega_s/2} \omega |H_z(\omega)|^2 d\omega, \quad (3.10-15)$$

and

$$\tau = \sum_{k=0}^{\infty} kT w_k^2 \quad (3.10-16)$$

Equivalent bandwidths and pulse widths are defined, then, by

$$\left[\frac{\omega}{2}\right]^2 = \frac{1}{\omega_s/2} \int_0^{\omega_s/2} (\omega - \bar{\omega})^2 |H(\omega)|^2 d\omega, \quad (3.10-17)$$

and

$$\left[\frac{\theta}{2}\right]^2 = \sum_{k=0}^{\infty} (kT - \tau)^2 w_k^2. \quad (3.10-18)$$

The proof that $\alpha\beta \geq 4$ can still be performed, but it is much more complicated and requires that

$$H(0) \approx H\left(\frac{\omega_s}{2}\right), \quad (3.10-19)$$

$$\approx 0.$$

This is not unreasonable as this proof is for filters, or spectra, that have significant values only at some frequency not equal to zero or $\omega_s/2$.

If the effective bandwidth, β , is converted to Hertz, or cycles per second: $\beta = 2\pi f_\Delta$ then

$$\alpha\beta \geq 4 \quad (3.10-20)$$

becomes

$$\alpha f_\Delta \geq \frac{2}{\pi};$$

the product of effective pulse width and effective spectrum width is greater than or equal to $2/\pi$. A sampled pulse with an envelope of 1/100 of a second must have a spectrum width

$$\begin{aligned} f_\Delta &\geq \frac{1}{\alpha} \frac{2}{\pi}, \\ &= \frac{2}{\pi} \times 100 \text{ Hertz.} \end{aligned}$$

To obtain an intuitive idea of the transient decay time evaluate the impulse response for a perfect rectangular filter (albeit unrealizable):

$$H(\omega) = \begin{cases} 1, & \left| \omega_0 - \frac{\Delta}{2} \right| \leq |\omega| \leq \left| \omega_0 + \frac{\Delta}{2} \right|, \\ 0, & \text{otherwise;} \end{cases} \quad (3.10-21)$$

$$\text{with } \omega_0 + \frac{\Delta}{2} \leq \frac{\omega_s}{2}.$$

The impulse response sequence is given (using equation 3.4-3) by

$$w_k = \frac{1}{\omega_s} \left[\int_{\omega_0 - \frac{\Delta}{2}}^{\omega_0 + \frac{\Delta}{2}} e^{i\omega kT} d\omega + \int_{-\omega_0 - \frac{\Delta}{2}}^{-\omega_0 + \frac{\Delta}{2}} e^{i\omega kT} d\omega \right], \quad (3.10-22a)$$

or

$$w_k = \frac{2\Delta}{\omega_s} \left[\frac{\sin\left(\frac{\Delta Tk}{2}\right)}{\left(\frac{\Delta Tk}{2}\right)} \right] \cos \omega_0 kT \quad (3.10-22b)$$

The familiar $\sin x/x$, $x = \Delta Tk_0/2$, term that the rectangular filters generate is found along with a cosinusoidal "ringing" at the center frequency of the filter. The amplitude of the impulse sequence is less than one ($2\Delta/\omega_s < 1$). The $\sin x/x$ term is substantially down to 10 percent of its value at about $\Delta Tk_0/2 = \pi$ or where

$$\tau \Delta = 2\pi; \quad \tau = k_0 T. \quad (3.10-23)$$

The product of filter bandwidth, Δ , and impulse decay time, τ , is equal to a constant 2π which differs only slightly from the general "effective bandwidth-transient decay time" (3.10-7). . A narrow bandwidth gives a long decay time. Concurrently, if the sampling time is short, many sampling intervals, k_0 , are necessary before the transient decays. If a digital computer is used, computer time and, therefore, cost is large for a narrow bandwidth as many computations are necessary before the filter arrives at its "steady state" conditions, if the steady state operation is of interest.

3.11 THE DISCRETE FOURIER TRANSFORM AS A DIGITAL FILTER

The discrete Fourier transform is a linear, complex coefficient, non-recursive filtering operation. That is, if the discrete Fourier transform is taken at a particular frequency, then input sequences that deviate from this frequency can contribute to the value of the output at this frequency via a transfer function. To see this compute the transform at a frequency $\omega_j = j \omega_s/N$ with an input sequence at an arbitrary frequency ω .

Let

$$v_k = e^{i\omega_k T} \quad (3.11-1)$$

then the discrete transform at frequency ω_j of v_k is

$$\begin{aligned} V_j(\omega) &= \sum_{k=0}^{N-1} e^{i(\omega - \omega_j)kT}, \quad T = \text{sampling interval;} \\ &= \frac{1 - e^{i(\omega - \omega_j)NT}}{1 - e^{i(\omega - \omega_j)T}}, \\ &= e^{i(\omega - \omega_j) \frac{(N-1)T}{2}} \cdot \frac{\sin\left[(\omega - \omega_j) \frac{NT}{2}\right]}{\sin\left[(\omega - \omega_j) \frac{T}{2}\right]} \end{aligned} \quad (3.11-2)$$

which is the transfer function of the Fourier coefficient computed at ω_j .

Now

$$\begin{aligned} \omega_j &= \frac{j\omega_s}{N} \\ &= \frac{j2\pi}{NT} \end{aligned}$$

Therefore, it can be seen that inputs at $\omega = \omega_k (k \neq j)$ have zero output. At $\omega = \omega_j$, $V_j(\omega) = N$. A typical plot of $V_j(\omega)$ is seen in Figure 3.11-1.

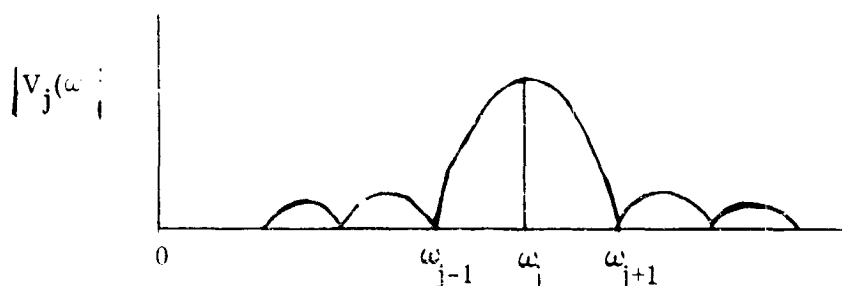


Figure 3.11-1 Response of a Discrete Fourier Transform

If the response of another coefficient is computed, for example $V_{j-1}(\omega)$ then its maximum occurs at the first zero of $V_j(\omega)$ to the left of the maximum of $V_j(\omega)$.

It should also be noted that $V_j(\omega)$ falls off approximately as $1/(\omega T/2)$ away from the center with the exception of the sharp zeros at harmonics of $\Delta \omega = \omega_s/N$.

3.12 A USEFUL SUMMATION RELATION

A quite useful relation involves the sum of a product of a "lagged sequence" with itself. When the "lag" is zero the sum particularizes to the sum of the square of the sequence. In many applications this relation allows "simple" evaluation by residues.

It will be shown that, if $x_k = 0$ for $k < 0$, then

$$\sum_{k=0}^{\infty} x_k^* x_{k+\tau} = \frac{1}{2\pi i} \oint_{\Gamma} X(z) X^*(z) z^{\tau} \frac{dz}{z} \quad (3.12-1)$$

where the path of integration is the unit circle, or in frequency form

$$= \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} X(\omega) X^*(\omega) e^{i\omega\tau T} d\omega$$

with $X(\omega)$ an obvious abridgement of $X(z)$, the z transform of x_k

If the sequence x_k is real then

$$X^*(z) = X\left(\frac{1}{z}\right)$$

on the unit circle and

$$\sum_{k=0}^{\infty} x_k x_{k+\tau} = \frac{1}{2\pi i} \oint X(z) X\left(\frac{1}{z}\right) \frac{dz}{z}$$

The proof is relatively simple and involves the orthogonality relations:

$$\frac{1}{2\pi i} \oint \frac{dz}{z^{j-k+1}} = \delta_{jk} \quad (3.12-2a)$$

with the contour enclosing the origin or

$$\frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} e^{i\omega(j-k)T} d\omega = \delta_{jk} \quad (3.12-2b)$$

Computing the integral using the transform of x_k

$$\frac{1}{2\pi i} \oint X(z) X^*(z) z^{\tau} \frac{dz}{z} = \sum_{j,k=0}^{\infty} x_j x_k^* \frac{1}{2\pi i} \oint \frac{dz}{z^{j-\tau-k}}$$

$$= \sum_{j,k=0}^{\infty} x_j x_k^* \delta_{j, \tau+k} \quad (3.12-3)$$

$$= \sum_{k=0}^{\infty} x_k^* x_{k+\tau}$$

as was to be shown. The frequency representation, integration along the unit circle, follows easily.

An example of interest involves the impulse response of a stable digital filter, in which case $x_k = w_k$ and $X(z) = H_z(z)$, and the sum of the squares is of interest, namely,

$$\sum_{k=0}^{\infty} |w_k|^2 = \frac{1}{2\pi i} \oint H_z(z) H_z\left(\frac{1}{z}\right) \frac{dz}{z} \quad (3.12-4)$$

In this example assume the transfer function to be

$$H_z(z) = \frac{1}{1 - e^{-aT} z^{-1}} \quad (3.12-5)$$

and

$$\sum_{k=0}^{\infty} |w_k|^2 = \frac{1}{2\pi i} \oint \frac{1}{1 - e^{-aT} z^{-1}} \cdot \frac{1}{1 - e^{-aT} z} \frac{dz}{z} \quad (3.12-6)$$

$$= \frac{1}{2\pi i} \oint \frac{-e^{aT} dz}{(z - e^{-aT})(z - e^{aT})}$$

as this is a stable system only the pole within the unit circle contributes to the evaluation by residues, therefore

$$\sum_{k=0}^{\infty} |w_k|^2 = \frac{1}{1 - e^{-2aT}} \quad (3.12-7)$$

4. NUMERICAL INTEGRATION OF LINEAR, CONSTANT COEFFICIENT INTEGRO-DIFFERENTIAL EQUATIONS

Digital Filters may be considered and designed as signal or data processing elements independently of analogue filter considerations. However, as considerable knowledge applicable to constant coefficient analogue filters is also useable for constant coefficient digital filter design it is desirable to know certain connections between the two domains. Also as digital filters are often used for numerical integration, or simulation, of analogue systems the connection is eminently worthwhile understanding.

Certain fundamental limitations in numerical integration, or simulation, are found. It will be seen that a successful numerical integration technique via a linear digital filter means equality of the analogue and digital transfer functions over the signal bandwidth of interest to a maximum frequency interval of zero to half the sampling frequency. A basic theorem to be proven shows that a general analogue transfer function cannot be synthesized by a digital filter without delay in the output of the digital filter. Therefore, a digital simulation inherently contains delay.

Pedagogically, it is usually easiest to fathom digital filter operation and usage through knowledge of counterparts in the analogue domain. Familiarity with both subjects blurs the distinction until they, in fact, cross-pollinate each other.

4.1 INTRODUCTION TO ESSENTIALS OF DESCRIPTION OF CONSTANT COEFFICIENT SYSTEMS

Linear, constant coefficient, differential (or integro-differential) equations describe a multitude of physical systems either exactly or approximately. Electrical circuits, dynamical systems, servo-mechanisms, and so on, often have either partial or complete descriptions in terms of these kinds of equations. To a degree of approximation linear equations with "slowly varying" coefficients may also be described within the framework of constant coefficient systems.

Historically these equations were solved by numerous techniques including the Fourier and Laplace transforms discussed previously. However, until the advent of the modern digital computer the primary interest was in finding solutions for "typical" idealized driving functions to obtain a physical "feeling" about the behavior of the system under discussion. "Closed form" solutions were obtained which involved the "Green's Function" of the mathematician and physicist or the same quantity by another name, the "impulse response" of the electrical engineer.

To solidify notation for further work pertinent constant coefficient, linear differential solution techniques are presented.

A linear, constant coefficient differential equation is defined by

$$\sum_{n=0}^N a_n \frac{d^n y}{dt^n} = \sum_{m=0}^N b_m \frac{d^m x}{dt^m} \quad (4.1-1)$$

where

y y(t), the system response; the "output",

x x(t), the system driving function; the "input".

$\frac{d^j u}{dt^j}$ = the j^{th} derivative of u with respect to time ($j = 0$ means the function u , itself),

q_j, p_j = constants determined by the system at hand.

Constant coefficient integro-differential equations can also be reduced to this form.

Assuming that the physical system is initially in the state at $t \leq 0$ where

$$\left. \frac{d^j y}{dt^j} \right|_{t \leq 0} = 0 \quad \text{for all } j,$$

then the Laplace (or Fourier) transform applied to both sides of equation (4.1-1) produces

$$Y(s) = H(s)X(s) \quad (4.1-2)$$

where

$$\begin{aligned} Y(s) &= \int_0^{\infty} y(t)e^{-st} dt \\ &= \text{Laplace transform of the output} \end{aligned}$$

$$\begin{aligned} X(s) &= \int_0^{\infty} x(t)e^{-st} dt \\ &= \text{Laplace transform of the input} \end{aligned}$$

and

$$H(s) = \frac{\sum_{m=0}^M p_m s^m}{\sum_{n=0}^M q_n s^n}$$

The relation

$$\int_0^{\infty} \frac{du(t)}{dt} e^{-st} dt = \frac{1}{s} \int_0^{\infty} u(t)e^{-st} dt \quad (4.1-3)$$

for zero initial conditions has been used repeatedly in obtaining (4.1-2). Non-zero initial conditions are discussed subsequently in section 6.4.

For physical systems, where an output is not to be expected before an input is applied, then it can be shown that the degree of the numerator in $H(s)$ must be less than or equal to that of the denominator: $M \leq N$. Furthermore for stability, bounded solutions, the poles of the denominator of $H(s)$ must lie in the left half of the "s" plane. $H(s)$ is frequently called the "system function" or "transfer function" by engineers, and the "eigenvalue" of the differential equation (4.1-1) by physicists and mathematicians.

From equation (2.2-7) it may be seen that $y(t)$ can be found by the Laplace inversion integral

$$\begin{aligned} y(t) &= \frac{1}{2\pi i} \int_{\gamma-i\infty}^{\gamma+i\infty} Y(s) e^{ist} ds, \\ &= \frac{1}{2\pi i} \int_{\gamma-i\infty}^{\gamma+i\infty} H(s) X(s) e^{ist} ds, \end{aligned} \quad (4.1-4)$$

where the integrand is assumed analytic to the right of γ .

This integral is of fundamental importance in the study of constant coefficient differential equations and through the use of complex variable techniques can be evaluated quite simply for many cases of analytical interest and a few cases of practical interest.*

If a sine wave (complex) $x(t) = Ae^{i\omega t}$ had been used for the driving function in (4.1-1) then the output would have been found to be

$$y(t) = H(i\omega) Ae^{i\omega t} \quad (4.1-5)$$

Complex sine waves are simply multiplied by the complex number $H(i\omega)$ when operated upon by the differential equation. $H(i\omega)$ as a function of ω is called the "frequency response" of the system described by the differential equation.

If the Fourier transform of equation (4.1-1) had been taken, then

$$Y(i\omega) = H(i\omega) X(i\omega) \quad (4.1-6)$$

and

$$y(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(i\omega) X(i\omega) e^{i\omega t} d\omega$$

where $x(t)$ is constrained to be Fourier transformable rather than Laplace transformable (see section 2.1), and zero initial conditions have been assumed for $t \leq 0$.

* A. Papoulis, The Fourier Integral and Its Application, McGraw-Hill

Defining the impulse response $\bar{w}(t)$ (Green's function) of the differential equation by the solution for an impulse input (equation 2.1.2-5), then this function is given by

$$\bar{w}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(i\omega) e^{i\omega t} d\omega \quad (4.1-7)$$

where $\bar{w}(t) = 0$ for $t < 0$, as implied by the condition for "causality" (no output before an input).

The solution can then be written, after some manipulation, as the "convolution" of the input with the impulse response (See Section 2.7.1):

$$y(t) = \int_{-\infty}^{\infty} \bar{w}(\tau) x(t - \tau) d\tau$$

or

$$y(t) = \int_{-\infty}^{\infty} \bar{w}(t - \tau) x(\tau) d\tau. \quad (4.1-8)$$

$\bar{w}(t)$ can be calculated analytically in closed form relatively easily.*

Not atypically this formula is used for numerical integration of the differential equation when $x(t)$ is not given analytically but as an empirical function whose sampled values are known. This situation would occur when using a digital computer. for example, as the computer must operate on discrete values rather than on a continuous record. Unfortunately, severe restrictions are found in using this formula for numerical calculations in the obvious form

$$y(kT) = T \sum_{j=0}^J \bar{w}(jT) x((k-j)T), \quad (4.1-9)$$

where J is some large number, and T is the sampling interval. Subsequently this problem will be discussed.

Henceforth, $H(\omega)$ will be written for $H(i\omega)$ for notational simplification whenever confusion will not arise.

4.2 NON-ZERO INITIAL CONDITIONS

Assuming that the initial conditions are zero is not a restrictive assumption as the solution of a linear, constant coefficient equation can be separated into the sum of the solutions for zero initial conditions and the "transient" solution depending on initial conditions.

Care must be taken, however, during numerical solution so that the "transient" solution which is computed analytically and sampled, once the initial conditions are given, is delayed in time corresponding to the solution for zero initial conditions which is solved numerically and therefore usually has a delay involved (See Section 7).

* IBID.

In general, the Laplace transform of the n^{th} derivative of a function $u(t)$ is given by

$$\int_0^{\infty} \frac{d^n u(t)}{dt^n} e^{-st} dt = \begin{cases} U(s), & n = 0 \\ s^n U(s) - \sum_{j=1}^n s^{(n-j)} u^{(j-1)}(0^+), & n > 0 \end{cases} \quad (4.2-1)$$

where

$$U(s) = \int_0^{\infty} u(t) e^{-st} dt$$

and

$$u^{(j)}(0^+) = \lim_{\epsilon \rightarrow 0} \left[\frac{d^j u(t)}{dt^j} \right]_{t=\epsilon}, \quad \epsilon > 0$$

Taking the Laplace transform of (4.1-1) results in this case

$$\begin{aligned} \left(\sum_{n=0}^N q_n s^n \right) Y(s) &= \sum_{n=1}^N q_n \sum_{j=1}^n s^{(n-j)} y^{(j-1)}(0^+) \\ &= \left(\sum_{n=0}^M q_n s^n \right) Y(s) = \sum_{m=1}^M p_m \sum_{k=1}^m s^{(m-k)} x^{(k-1)}(0^+) \end{aligned} \quad (4.2-2)$$

where non-zero initial conditions are allowed.

Therefore:

$$\begin{aligned} Y(s) &= H(s) X(s) \\ &= \frac{\sum_{n=1}^N q_n \sum_{j=1}^n s^{(n-j)} y^{(j-1)}(0^+) + \sum_{m=1}^M p_m \sum_{k=1}^m s^{(m-k)} x^{(k-1)}(0^+)}{Q(s)} \end{aligned} \quad (4.2-3)$$

where

$$Q(s) = \sum_{n=0}^N q_n s^n,$$

and

$$H(s) = \frac{\sum_{m=0}^M p_m s^m}{Q(s)}$$

The first term $H(s) X(s)$ is the Laplace transform of the solution for zero initial conditions discussed previously; the second term is the "transient" depending upon initial conditions at $t = 0^+$.

Obviously, due to linearity of the original differential equation, the solution for zero initial conditions, the inverse transform of $H(s) X(s)$, can be considered independently for numerical solutions.

The part of the solution corresponding to

$$F(s) = \frac{\sum_{n=1}^N q_n \sum_{j=1}^n s^{n-j} y^{(j-1)}(0^+) - \sum_{m=1}^M p_m \sum_{k=1}^m s^{m-k} x^{(k-1)}(0^+)}{Q(s)} \quad (4.2-4)$$

can be found quite simply by taking the inverse transform using familiar Laplace transform tables. However, the initial conditions are quite often not given directly; the derivatives $(x^{(j)}(0^+), y^{(j)}(0^+))$ must be estimated from samples of $x(t)$ and $y(t)$. The Gregory-Newton polynomial interpolation formula may be used.

Mechanization of the initial condition "transient" computation is deferred until section 6.4 after the basic problem of computing $y(t)$ for zero initial conditions is discussed.

4.3 CONVERSION OF A "HIGH PASS" TRANSFER FUNCTION TO A CONSTANT PLUS "LOW PASS" TRANSFER FUNCTION

A few words about the transfer function $H(s)$ are in order here. Subsequent analysis will show that transfer functions that approach zero as s (or $i\omega$) approaches infinity are of considerable interest when considering sampled data. The general form for $H(s)$,

$$H(s) = \frac{\sum_{m=0}^M p_m s^m}{\sum_{n=0}^N q_n s^n} ; \quad M \leq N \quad (4.1-2)$$

indicates that when $M < N$, $H(s)$ approaches zero for large s . However, for $M = N$

$$\lim_{s \rightarrow \infty} H(s) = \frac{p_N}{q_N} \quad (4.3-1)$$

which is not zero.

Fortunately, $H(s)$ may be arranged in the form

$$H(s) = \frac{p_N}{q_N} + H_1(s) \quad (4.3-2)$$

where

$$H_1(s) = \frac{\sum_{n=0}^{N-1} \left(p_n - \frac{p_N}{q_N} q_n \right) s^n}{\sum_{n=0}^N q_n s^n}$$

Another, equivalent, useful form of $H(s)$ is the partial fraction expansion of section 6.2.

In block diagram form the system is arranged in two equivalent forms (fig. 4.3-1).

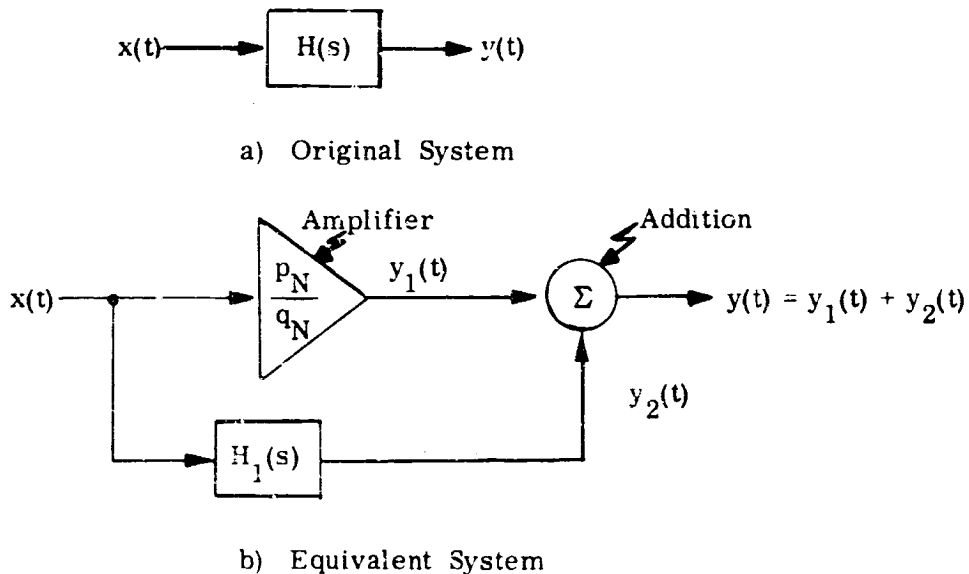


Figure 4.3-1 Two Equivalent Forms of a Linear System for $M = N$

In effect, $H(s)$ for $M = N$ can be considered as a network for $M < N$ (the response approaches zero as $H(\omega)$ becomes large) whose output is added to that of a simple amplifier of gain p_N/q_N . The input to both the network $H_1(s)$ and the amplifier is the driving function $x(t)$ and the system output is the summation of their outputs.

4.4 FILTERING OF A BANDLIMITED FUNCTION REPRESENTED IN TERMS OF SAMPLED VALUES

Having discussed the basic properties of linear, constant coefficient (or time invariant) differential equations, it is of considerable interest to find their effect upon an arbitrary bandlimited function represented in terms of samples at equi-spaced intervals of time. Operation upon a bandlimited function is of interest because it is not "damaged" by sampling and as a sequence of sampled data may be considered equivalent to a bandlimited function (section 2.1.1). At first assume for simplicity that the numerator order of $H(s)$ is less than that of the denominator. From previous work it was found that the Fourier transform of a bandlimited signal could be written as

$$X(\omega) = \begin{cases} T \sum_{n=-\infty}^{\infty} x(nT) e^{-i\omega nT}, & |\omega| \leq \omega_s/2 \\ 0, & |\omega| > \omega_s/2 \end{cases} \quad (2.1.1-7)$$

where the bandlimit of $x(t)$ is assumed, in this section, to extend to half the sampling frequency. $x(t)$ also has been set to equal for $t < 0$.

The output of the filter $H(\omega)$ was given in equation (4.1-6) as

$$y(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\omega) X(\omega) e^{i\omega t} d\omega \quad (4.1-6)$$

and the transform $Y(\omega)$ of $y(t)$ as

$$Y(\omega) = H(\omega) X(\omega)$$

It can be seen that if $X(\omega)$ is bandlimited, then the filter output has the same bandlimit. The transform can, therefore, be written as

$$Y(\omega) = \begin{cases} T \sum_{n=-\infty}^{\infty} y(nT) e^{-i\omega nT}, & |\omega| \leq \omega_s/2 \\ 0, & |\omega| > \omega_s/2 \end{cases} \quad (4.4-1)$$

where $y(nT)$ are the sampled values of the output. From expression (4.1-6) it is found that

$$y(t) = \frac{T}{2\pi} \sum_{n=-\infty}^{\infty} x(nT) \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H(\omega) e^{i\omega(t-nT)} d\omega \quad (4.4-2)$$

At a sampling instant $t = kT$, the output is given by

$$y(kT) = \sum_{n=-\infty}^{\infty} x(nT) \left[\frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H(\omega) e^{i\omega(k-n)T} d\omega \right] \quad (4.4-3)$$

The action of the filter at the sampled output times is completely specified by the quantities in the brackets. Defining these quantities, the Fourier expansion coefficients of $H(\omega)$ over $0 \leq |\omega| \leq \omega_s/2$, as

$$w^c(jT) = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H(\omega) e^{i\omega jT} d\omega \quad (4.4-4)$$

then

$$y(kT) = \sum_{n=-\infty}^{\infty} x(nT) w^c((k-n)T) \quad (4.4-5a)$$

or suppressing the functional notation

$$y_k = \sum_{n=-\infty}^{\infty} w_{k-n}^c x_n \quad (4.4-5b)$$

which is a fundamental formula for numerical integration.

It is important to note that the $w^c(jT)$ defined is not generally the sampled impulse response $\bar{w}(jT)$ (4.1-7), as $\bar{w}(t)$ requires integration over ω from $-\infty$ to $+\infty$ rather than from $-\omega_s/2$ to $\omega_s/2$. If the transfer function is bandlimited to the interval

$-\omega_s/2$ to $\omega_s/2$, the $w^c(jT)$ will be seen to be proportional to $\bar{w}(jT)$. Furthermore, the

$w^c(jT)$ are generally non-zero for $j < 0$.

If x_n is applied only at $t = 0$ ($x_0 = 1$), then $y_{-k} = w_{-k}^c$. The output is apparently non-zero at the sampling instants even if no input has been applied. This apparent contradiction is caused by the assumption of "bandlimiting" on $x(t)$. While $x(t)$ is zero at $t = -kT$, $k = 1, 2, \dots$; the function $x(t)$ is not generally zero between sampling instants as can be seen from equation 2.1.1-9. Therefore, output from the continuous filter for $t < 0$ is present at the sampling instants. Usually these outputs are quite small.

The relation between the quantities $w^c(jT)$ and the filter impulse response $\bar{w}(t)$ (4.1-7) is also of interest. The transfer function is given in terms of $\bar{w}(t)$ by

$$H(\omega) = \int_{-\infty}^{\infty} \bar{w}(t') e^{-i\omega t'} dt' \quad (4.4-6a)$$

or as $\bar{w}(t) = 0, t < 0$, for physically realizable transfer functions it can also be written:

$$H(\omega) = \int_0^{\infty} \bar{w}(t') e^{-i\omega t'} dt' \quad (4.4-6b)$$

Inserting (4.4-6) into (4.4-4) gives

$$\begin{aligned} w^c(jT) &= \int_{-\infty}^{\infty} dt' \bar{w}(t') \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} e^{-\omega(t' - jT)} d\omega \\ &= \int_{-\infty}^{\infty} dt' \bar{w}(t') \frac{\sin\left[\frac{\omega_s}{2}(t' - jT)\right]}{\left[\frac{\omega_s}{2}(t' - jT)\right]} \end{aligned} \quad (4.4-7)$$

If $\bar{w}(t')$ is "bandlimited" ($H(\omega) = 0, |\omega| > \omega_s/2$) then it can be expressed as

$$\bar{w}(t') = \sum_{n=-\infty}^{\infty} \bar{w}(nT) \frac{\sin\frac{\omega_s}{2}(t' - nT)}{\frac{\omega_s}{2}(t' - nT)} \quad (4.4-8)$$

and

$$w^c(jT) = \sum_{n=-\infty}^{\infty} \bar{w}(nT) T \delta_{n-j}; \delta_k = \begin{cases} 1, & k = 0, \\ 0, & k \neq 0 \end{cases} \quad (4.4-9)$$

as

$$\int_{-\infty}^{\infty} \frac{\sin x}{x} \frac{\sin (x + k\pi)}{x + k\pi} dx = \pi \delta_k \quad (4.4-10)$$

Finally,

$$w^c(jT) = T \bar{w}(jT) \quad (4.4-11)$$

as expected when $H(\omega)$ is zero for $|\omega| > \omega_s/2$. Further, $w^c(jT) = T \bar{w}(jT)$ is zero for $j < 0$ in this case, as it is assumed $H(\omega)$ is physically realizable. Actually, "physical realizability" and "bandlimiting" are contradictory ideas. No real linear network is actually bandlimited as it will have non-zero response to $\omega = \pm\infty$. However, in practice many transfer functions are essentially zero above some frequency.

It is not necessary that the sampled impulse response $\bar{w}(jT)$ equal the $w^c(jT)$'s, but if delay between the output sequence y_k and the input sequence x_j is to be minimized then the sampling frequency ω_s should be large so that $w^c(jT) \approx 0$ for j less than zero (see Section 4.5.2).

4.4.1 Extension to High Pass Transfer Functions

It would appear that high pass transfer functions are excluded from consideration as they do not approach zero for ω large. The work of Section 4.3 indicates that a high pass transfer function ($M = N$) can be written as

$$H(s) = \frac{p_N}{q_N} + H_1(s) \quad (4.3-2)$$

where $H_1(\omega)$ approaches zero for ω large. Therefore, this form is used to develop the effect of $H(s)$, ($H(\omega)$), on a bandlimited function.

As in equation (4.4-5) the result of $H(\omega)$ operating on a bandlimited function is

$$y(kT) = Ax(kT) + \sum_{n=-\infty}^{\infty} w_1^c((k-n)T) x(nT) \quad (4.4.1-1)$$

where

$$A = p_N/q_N$$

and

$$w_1^c(jT) = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H_1(\omega) e^{i\omega jT} d\omega$$

This equation may also be written

$$y(kT) = \sum_{n=-\infty}^{\infty} \left[A \delta_{kn} + w_1^c((k-n)T) \right] x(nT) \quad (4.4.1-2a)$$

or with functional notation suppressed

$$y_k = \sum_{n=-\infty}^{\infty} \left[A \delta_{kn} + w_{(1)k-n}^c \right] x_n. \quad (4.4.1-2b)$$

As shown previously, if $H_1(\omega)$ is bandlimited by $\omega_s/2$ then

$$w_{(1)j}^c = T \bar{w}_{(1)}(jT) \quad (4.4.1-3)$$

where $\bar{w}_{(1)}(jT)$ is the sampled impulse response of $H_1(s)$, ($H_1(\omega)$):

$$\bar{w}_{(1)}(jT) = \frac{i}{2\pi} \int_{-\infty}^{\infty} H_1(\omega) e^{i\omega jT} d\omega \quad (4.4.1-4)$$

4.5 DEFINITION OF SIMULATION OF A LINEAR CONSTANT COEFFICIENT SYSTEM

Simulation is defined here to mean the exact (or highly approximate) duplication of the simultaneously sampled input and output of a continuous system by processing on a digital computer. Computation time of the computer is considered as less than the sampling interval between input samples. As long as the correct output is available before the next input is applied then simulation has been achieved.

Simulation can be used either as a replacement for part of an operating physical system, where a digital computer filters a number stream that is taken from samples of a continuous function, or it can be used as an analytic simulation whereby the interest is in conceptually replacing a continuous system, or part of a system, by a program which accepts a sample of the input and generates an output before the next input is accepted. Obviously, if the simulated system is part of an operational system, then the successive outputs must be computed in "real-time;" that is, before the next input arrives. For analytic simulation it is only necessary that the simulated output be approximately equal to the continuous output in "simulated time." In diagrammatic form simulation is indicated by Figure 4.5-1.

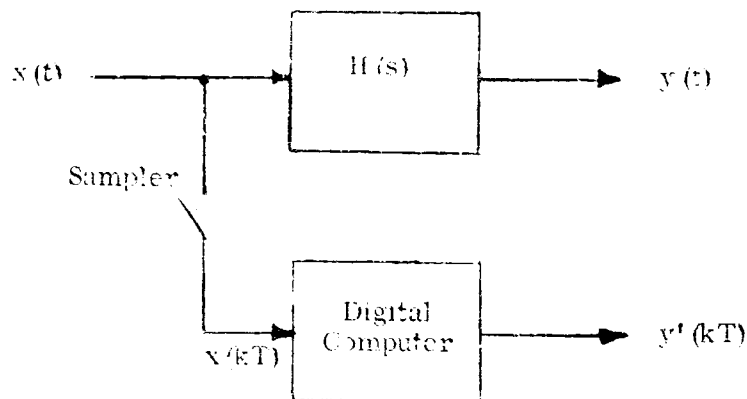


Figure 4.5-1. Simulation Illustrated

when simulation is achieved exactly then $y'(kT) = y(kT)$.

4.5.1 Simulation and Numerical Integration

Numerical integration is the process of solving a system of differential equations when the driving function is given by values taken at discrete instants. These instants are usually taken at equal spacing. When the numerical solution corresponding in time to the input driving function at that time can be obtained by the mathematical procedure used before the next input sample value then this procedure is termed simulation. In other words, numerical integration without time delay is simulation. Usually some delay, in terms of the number of sampling intervals, is necessary in any numerical integration scheme though it can be made small by choosing the sampling frequency sufficiently large. Simulation is then approximate, though in practice the delay can quite often be made unnoticeable.

Again it should be emphasized that the delay is inherent in the mathematics of numerical integration and not in the time to perform the computations.

4.5.2 Exact Simulation is Impossible in Principle Except for Bandlimited Transfer Functions.

It is a surprising fact that exact simulation of a linear constant coefficient system by a linear digital system is impossible unless the transfer function is bandlimited by half the sampling frequency. If delay in the output sequence $y(kT)$ with respect to the input sequence is tolerable then the sampling time is constrained only by the signal bandlimit.

A formal proof will be given demonstrating the general impossibility of exact simulation, or "real-time" sampled data filtering without delay as it points up pertinent aspects of numerical integration.

THEOREM: A "real-time," linear, constant coefficient digital filter (or numerical) integration procedure) cannot be found which produces an identical sampled output as the linear continuous system at the same instant, even upon assuming that the computing time is negligible, if the signal bandlimit extends to half the sampling frequency unless $H(s)$ is bandlimited by half the sampling frequency. Delay by the digital filter is the fundamental limit.

PROOF: First, assume that the numerator is of lower order than the denominator in the transfer function $H(s)$, without loss of generality (Section 4.3), so that high pass systems need not be considered here.

In Section 4.4 it was found that the sampled output of a continuous filter with a bandlimited input is

$$y_k = \sum_{n=-\infty}^{\infty} w_{k-n}^c x_n \quad (4.4-5)$$

with

$$w_j^c = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H(\omega) e^{i\omega jT} d\omega \quad (4.4-4)$$

It can be seen that for a given time $t = kT$ that all future inputs are required as w_j^c is not generally zero for $j < 0$ unless $H(\omega)$ is bandlimited by $\omega_s/2$. In this case (See equation 4.4-11)

$$w_j^c = T \bar{w}(jT) \quad (4.4-11)$$

where $\bar{w}(t)$ is the "impulse response" of $H(\omega)$.

Now for a stable system the w_j^c 's will be practically zero for some $j < -M$ (eq. 4.4-7). Therefore, the output can be written as

$$\begin{aligned} y_k &= \sum_{n=-\infty}^{\infty} w_{k-n}^c x_n \\ &= \sum_{m=-\infty}^{\infty} w_m^c x_{k-m} \\ &= \sum_{m=-M}^{\infty} w_m^c x_{k-m} \end{aligned} \quad (4.5.2-1)$$

A delay of MT seconds exists between the output and input sequences. Now equation (4.5.2-1) is an "illegal" expression for a digital filter as it expresses an output which depends on future inputs. See Section 5.1.

By defining $y'_{k+M} = y_k$ and adjusting indices in equation (4.5.2-1)

$$y_k = \sum_{n=0}^{\infty} w_{n-M} x_{k-n} \quad (4.5.2-2)$$

a realistic, in the sense that time relation is considered, computing expression has been obtained. The output y_k depends only on the present and past inputs. Obviously, the delay of MT seconds is still present by virtue of the redefinition of the output sequence. MT may be made as small as we like by taking T small. However, in principle simulation is impossible.

Though it has been shown that equation 4.4-5, when used for numerical integration or for simulation, of the differential equation described by $H(\omega)$ inherently contains delay, it has not been shown that some one of all possible linear digital filters cannot perform the "real-time" integration or simulation. This situation will now be remedied.

In terms of the filter impulse response the digital filter input-output relation can be written as

$$y_k = \sum_{n=-\infty}^{\infty} w_{k-n} x_n \quad (3.4.1-3)$$

As the function (sequence) being filtered is arbitrary it must be true that the "impulse response" of the digital filter is related to the numerical integration formula by

$$w_j = w_{j-M}^c \quad (4.5.2-3)$$

If identical outputs are desired. Therefore, there is only one linear digital filter ($w_j = w_{j-M}^c$) of all possible linear digital filters for numerical integration and it possesses MT seconds of delay.

The theorem is proved!

An analog transfer function bandlimited by $\omega_s/2$ is the only means of obtaining accurate simulation over the entire region $0 < |\omega| < \omega_s/2$ of the bandlimited input.

In Section 7, it will be shown that conditions on the digital filter can be relaxed considerably if the input data bandlimit is less than half the sampling frequency. It will then not be necessary that the transfer function $H(\omega)$ be bandlimited by $\omega_s/2$, though usually artificial means will be used to effectively bandlimit the given transfer function.

4.6 RELATION BETWEEN ANALOG AND DIGITAL TRANSFER FUNCTIONS FOR FOURIER COEFFICIENT INTEGRATION FORMULA

A question of interest is the relation between the analog transfer function via the Fourier coefficient numerical integration formula (4.4.1-2) and the corresponding digital transfer function. The case for $H(\omega)$ with $M < N$ will be considered for simplicity.

By definition, the digital transfer function is (with $z = e^{i\omega T}$)

$$H_z(\omega) = \sum_{j=0}^{\infty} w_j e^{-i\omega j T}. \quad (3.4-2)$$

For this particular integration formula the w_j 's are given by

$$w_j = w_{j-M}^c \quad (4.5.2-3)$$

$$= \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H(\omega') e^{i\omega'(j-M)T} d\omega' \quad (4.4-4)$$

and it follows that

$$H_z(\omega) = \sum_{j=0}^{\infty} \left[\frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H(\omega') e^{-i\omega' M T} e^{i\omega' j T} d\omega' \right] e^{-i\omega j T}. \quad (4.6-1)$$

Now M was chosen such that it was large enough to make $w_{j-M}^c \approx 0$; $j < 0$. Therefore, equation (4.6-1) can be recognized as the Fourier series expansion of $H(\omega)e^{-i\omega M T}$; that is

$$H_z(\omega) \approx H(\omega) e^{-i\omega M T}, \quad 0 < |\omega| < \omega_s/2. \quad (4.6-2)$$

The digital filter has precisely the same characteristics as the analog filter except for the additional phase term $e^{-j\omega MT}$ which corresponds to the aforementioned delay of MT seconds.

In practice, the upper limit of the summation must be finite; enough terms in the Fourier series are kept to maintain an "adequate" approximation. Section 5.6 discusses methods for reducing the number of terms (improving convergence) and the delay M . This technique is usually used only when the data bandwidth approaches half the sampling frequency. Recursive methods are preferable when higher sampling frequencies are allowable or the data bandwidth is "small." The more closely that the analog transfer function approaches bandlimiting, the smaller the delay MT becomes and less w_j^c 's are needed for $j < 0$.

5. NON-RECURSIVE FILTERS AND FILTER LIMITATIONS

Non-recursive filters, as mentioned previously, have no dependence on past outputs. The transfer function takes the form of a truncated one-sided Fourier series with only negative exponents. Synthesis of digital filters of this form will be seen to be closely akin to determining coefficients of Fourier series. In fact, subsequent discussion of Fourier series and conditions for convergence will show that arbitrary transfer functions cannot be synthesized without considerable delay, in general, irrespective of the form of the filter, recursive or non-recursive.

Independent of any general theoretical considerations non-recursive filters have a certain practical advantage besides being found naturally in some physical problems. These filters allow synthesis of arbitrary transfer functions satisfying mild regularity conditions arbitrarily well over the entire bandwidth from zero to half the sampling frequency albeit with delay generally determined by the accuracy of approximation.

It will be seen that non-recursive filter synthesis becomes a matter of curve fitting and that techniques used historically can be made available. Least squares curve fitting will be discussed along with exact fitting at arbitrarily spaced points using polynomial techniques. Convergence methods used upon Fourier series can reduce the number of coefficients in the filters and will be discussed.

Two major disadvantages of non-recursive filters are (1) the large number of coefficients, or weights, that are often needed entailing long computation time and/or complex hardware when implementing them, and (2) the lack of ability to draw upon the voluminous work on analog filter synthesis that recursive filters will be seen to allow.

5.1 REDUCTION OF THE GENERAL LINEAR DIGITAL FILTER FORM TO NON-RECURSIVE FORM.

Development here, reduces the general, linear digital filter transfer function form to a general "delayed" non-recursive form. The general filter form is

$$H_z(z) = \frac{\sum_{m=0}^M a_m z^{-m}}{\sum_{n=0}^N b_n z^{-n}}, \quad b_0 = 1 \quad (3.2-6)$$

Setting all the b_n to zero ($b_0 = 1$) produces (by definition) the general non-recursive filter input-output relation

$$H_z(z) = \sum_{m=0}^M a_m z^{-m}, \quad (5.1-1)$$

whose realization is

$$y_j = \sum_{m=0}^M a_m x_{j-m} \quad (5.1-2)$$

Obviously, the transfer function of the non-recursive filter when evaluated at $z = e^{i\omega T}$, is the frequency response function

$$H_z(\omega) = \sum_{m=0}^M a_m e^{-i\omega m T}, \quad (5.1-3)$$

where the notation is abridged as before, setting $H_z(e^{i\omega T}) = H_z(\omega)$.

It should also be noted that $H_z(\omega)$ is a partial Fourier series: that is, it contains a finite number of terms with negative exponents. The most general linear filter has an infinite number of terms, though only with negative exponents. As shall be seen, the fact that there are no positive terms in the Fourier series will not allow synthesis of an arbitrary transfer function without delay, which is true of linear, digital (and analog) filters in general. The finite number of terms in the Fourier series for non-recursive filters causes ripple in the approximations not generally found in recursive filters unless specifically desired.

Alternatively, the impulse response form of the transfer function

$$H_z(z) = \sum_{n=0}^{\infty} w_n z^{-n} \quad (3.4-2)$$

can be truncated to $M + 1$ terms to define the non-recursive filter.

Rearrangement of the transfer function into another form allows discussion of representation of a given arbitrary transfer function as a digital filter. This form also is used in transfer function fitting techniques for non-recursive filters.

Multiplication of the original transfer function by unity, $1 = z^{-\tau} z^{\tau}$, with τ an integer produces

$$\begin{aligned} H_z(z) &= z^{-\tau} \left[z^{\tau} \sum_{m=0}^M a_m z^{-m} \right] \\ &= z^{-\tau} \sum_{m=-\tau}^{M-\tau} a_{m+\tau} z^{-m} \end{aligned} \quad (5.1-4)$$

with a change of indices.

Defining the bracketed term

as

$$\begin{aligned} H_z^{(\tau)}(z) &= z^{\tau} \sum_{m=0}^M a_m z^{-m} \\ &= \sum_{m=-\tau}^{M-\tau} a_{m+\tau} z^{-m} \end{aligned} \quad (5.1-5)$$

then in another form

$$H_z(z) = z^{-\tau} H_z^{(\tau)}(z) \quad (5.1-6)$$

Corresponding to the "transfer function" $H_z^{(\tau)}(z)$ a time relationship can be written:

$$y_{j-\tau} = \sum_{m=0}^M a_m x_{j-m} \quad (5.1-7a)$$

or equivalently:

$$y_j = \sum_{m=0}^M a_m x_{j+\tau-m} \quad (5.1-7b)$$

This obviously (Section 3.4) is an unrealizable filtering operation as future data is required to produce the present (j^{th}) output. While the idea of time relationship between input and output is somewhat misplaced, the unrealizable transfer function is quite useful in approximations.

If $H_z^{(\tau)}(z)$ is used as an approximation to an arbitrary transfer function, $H(\omega)$, then a delay of τ will be involved in the actual filter as can be seen from eq. (5.1-6).

The desirability of $H_z^{(\tau)}(z)$ as the approximation for a given arbitrary transfer function follows from (5.1-5), where it can be seen that τ positive exponents are allowed. This produces a truncated Fourier series which can approximate an arbitrary transfer function under quite general conditions. These matters, and the choice of τ , will be discussed in Section 5.2 et al.

Hereafter the "transfer function" $H_z^{(\tau)}(z)$ will be called the "delayed" (by τ) transfer function for obvious reasons.

Certain properties of the "delayed" transfer function should be mentioned:

(a) Periodicity with period ω_s :

$$H_z^{(\tau)}(\omega) = H_z^{(\tau)}(\omega + \omega_s) \quad (5.1-8)$$

and

(b) Specification of real coefficient filters by positive frequencies:

$$\left[H_z^{(\tau)}(\omega) \right]^* = H_z^{(\tau)}(-\omega) \quad (5.1-9)$$

As a non-recursive filter is already in the impulse response form of the transfer function hereafter set $w_m = a_m$ and have (again)

$$H_z(z) = \sum_{m=0}^M w_m z^{-m}, \quad (5.1-10)$$

and

$$H_z^{(\tau)}(z) = z^{\tau} \sum_{m=0}^M w_m z^{-m} \quad (5.1-11)$$

$$= \sum_{m=-\tau}^{M-\tau} w_{m+\tau} z^{-m}$$

Finally, when the "delayed" transfer function is used for approximations with $\tau = N$, where the number of coefficients or "weights" is odd: $M = 2N + 1$, it is called a "central filter." This particular name takes on significance when expression (5.1-11) is examined. The summation is seen to be symmetrically, or "centrally," located within the limits $\pm N$.

5.1.1 "Cosine" and "Sine" Central Filter Simplifications

The particular case $\tau = N$, the "central filter," has a pair of simplified forms which allow easier computation, both from the point of view of weight computation and in speed of data processing. Central filters often have, however, more weights than necessary when approximating an arbitrary transfer function.

The first form is for "delayed" real transfer functions, "cosine" filters: As the real part of $H_z^{(\tau)}(\omega)$ is an even function it is easy to show that

$$w_{N-k} = w_{N+k}; \quad k = 0, 1, \dots, N; \quad (5.1.1-1)$$

the weights are symmetrical about the center one. Therefore, only $N+1$ weights must be computed. The filter speed can be increased by noting that the realization is

$$y_j = \sum_{n=-N}^N w_{N-n} x_{n+j-N} \quad (5.1.1-2)$$

$$= w_N x_{j-N} + \sum_{n=1}^N w_{N-n} (x_{n+j-N} + x_{-n+j-N}).$$

The number of multiplications is reduced by N . The filter, of course, delays automatically with no other phase shift to the center of the filtered data span.

Similarly, for $H_z^{(\tau)}(\omega)$ pure imaginary the filter is called a "sine" filter. The phase shift is automatically 90 degrees plus the delay to the midpoint. Also, the transfer function must be odd. It can easily be shown then, that

$$w_N = 0 \quad (5.1.1-3)$$

$$w_{N-k} = -w_{N+k}, \quad k = 1, 2, \dots, N,$$

which simplifies the weight calculations to N rather than $2N+1$.

The realizable filter reduces to

$$y_j = \sum_{n=1}^N w_{N-n} (x_n + j - N - x_{-n + j - N}) \quad (5.1.1-4)$$

with only N multiplications needed.

Developments in Section 5.8 illustrate usage of "cosine" and "sine" filters for implementing various bandpass filters.

5.2 FOURIER SERIES FUNDAMENTALS

Understanding of the basic elements of Fourier series is essential to an understanding of the limitations upon synthesis of a digital filter with an arbitrary transfer function. It will be seen that allowance of positive exponents in the "delayed" transfer function by the delay of τ sampling units is critical to synthesis.

Quite generally it will be proven in this section that an arbitrary transfer function with modest mathematical constraints can be synthesized arbitrarily well if sufficient delay is allowed. If the transfer function undergoing synthesis is analog and bandlimited by half the sampling frequency then no delay is necessary ($\tau = 0$) (Section 4.4).

The "delayed" non-recursive filter form (5.1-11) can be put in a slightly more general form

$$G(\omega T) = \sum_{n=-N_1}^{N_2} c_n e^{-i\omega n T} \quad (5.2-1)$$

where N_2 and N_1 can be related to M and τ . The general filter transfer function (5.1-10) is simply $N_1 = 0$ above. Therefore, discussion of this form and its properties is pertinent to our synthesis problem.

After discussing the properties of (5.2-1) its coefficients c_n will be related to the w_n of the non-recursive filters (Section 5.3.2).

The filter form has the variable ωT as an argument where $-\pi \leq \omega T \leq \pi$. To simplify matters let ω be normalized so that

$$\bar{\omega} = \omega T, \quad -\pi \leq \bar{\omega} \leq \pi$$

and have

$$G(\bar{\omega}) = \sum_{n=-N_1}^{N_2} c_n e^{-i\bar{\omega}n} \quad (5.2-2)$$

The bar on ω is superfluous and is dropped in further discourse. Effectively, the frequency axis has been scaled so that the sampling interval is unity and angular frequency ranges between minus and plus π .

The basic theorem on Fourier series states that if both the real and imaginary parts of $G(\omega)$ satisfy the conditions (the Dirichlet conditions):

- a. $G(\omega)$ is defined at every point of the interval $-\pi \leq \omega \leq \pi$;
- b. $G(\omega)$ is everywhere single-valued, finite, and sectionally continuous; $G(\omega)$ can have a finite number of discontinuities only, and two consecutive discontinuities must be separated by a finite interval;
- c. $G(\omega)$ is of "bounded variation," that is, it has a finite number of maxima and minima;

then, $G(\omega)$ can be expanded into a convergent infinite series which converges uniformly except at a point of discontinuity.

$$G(\omega) = \lim_{\substack{N_1 \rightarrow \infty \\ N_2 \rightarrow \infty}} \sum_{n=-N_1}^{N_2} c_n e^{-i\omega n} \quad (5.2-3)$$

where

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(\omega) e^{i\omega n} d\omega$$

or equivalently;

$$G(\omega) = a_0 + \lim_{\bar{N} \rightarrow \infty} \sum_{n=1}^{\bar{N}} (a_n \cos \omega n - i b_n \sin \omega n) \quad (5.2-4)$$

with

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(\omega) d\omega,$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} G(\omega) \cos n\omega d\omega; \quad n = 1, 2, \dots,$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} G(\omega) \sin \omega n d\omega; \quad n = 1, 2, \dots,$$

At a point of discontinuity $\omega = \omega_0$, the series converges to the arithmetic mean of the two limiting ordinates:

$$G(\omega_0) = \frac{1}{2} \left[G(\omega_0^+) + G(\omega_0^-) \right] \quad (5.2-5)$$

The coefficients a_n and b_n can be obtained through the well known "least squares" method (Section 5.2.1). Alternatively it can be assumed that the series (5.2-3) exist with coefficients c_n , a_n , and b_n as given, and investigate convergence and sufficient conditions upon $G(\omega)$ to ensure this convergence, (the Dirichlet conditions).

The coefficients of the two representations (5.2-3) and (5.2-4) are related by

$$a_0 = c_0,$$

$$a_n = c_n + c_{-n}; \quad n > 0, \quad (5.2-6a)$$

$$b_n = c_n - c_{-n}; \quad n > 0;$$

and

$$\begin{aligned} c_0 &= a_0, \\ c_n &= (a_n + b_n)/2; \quad n > 0, \\ c_{-n} &= (a_n - b_n)/2; \quad n > 0. \end{aligned} \quad (5.2-6b)$$

If, for a moment, the truncated Fourier series is written in exponential form with both indices equal:

$$G^N(\omega) = \sum_{n=-N}^N c_n e^{-i\omega n} \quad (5.2-7)$$

where the superscript N on $G^N(\omega)$ indicates the truncation at $n = \pm N$ and also (as before):

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(\omega') e^{i\omega' n} d\omega' \quad (5.2-8)$$

the c_n can be inserted back into the expression for $G^N(\omega)$. Interchanging the orders of integration and summation:

$$G^N(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(\omega') \left[\sum_{n=-N}^N e^{i(\omega - \omega') n} \right] d\omega' \quad (5.2-9)$$

or

$$G^N(\omega) = \int_{-\pi}^{\pi} G(\omega') \frac{\sin \left[\left(N + \frac{1}{2} \right) (\omega - \omega') \right]}{2\pi [\sin (\omega - \omega')/2]} d\omega' \quad (5.2-10)$$

It can be shown that the so-called "kernel" (the "Dirichlet kernel")

$$K_N(\omega - \omega') = \frac{\sin \left[\left(N + \frac{1}{2} \right) (\omega - \omega') \right]}{2\pi \sin (\omega - \omega')/2} \quad (5.2-11)$$

has the property that

$$\begin{aligned} \lim_{N \rightarrow \infty} G^N(\omega) &= \lim_{N \rightarrow \infty} \int_{-\pi}^{\pi} G(\omega') K_N(\omega - \omega') d\omega' \\ &= G(\omega); \end{aligned} \quad (5.2-12)$$

if the Dirichlet conditions are satisfied.

Unfortunately, when $K_N(\omega - \omega')$ is plotted the function becomes sharply peaked as N becomes large and the first peaks on either side of the main one are significant and approach a constant fraction of the main peak.

If an approximation of a function with a discontinuity is tried, say,

$$H(\omega) = \begin{cases} 1, & |\omega| \leq \omega_c \\ 0, & |\omega| > \omega_c \end{cases}$$

then these peaks, even for large N , cause the "Gibbs phenomena" severe oscillations about the point $\omega = \omega_c$ (Figure 5.2-1). As more general functions with discontinuities can be represented as the sum of a step function plus a continuous function, the "Gibbs phenomena" is still present for these functions (Section 5.6).

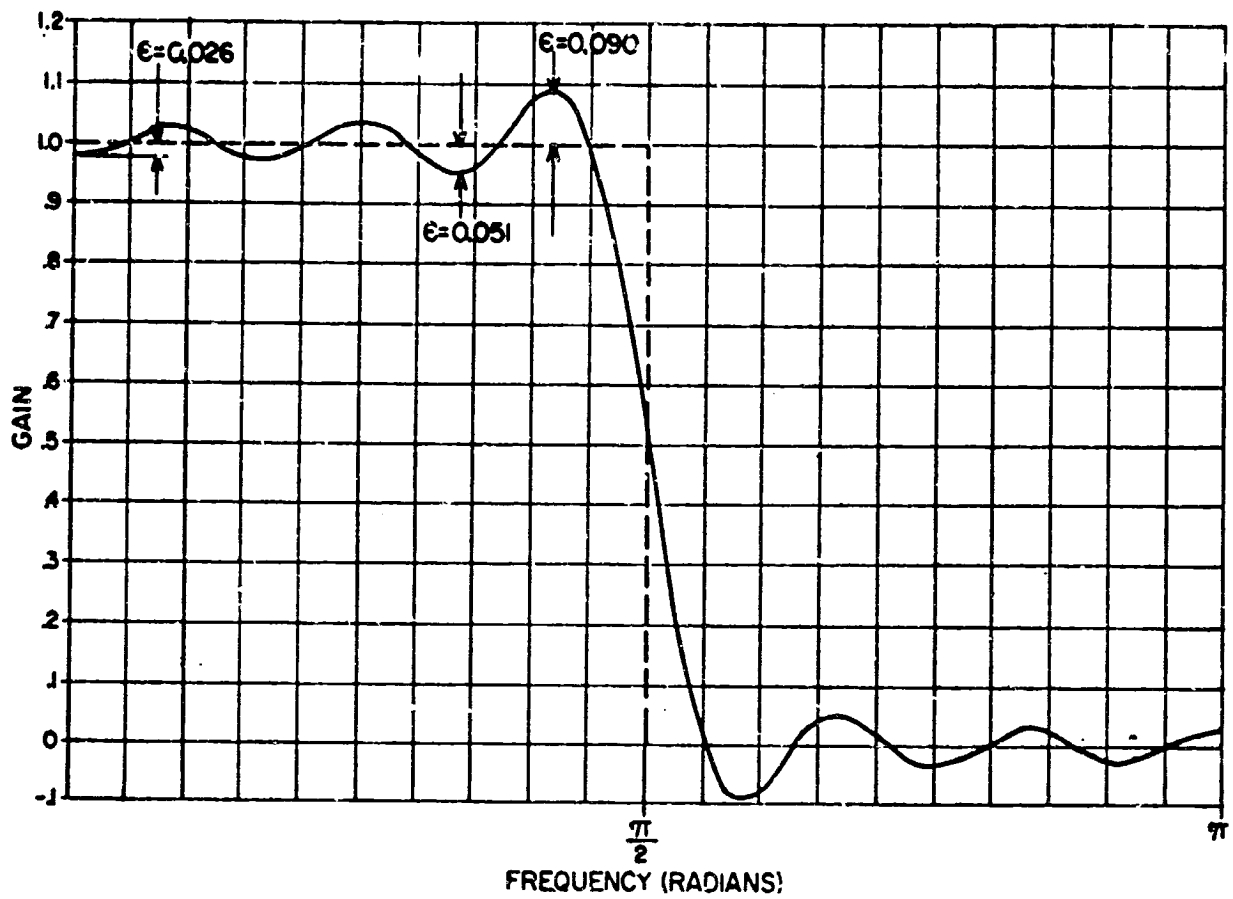


Figure 5.2-1. Gibbs Oscillations ($N = 12$)

It can be shown that the first overshoot approaches a limit $\neq 1$ for increasing N and similarly for the first undershoot. In many practical circumstances these "ripples" cause unwanted non-negligible effects in the filtering process. The limits are .08949 for the overshoot and -.08949 for the undershoot.

A point of view that may be taken towards the "Gibbs phenomena" is that the kernel $K_N(\omega - \omega') = K_N(-(\omega - \omega'))$ does not have sufficient "focusing power." It does not fall-off sufficiently fast away from $\omega = \omega'$. Its rather large oscillations outside of the immediate region of $\omega = \omega'$ cause the large oscillations in the truncated Fourier series $G^N(\omega)$ at a point of discontinuity.

The fact that the truncated Fourier series approaches $G(\omega)$ uniformly except at points of discontinuity is of exceptional interest as this allows approximations within a given error over the entire range $0 \leq \omega \leq \pi$ for some N .

Methods of suppressing the "Gibbs phenomena" will be discussed later.

5.2.1 Orthonormal Expansions and their Relation to Fourier Series

An expansion of a function $G(\omega)$ over a region $a \leq \omega \leq b$ in terms of an "orthonormal" function set is often possible. An "orthonormal" set is a set of linearly independent functions $\varphi_j(\omega)$ which have the properties

$$\int_a^b |\varphi_j(\omega)|^2 d\omega = 1 \quad (5.2.1-1a)$$

and

$$\int_a^b \varphi_j(\omega) \varphi_n^*(\omega) d\omega = 0; j \neq n \quad (5.2.1-1b)$$

An "orthogonal" set is one which has not been normalized so that (5.2.1-1a) is true. It, of course, can be so normalized.

An expansion of $G(\omega)$ in the form

$$G(\omega) = \lim_{N \rightarrow \infty} \sum_{j=0}^{\overline{N}} A_j \varphi_j(\omega). \quad (5.2.1-2)$$

is often desired. Usually it is also desired that this expansion converge over the region $a \leq \omega \leq b$ for a very general class of functions $G(\omega)$. That is, restrict $G(\omega)$ by as few conditions as possible. The Dirichlet conditions of Section (5.2) are as general as needed in most practical problems.

Assuming the expansion holds both sides may be multiplied by $\varphi_n^*(\omega)$ and integrated from a to b . Using the orthonormality conditions it is found that

$$A_n = \int_a^b G(\omega) \varphi_n^*(\omega) d\omega. \quad (5.2.1-3)$$

Letting

$$\begin{aligned} \varphi_0(\omega) &= \frac{1}{\sqrt{2\pi}} \\ \varphi_{2n}(\omega) &= \frac{1}{\sqrt{\pi}} \cos n\omega; n = 1, 2, \dots; \end{aligned} \quad (5.2.1-4)$$

and

$$\varphi_{2n+1}(\omega) = \frac{1}{\sqrt{\pi}} \sin n\omega, n = 1, 2, \dots;$$

over the region $-\pi \leq \omega \leq \pi$, then the Fourier series (equation 5.2-3) with slightly different normalization would be obtained.

If the general orthonormal expansion had been truncated at $\bar{N} + 1$ terms

$$\bar{G}^{\bar{N}}(\omega) = \sum_{j=0}^{\bar{N}} A_j \varphi_j(\omega), \quad (5.2.1-5)$$

then the coefficients of A_j could be determined so that they minimize the square error:

$$\epsilon^2 = \int_a^b \left| G(\omega) - \sum_{j=0}^{\bar{N}} A_j \phi_j(\omega) \right|^2 d\omega \quad (5.2.1-6)$$

Minimization is obtained by taking the partial derivatives with respect to the coefficients and setting them equal to zero. They are

$$A_j = \int_a^b G(\omega) \phi_j^*(\omega) d\omega. \quad (5.2.1-7)$$

Amazingly, the coefficients are the same as obtained before when deriving equation (5.2.1-3).

It should be kept in mind that least squares fitting only minimizes the integrated square error. The error at a given point can be quite large. Fortunately, when sines and cosines are used starting with the 0th harmonic and adding in the next sine and cosine harmonic, as $\bar{N}$ is increased by one, the set $\phi_j(\omega)$ has been ordered in such a way that the series uniformly converges and the absolute error also decreases.

The exponential form of the trigonometric series which is also composed of orthogonal functions is not ordered as simply, so that the expansion does not converge if the limiting is not performed properly. For example, the general expansion

$$G(\omega) = \lim_{\substack{N_1 \rightarrow \infty \\ N_2 \rightarrow \infty}} \sum_{n=-N_1}^{N_2} c_n \frac{e^{-i\omega n}}{\sqrt{2\pi}}; \quad -\pi \leq \omega \leq \pi; \quad (5.2.1-8)$$

will not converge, in general, to $G(\omega)$ if say, $N_2 \rightarrow \infty$ only is allowed. Of course, if $G(\omega)$ does not contain the members of $e^{i\omega n}$ for $n < -N_1$, then convergence is possible. This exceptional case happens to be the case of interest for realizable digital filters.

Also, it should be noted that the orthonormal trigonometrical exponential set can be arranged in the manner of the $\phi_j(\omega)$'s so that, as each term (or pair) is added, convergence to $G(\omega)$ is uniform.

Writing

$$\varphi_0(\omega) = \frac{1}{\sqrt{2\pi}}$$

$$\varphi_{2n}(\omega) = \frac{e^{-in\omega}}{\sqrt{2\pi}}, \quad n > 0 \quad (5.2.1-9)$$

$$\varphi_{2n+1}(\omega) = \frac{e^{in\omega}}{\sqrt{2\pi}}, \quad n > 0$$

and the form equivalent to (5.2.1-2) becomes

$$G(\omega) = \lim_{\bar{N} \rightarrow \infty} \sum_{j=0}^{\bar{N}} A_j \varphi_j(\omega) \quad (5.2.1-10)$$

$$= \frac{A_0}{\sqrt{2\pi}} + \lim_{N \rightarrow \infty} \sum_{j=1}^N (A_{2j} \frac{e^{-ij\omega}}{\sqrt{2\pi}} + A_{2j+1} \frac{e^{ij\omega}}{\sqrt{2\pi}})$$

Examination of this form, using the identities

$$e^{\pm ij\omega} = \cos j\omega \pm i \sin j\omega$$

will disclose the fact that, as $\bar{N}$ increases, successive harmonics of both sines and cosines are being added allowing the approximation error to decrease uniformly. If the A_{2n} terms should be ignored and approximation by a series of the form

$$G^{\bar{N}}(\omega) \simeq \sum_{j=0}^{\bar{N}} A_{2j+1} \frac{e^{-ij\omega}}{\sqrt{2\pi}} \quad (5.2.1-11)$$

failure is imminent in general, except for the special case where $G(\omega)$ does not contain the positive exponential terms.

5.2.2 Completeness of an Orthonormal Set

The last few pages have been a specialization of a particular fact about "completeness" of orthonormal sets. If the orthonormal set (in this case the set consists of the functions $e^{\pm i k \omega}$) is not "complete," then a function cannot be fitted satisfying the Dirichlet conditions in a converging manner.

Mathematically, completeness means that the set $\varphi_j(\omega)$ possesses the property:

$$\begin{aligned} K(\omega, \omega') &\equiv \sum_{j=0}^{\infty} \varphi_j(\omega) \varphi_j^*(\omega') \\ &= \delta(\omega - \omega'). \end{aligned} \quad (5.2.2-1)$$

where $\delta(\omega - \omega')$ is the "Dirac delta function." $K(\omega, \omega')$ is a "kernel" with the property:

$$G(\omega) = \int_a^b G(\omega') \delta(\omega - \omega') d\omega'; \quad a \leq \omega \leq b \quad (5.2.2-2)$$

The "completeness" property is obtained by inserting the coefficients

$$A_j = \int_a^b G(\omega') \varphi_j^*(\omega') d\omega' \quad (5.2.1-3)$$

back into the expansion that was assumed to exist:

$$G(\omega) = \lim_{\bar{N} \rightarrow \infty} \sum_{j=0}^{\bar{N}} A_j \varphi_j(\omega). \quad (5.2.1-2)$$

After interchanging the orders of integration and summation, it is found $\varphi_j(\omega)$ must have the property (5.2.2-1).

The preceding discussion about the functions $\frac{1}{\sqrt{2\pi}} e^{\pm in\omega}$ simply reiterates the fact that a general transfer function can be synthesized (approximated) in a converging manner only if the exponentials for all $n = 0, \pm 1, \pm 2, \dots$ are present as N increases. A "complete" set must be used!

A fact of importance that should be remembered in later work, occurs when filters are synthesized to fit exactly at $M+1$ given points of the desired transfer function. While the fit is exact at the $M+1$ points in both amplitude and phase, interpolation points between the fitted points can deviate wildly from the desired function. This can happen as the function for which a fit is desired can contain members not in the approximating set.

5.3 RELATION BETWEEN FOURIER SERIES EXPANSIONS AND LINEAR DIGITAL FILTER TRANSFER FUNCTIONS (INCLUDING RECURSIVE FILTERS)

The general non-recursive, linear filter frequency response was expressed as

$$H_z(\omega) = \sum_{m=0}^M w_m e^{-i\omega mT} \quad (5.1-3)$$

It is immediately seen from Section 5.2.2 that legitimate transfer functions belong to those functions which can be expressed with negative (and zero) trigonometric exponentials.

Extending M to infinity includes all linear, time invariant digital filters which may be recursive.

When the transfer function of a non-recursive filter was written as

$$H_z(z) = z^{-\tau} H_z^{(\tau)}(z), \quad (5.1-6)$$

with

$$H_z^{(\tau)}(z) = \sum_{m=-\tau}^{M-\tau} w_{m+\tau} z^{-m} \quad (5.1-5)$$

it was simply on algebraic manipulation which still had only negative and zero powers of z (or $e^{i\omega T}$). Equation (5.1-5) above was called the "delayed" non-recursive filter form. At this point it should be more clear why the non-recursive transfer function was written in this manner. For example, when $\tau = N$ in (5.1-5) above, with $M = 2N$ for expository purposes:

$$H_z(\omega) = e^{-i\omega NT} \sum_{m=-N}^N w_{N+m} e^{-i\omega mT} \quad (5.3-1)$$

when $z = e^{i\omega T}$ which gives the frequency response. It can be seen that if a delay τ of N units is allowed (NT seconds) corresponding to the factor $e^{-i\omega NT}$ then the second factor contains both positive and negative exponents in equal quantities. Therefore, with $\tau = N$ an arbitrary transfer function $H(\omega)$ can be approximated satisfying the Dirichlet conditions (Section 5.2) by (5.1-5), by allowing N to be sufficiently large. However this means the delay also gets larger.

A pertinent example of the relation between delay and phase is the approximation of an ideal bandpass filter by a non-recursive digital filter.

For example, if a low-pass filter whose amplitude is defined by

$$H(\omega) = \begin{cases} 1, & |\omega| \leq \omega_0 < \omega_s/2, \\ 0, & \omega_0 < |\omega| \leq \omega_s/2, \end{cases}$$

is desired, then the closer the approximation the more delay required. In Section 5.7.4 it will be found that the "fall-off" interval (the interval where the approximation falls from 1 to 0, approximately) is given by

$$\Delta\omega = \frac{1}{N+1} \omega_s/2,$$

for Fourier approximations without modification for the Gibbs phenomena. If rapid "fall-offs" are desired ($\Delta\omega$ small) then N must be concomitantly large and, therefore, the delay must be large, (NT).

5.3.1 The Transfer Function Approximation Problem and the Determination of τ .

The approximation problem for an arbitrary transfer function, with possibly delay, is now clear.

Definition of $H_z^{(\tau)}(z)$ was obviously made so that it possessed terms with positive and negative exponentials; the quantity of each controlled by the integer variable τ . Assuming τ has been selected, at first, near the center of the range 0 to M , so that the approximating exponentials are present in relatively equal quantities, the approximation problem follows.

Given a transfer function $H(\omega)$ for which an approximation is desired, with $H(\omega)$ given either in numerical or functional form, determine the numbers w_n so that the error

$$\epsilon(\omega) = H(\omega) - H_z^{(\tau)}(\omega)$$

(5.3.1-1)

$$= H(\omega) - \sum_{m=-\tau}^{M-\tau} w_{m+\tau} e^{-i\omega mT}$$

is minimized in some sense.

Using the least-squares method of approximation (or any other) along with convergence factors for removal of the Gibbs phenomena to be discussed in Section 5.7 the method of approximation is relatively straight forward.

1. Determine by some manner (trial-and-error, convergence rate of Fourier coefficients (Section 5.4), analytic properties of $H(\omega)$, etc.) the initial number of weights $M+1$.
2. Set τ to $(M+1)/2$
3. Compute the weights w_k
4. Examine the fitting to $H(\omega)$ by $H_z^{(\tau)}(\omega)$
5. If the result of 4) is adequate examine the weights w_n .

Quite often in practical problems the weights will diminish rapidly to insignificance for n near zero. If they are small near zero, τ may be decreased by an amount equal to the number of insignificant weights with an attendant decrease in the number $(M+1)$ of weights. Step 4) should be repeated as a check on removal of the "insignificant" weights.

5.3.2 Fourier Least-Squares or "Dirichlet" Weights

Using the least-squares methods of Section 5.2 the "Dirichlet" weights are easily found. Comparison of equation (5.2-1) with (5.1-5) ($z = e^{i\omega T}$) gives the relationship between the "Dirichlet" or least-squares coefficients and the filter weights.

$$c_n = w_{n+\tau}, \quad -\tau \leq n \leq M-\tau, \quad (5.3.2-1)$$

or

$$w_n = c_{n-\tau}, \quad 0 \leq n \leq M$$

with

$$c_n = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H(\omega) e^{i\omega n T} d\omega \quad (5.3.2-2)$$

and $H(\omega)$ is the transfer function undergoing approximation.

Generally, $H(\omega)$ may be the transfer function of a complex filter (see Section 8) producing complex weights. If $H(-\omega) = H^*(\omega)$, as is often the case, then the weights are real numbers.

If $H(\omega)$ is given in amplitude and phase form and has a real impulse response ($H(-\omega) = H^*(\omega)$) then

$$H(\omega) = A(\omega) e^{i\phi(\omega)} \quad (5.3.2-3)$$

where

$A(\omega)$ = amplitude, an even function

and

$\phi(\omega)$ = phase (in radians), an odd function

then the "least squares" weights can be written as

$$w_n = \frac{1}{\pi} \int_{-\pi}^{\pi} A\left(\frac{\omega}{T}\right) \cos\left(\phi\left(\frac{\omega}{T}\right) + \omega(n-\tau)\right) d\omega. \quad (5.3.2-4)$$

Many limitations present were mentioned previously, but will be reiterated for completeness; others are included.

- a. $H(\frac{\omega}{T})$ must be known analytically between its discontinuities, if any.
 - b. Numerical integration is required in most cases for the w_n , where at least $N + 1$ complex sample values of the desired transfer function are needed. The coefficients will become periodic with period $2Q + 1$, where $2Q + 1$ is the number of real and imaginary (amplitude and phase) transfer function samples available, with most numerical integration techniques. This means that $2Q + 1$ is the maximum number of coefficients that can be obtained using typical numerical integration schemes. Fortunately, the integrations can be simplified by a discrete transform technique. (Section 5.5).
 - c. The total square error is minimized and not the maximum error. This effect is seldom pronounced when the "Gibbs phenomena" is properly handled.
 - d. For delays small, $\tau \simeq 0$, the convergence of the approximation is usually poor. This is due to lack of convergence to a "complete" set of functions $e^{\pm i\omega k}$. Details are given in Section 5.2.
1. The advantages of the technique are pronounced:
 - a. For $\tau \simeq (M+1)/2$ there will be convergence to a "complete" set for increasing M ; convergence is as fast as the coefficients w_{N-n} decrease with n (as fast as n^{-1} , even with discontinuities at $\omega = 0$ and π in the imaginary part of the transfer function). If the imaginary part of $H(\omega) = 0$ at $\omega = 0$ and π then convergence is at least as fast as n^{-2} . (See Section 5.4).
 - b. The "Gibbs phenomena" can be removed simply (Section 5.6) with a small price to pay in accuracy. Usually better accuracy for a given class of filters with a smaller N is possible; for example, "bandpass" filters.
 - c. w_n 's that contribute relatively little can be recognized by examining the ratio:

$$\frac{|w_n|}{|H_z(\omega)|}$$

If this ratio be less than the tolerable error for important ω this term can be dropped as the $e^{ik\omega}$ are "equi-ripple" functions with a maximum ripple of 1.

5.4 CONVERGENCE RATES OF THE REAL AND IMAGINARY PARTS OF THE TRANSFER FUNCTION, OR MAGNITUDE OF FOURIER COEFFICIENTS IN GENERAL.

In Section 5.1 it was found that $H_z^{(\tau)}(\omega)$ possessed the property

$$H_z^{(\tau)}(-\omega) = H_z^{(\tau)}(\omega)^* \quad (5.1-9)$$

for real filters, or weights which, if $H_z^{(k)}(\omega)$ is written in terms of its real and imaginary parts:

$$H_z^{(\tau)}(\omega) = f(\omega) + ig(\omega), \quad (5.4-1)$$

implies that $f(\omega)$ is even:

$$f(-\omega) = f(\omega) \quad (5.4-2a)$$

and $g(\omega)$ is odd:

$$g(-\omega) = -g(\omega) \quad (5.4-2b)$$

Furthermore, the correspondence developed in Section 5.3.2 relates the transfer function weights to the Fourier expansion coefficients. How the expansion coefficients behave for large indices, therefore, becomes important because it informs about relative error magnitudes in transfer function synthesis.

Because of the evenness of $f(\omega)$ and oddness of $g(\omega)$ it is found that the coefficients a_n and b_n are given by:

$$\begin{aligned} a_n &= \frac{2}{\pi} \int_0^{\pi} f(\omega) \cos \omega n \, d\omega, \quad n > 0; \\ b_n &= \frac{2}{\pi} \int_0^{\pi} g(\omega) \sin \omega n \, d\omega. \end{aligned} \quad (5.4-3)$$

The order of magnitude of the coefficients a_n and b_n can be estimated. Consider, first the coefficients a_n ; for $f(\omega)$ continuous and n small it is difficult to estimate their behavior in general. With large n the factor $f(\omega)$ changes relatively slowly. Integration by parts may be performed:

$$\begin{aligned}\frac{\pi}{2} a_n &= \left. \frac{f(\omega) \sin \omega n}{n} \right|_0^\pi - \frac{1}{n} \int_0^\pi f'(\omega) \sin n\omega \, d\omega \\ &= -\frac{1}{n} \int_0^\pi f'(\omega) \sin n\omega \, d\omega\end{aligned}\quad (5.4-4)$$

Integration again by parts results in

$$\frac{\pi}{2} a_n = \frac{1}{n^2} \left. f'(\omega) \cos n\omega \right|_0^\pi - \frac{1}{n^2} \int_0^\pi f''(\omega) \cos n\omega \, d\omega. \quad (5.4-5)$$

It is now possible to say that for large n the coefficients a_n fall off at a rate of at least $1/n^2$ for continuous functions.

Similarly, the behavior of the coefficients b_n can be investigated for continuous functions $g(\omega)$. Integrating by parts:

$$\begin{aligned}\frac{\pi}{2} b_n &= - \left. \frac{g(\omega) \cos \omega n}{n} \right|_0^\pi + \frac{1}{n} \int_0^\pi g'(\omega) \cos \omega n \, d\omega \\ &= -((-1)^n g(\pi) - g(0)) + \frac{1}{n} \int_0^\pi g'(\omega) \cos \omega n \, d\omega\end{aligned}\quad (5.4-6)$$

Unless, $g(\pi) = g(0) = 0$, the b_n descend only at the rate of n^{-1} . If $g(\pi) = g(0) = 0$, then further integrations by parts demonstrate that the b_n 's, the sine series, converges at the rate n^{-3} .

Analogous considerations, using integration by parts, for general discontinuities (assuming a finite number within $0 \leq \omega \leq \pi$), demonstrate that both sine and cosine coefficients, then only fall-off as $1/n$ for large n .

Some other properties of expansions in Fourier series aid in convergence, but remove discussion from the interval $-\pi \leq \omega \leq \pi$, and will not be considered here.

In summary it can be seen that sine expansions are preferred if the constant 90 degree phase shift can be tolerated and the desired transfer function is zero at $g(\pi)$ and $g(0)$ and is odd, as the convergence rate for large n is n^{-3} . If $g(\pi)$ and $g(0)$ are not zero then the convergence is only as n^{-1} . A method of assisting in making $g(\pi)$ and $g(0)$ equivalent to zero is discussed in Section 5.6 though the discontinuities are still left to be expanded.

Real transfer functions, cosine transfer functions, converge as n^{-2} if discontinuities are not present.

If a restriction to sine or cosine expansions is made, then varying phase shifts cannot be synthesized.

5.5 FOURIER LEAST SQUARES APPROXIMATION USING DISCRETE FOURIER TRANSFORMS INCLUDING EXACT FITTING AT $N + 1$ EQUI-SPACED POINTS

A major limitation discussed in a previous section, in "least-squares" fitting using an analytical expression for $H(\omega)$ was the difficulty in performing the integrations necessary for computing the weights either by analytical or numerical integration techniques. Fortunately, a simple method is available in the numerical case which converges to the exact least squares weights for a sufficient number of complex sample values of the transfer function. The methods needed to remove the "Gibbs phenomena" are also applicable here.

Generally, complex coefficient non-recursive filters can be computed by the method to be described. When the input transfer function is restricted to having the property $H(-\omega) = H^*(\omega)$ real weights result. The IDFT2 of Section 2.6.1 will be useful in subsequent paragraphs.

Trigonometric functions possess integral orthonormality over the normalized region $-\pi \leq \omega \leq \pi$ when ω is a continuous variable as discussed previously. Amazingly, they also possess this relatively unique property over equally spaced discrete samples. As complex exponential forms are easier to manipulate only these forms will be discussed. Equivalent results can be deduced from these forms for sines and cosines.

As in the continuous case with "least-squares" fitting the magnitude of the error over the range $-\frac{\omega_s}{2} \leq \omega \leq \frac{\omega_s}{2}$ is squared to avoid cancellation and summed. The weights w_n are chosen to minimize the summed squared error, hence the name "least-squares."

The total square error between the desired transfer function $H(\omega)$ and the approximation $H_z^{(\tau)}(\omega)$ at the frequency sampling points ω_q is

$$\epsilon^2 = \sum_{q=-Q_0}^{Q_0} \left| H(\omega_q) - H_z^{(\tau)}(\omega_q) \right|^2 \quad (5.5-1)$$

where

$$\omega_q = q \frac{2\pi}{QT},$$

$$Q = \begin{cases} 2Q_0, & Q \text{ even,} \\ 2Q_0+1, & Q \text{ odd,} \end{cases}$$

$$Q'_0 = Q - Q_0 - 1$$

$$= \begin{cases} Q_0 - 1, & Q \text{ even} \\ Q_0, & Q \text{ odd} \end{cases}$$

Inserting the form (5.1-5) for $H_z^{(\tau)}(\omega)$ and taking the partial derivative with respect to $w_{m+\tau}$ (or its conjugate):

$$0 = \frac{\partial \epsilon^2}{\partial w_{m+\tau}^*} \quad (5.5-2)$$

$$= - \sum_{q=-Q'_0}^{Q_0} H(\omega_q) e^{i\omega_q mT} + Q w_{m+\tau};$$

where the discrete orthogonality of Section 2.6.1 has been used. Therefore, $w_{m+\tau}$ is given by

$$w_{m+\tau} = \frac{1}{Q} \sum_{q=-Q'_0}^{Q_0} H(\omega_q) e^{i\omega_q mT} \quad (5.5-3)$$

with $\omega_q = q\omega_s/Q$ and can be computed directly using complex arithmetic procedures. When $H(\omega)$ is the transfer function of a real filter ($H(-\omega) = H^*(\omega)$) the weights are real. It may be noticed that the weights are given by the IDFT2 of Section 2.6.1.

If the transfer function is that of a real filter and given in amplitude and phase form

$$H(\omega) = A(\omega) e^{i\varphi(\omega)}$$

then the weights can be written:

$$\underline{Q = 2Q_0 \text{ (even)}}$$

$$w_{m+\tau} = \frac{1}{Q} A(0) \cos \varphi(0) + \frac{1}{Q} A\left(\frac{\omega_s}{2}\right) \cos\left(\frac{\omega_s}{2} Q_0 T + \varphi\left(\frac{\omega_s}{2}\right)\right) + \frac{2}{Q} \sum_{q=1}^{Q_0-1} A(\omega_q) \cos(\omega_q mT + \varphi(\omega_q)) \quad (5.5-4a)$$

$$\underline{Q = 2Q_0 + 1 \text{ (odd)}}$$

$$w_{m+\tau} = \frac{1}{Q} A(0) \cos \varphi(0) + \frac{2}{Q} \sum_{q=1}^{Q_0-1} A(\omega_q) \cos(\omega_q mT + \varphi(\omega_q)) \quad (5.5-4b)$$

For real filters it should also be noted that $\varphi(0)$ and $\varphi(\omega_s/2)$ are restricted to multiples of π .

The choice of τ for the weights computed by discrete sampling of the transfer function is performed in a manner identical to that for continuous least squares fitting (Section 5.3.1).

When the number of transfer function samples, Q , is equal to the number of filter weights, $M+1$, then the fitting will be exact at the points ω_q as(5.5.3)

may be recognized as the IDFT2 of Section 2.6.1. However as τ approaches zero the fit between points often goes completely awry. This occurs because of the "incomplete set" again.

Determination of Q so that the discrete coefficients approach the continuous ones is most clearly understood by consulting Section 2.6.6. As work there in the time domain the relation is converted here to the frequency domain.

If $H(\omega)$ is expanded in a Fourier series as in Section 5.3.2.

$$H(\omega) = \sum_{m=-\infty}^{\infty} c_m e^{-i\omega mT} \quad (5.5-5)$$

with

$$c_m = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H(\omega) e^{i\omega mT} d\omega, \quad \omega_s = \frac{2\pi}{T}$$

then the discrete coefficients, C_m , are given in terms of the c_m by

$$C_m = \sum_{j=-\infty}^{\infty} c_{m-jQ} \quad (5.5-6)$$

The method of Section 2.6.6 is used identically to derive(5.5-6).

allow Q should be chosen so that the "aliasing" indicated is minimized to

$$C_m \approx c_m \quad (5.5-7)$$

In either case the digital filter coefficients are related by

$$w_m = C_{m-\tau}; \quad m = 0, 1, \dots, M, \quad (5.5-8)$$

or

$$w_m = c_{m-\tau}; \quad m = 0, 1, \dots, M.$$

5.6 SUPPRESSION OF THE "GIBBS PHENOMENA"

The "Gibbs Phenomena", as mentioned previously (Section 5.2), is due to the so-called Dirichlet "Kernel", that is, if the approximation to the function $G(\omega)$ is represented as

$$G^N(\omega) = \sum_{n=-N}^N c_n e^{i\omega n} \quad (5.2-7)$$

with the usual Fourier coefficients

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(\omega') e^{i\omega' n} d\omega', \quad (5.2-8)$$

and these are inserted back into the expression for $G^N(\omega)$,

$$G^N(\omega) = \int_{-\pi}^{\pi} G(\omega') \frac{\sin[(N + \frac{1}{2})(\omega - \omega')]}{2\pi \sin[(\omega - \omega')/2]} d\omega' \quad (5.2-10)$$

results. Integration and summation have been reversed in obtaining this equation.

Also, as mentioned in Section 5.2, if the Dirichlet conditions are satisfied, then

$$\lim_{N \rightarrow \infty} G^N(\omega) = \lim_{N \rightarrow \infty} \int_{-\pi}^{\pi} G(\omega') \frac{\sin[(N + \frac{1}{2})(\omega - \omega')]}{2 \sin[(\omega - \omega')/2]} d\omega' \quad (5.2-12)$$

at all points where $G(\omega)$ is continuous and is equal to

$$1/2 [G(\omega_0^+) + G(\omega_0^-)]$$

at a point of discontinuity ω_0 , where ω_0^+ and ω_0^- indicate limits from either side of the discontinuity. If the function $G(\omega)$ is continuous in the interval

$$a \leq \omega \leq b,$$

the convergence of $G^N(\omega)$ to $G(\omega)$ is uniform in the open interval

$$a < \omega < b$$

if the points a and b are points of discontinuity. As a point of discontinuity is approached; the series converges to the value $G(\omega_0^-) + .08949 (G(\omega_0^-) - G(\omega_0^+))$, as ω_0 is approached from the left. An overshoot, regardless of how large N becomes, of approximately 9 percent exists! Similarly, at the right of the discontinuity an undershoot of 9 percent exists. The phenomena is illustrated in Figure 5.2-1.

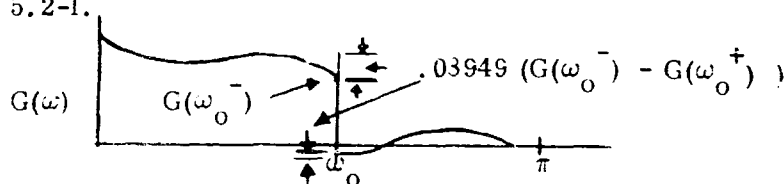


Figure 5.6-1. The "Gibbs Phenomena"

5.6.1 Separation of Discontinuities from Arbitrary Functions

A function other than a simple rectangle, as those in Figure 5.6.1-1

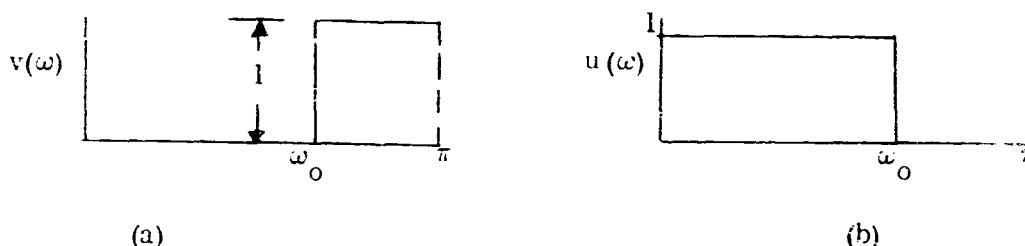


Figure 5.6.1-1 Simple Rectangles

which has a discontinuity at $\omega = \omega_0$ as in Figure 5.6-1 can be decomposed into a continuous function plus a simple "square wave" or rectangle.

For example, removal of the discontinuity in the function $G(\omega)$ of Figure 5.6-1 can be accomplished by adding the rectangle $v(\omega)$ (Figure 5.6.1-1a) weighted by $(G(\omega_0^-) - G(\omega_0^+))$. That is, the function with the discontinuity removed, $G^R(\omega)$, can be written

$$G^R(\omega) = G(\omega) + (G(\omega_0^-) - G(\omega_0^+)) v(\omega) \quad (5.6.1-1)$$

It can be seen that $G^R(\omega)$ is continuous at $\omega = \omega_0$ and can be expanded in a rapidly converging Fourier series (see Section 5.4). Of course, the expansion of

$$G(\omega) = G^R(\omega) - (G(\omega_0^-) - G(\omega_0^+)) v(\omega) \quad (5.6.1-2)$$

is actually desired. Assuming the expansion of $G^R(\omega)$ has been obtained only $v(\omega)$ need be expanded. A function which has a discontinuity that jumps positively as ω passes the point $\omega = \omega_0$ would involve $u(\omega)$ (Figure 5.6.1-1b).

If convenient methods can be found for expanding $u(\omega)$ (Figure 5.6.1-1b) and $v(\omega)$ with high accuracy and a relatively small number of terms, then $G(\omega)$ can be expanded much more accurately. Furthermore, as appropriate analytical forms can be found for the expansions of $u(\omega)$ and $v(\omega)$, the numerical accuracy of the expansion is also improved as only the expansion coefficients of $G^R(\omega)$ must be computed numerically.

Subsequently, expansions of $u(\omega)$ and $v(\omega)$ will be discussed with attention to the "Gibbs phenomena" (Section 5.3)

One other type of discontinuity exists which should be removed before calculation of coefficients is attempted numerically. The even part of the function $G^N(\omega)$ is continuous at $\omega = \pm\pi$ by virtue of its evenness. However, the odd part (the imaginary part) is automatically discontinuous at $\omega = \pm\pi$ unless $G(\omega)$ is zero there. The asymptotic behaviour ($N \rightarrow \infty$) of $G^N(\omega)$ has been assumed.

This odd function, say $g(\omega)$, has the value $-g(\omega)$ at $\omega = -\pi$. Graphically, the function can be presented as in Figure 5.6.1-2:

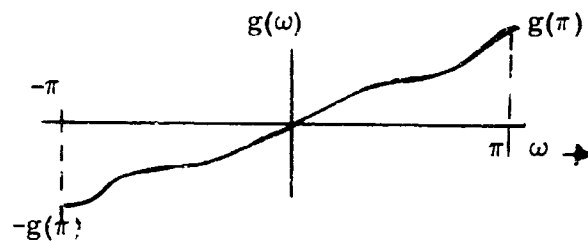


Figure 5.6.1-2. An Odd Function

Assume, of course, that all other discontinuities have been removed by the previous procedure. Removal of the discontinuity at $\pm\pi$ can be performed by decomposing $g(\omega)$ in a part that does not have the discontinuity at $\pm\pi$, $g^{RO}(\omega)$, plus a linear function; that is, the representation

$$g^{RO}(\omega) = g(\omega) - \frac{g(\pi)}{\pi} \omega \quad (5.6.1-3)$$

can be written, or

$$g(\omega) = g^{RO}(\omega) + \frac{g(\pi)}{\pi} \omega \quad (5.6.1-4)$$

It can be seen that $g^{RO}(\omega)$ has no discontinuities and may be expanded in a rapidly converging sine series (Section 5.4).

The preceding discourse has indicated that three fundamental functional forms must be considered with respect to the "Gibbs phenomena": Consideration, of course, must be given to whether these fundamental functions are needed in the real (even) part of the transfer function or in the imaginary (odd) part. The third of the functions $v(\omega)$ is only of use in the imaginary part as it is an odd function. The first two functions $v(\omega)$ and $u(\omega)$ can also be considered as ideal bandpass functions; $v(\omega)$ is a high-pass filter and $u(\omega)$ is a low-pass filter. Linear combinations of $v(\omega)$ for different values

of ω_0 give bandpass filters. $i\omega$, by itself is a "differentiator."

Note that $v(\omega)$ and $u(\omega)$ are related by the simple equation

$$u(\omega) = 1 - v(\omega),$$

or

$$(5.6.1-5)$$

$$v(\omega) = 1 - u(\omega),$$

so that only one of them need be synthesized. Unfortunately, if the discontinuity being removed from a transfer function is in the imaginary part (the sine expansion part), then there will be difficulties if $v(\omega)$ is needed. A discontinuity exists at $\omega = \pm \pi$, as $v(\omega)$ will be odd in this case.

To (1) remove the discontinuities in the real or imaginary parts of a transfer function or (2) design low, high, or bandpass filters discussion of various sine and cosine expansions of $v(\omega)$, $u(\omega)$, and ω will be presented.

5.6.2 Basic Approaches to Removal of the "Gibbs Phenomena"

Two fundamental techniques have been used to reduce the oscillations induced by the "Gibbs Phenomena".

1. Replacing of the discontinuous function by a continuous approximation a la J. Ormsby.*

For example, Ormsby replaced $u(\omega)$, the ideal low-pass filter, by

$$u(\omega) = \begin{cases} 0, & |\omega| > \omega_T \\ 1, & |\omega| \leq \omega_c \\ \left(\frac{\omega + \omega_T}{\omega_T - \omega_c} \right)^p, & -\omega_T \leq \omega \leq -\omega_c \\ \left(\frac{\omega - \omega_T}{\omega_T - \omega_c} \right)^p, & \omega_c < \omega \leq \omega_T \end{cases} \quad (5.6.2-1)$$

which looks approximately like Figure 5.6.2-1.

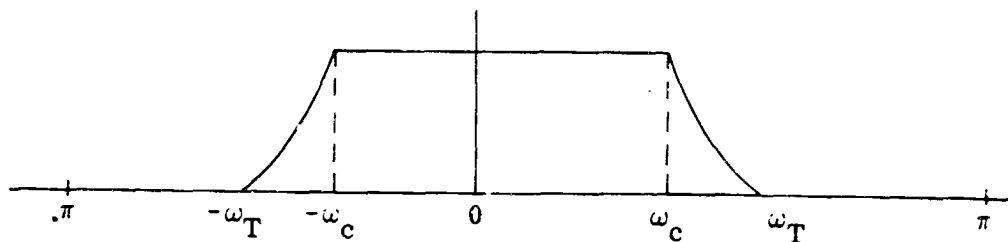


Figure 5.6.2-1. Ormsby Replacement of Low-Pass Filters

*Ormsby, J. F. Design of Numerical Filters with Applications to Missile Data Processing; JACM; Vol. 8, 1961 pp. 440-466.

The integer "p" is the degree of the roll off. Obviously, this is a "cosine" or zero phase shift filter. (It is an even function.)

Other writers have used other roll-off functions.

This general technique of replacement suffers difficulties in the determination of the Fourier coefficients analytically. The integrations are often quite complicated and tedious, and error evaluation as a function of the number of coefficients is difficult to determine. As the "replacement" function is arbitrary an unnecessarily strict requirement is made. Better filters are obtained by a method to be described.

2. Use a filtering technique on the ideal transfer function to remove its undesirable features; namely, the discontinuities.

Many writers have performed this maneuver though they have not given it a common basis. Various names used are:

- a. Use of the Fourier convolution theorem to improve approximations
- b. Use of "apodizing" functions
- c. Weighted Fourier Coefficients
- d. Lanczos factors
- e. Fejer summation (and Cesaro summation in general)

While the first approach to removal of the "Gibbs phenomena" received only a short discussion, as actual designs depend on the method of ideal filter replacement and has received a good deal of attention elsewhere, the filtering technique will be developed fully here.

Definition of "Fall-off" for the Filtering Technique

When rectangular filters are designed it will be found that they have an approximately linear "fall-off" for the useful ones. The middle of the filter "fall-off" usually occurs near $\omega = \omega_0$, where ω_0 is the step in the ideal rectangular filter normalized to unit gain in the passband. By computing the slope, m_s , of the approximating filter, $H_A(\omega)$, at $\omega = \omega_0$ "fall-off" can be found accurately, and as a function of $2N + 1$, the number of weights.

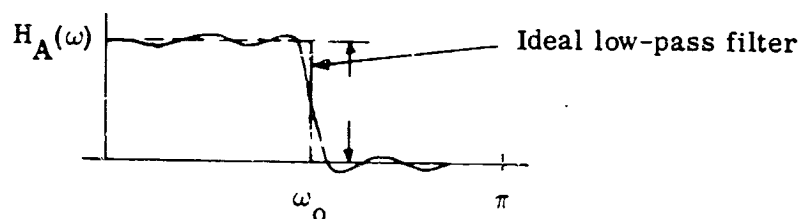


Figure 5.6.2-2. Illustration of "Fall-Off"

The "fall-off" for the filtering technique (and it can also be used for the "replacement" filters) is defined as

$$\lambda_F = \frac{1}{2\pi m_s}$$

where

(5.6.2-2)

$$m_s = \left. \frac{d}{d\omega} (H_A(\omega)) \right|_{\omega = \omega_0}$$

The sampling frequency has been assumed normalized to 2π in radians or unity in Hertz.

The "fall-off" λ_F is the distance in ω between points where a line of slope m_s intersects the ordinates 1 and 0, divided by the sampling frequency. In "non-normalized" frequencies

$$\lambda_F = \frac{1}{\omega_s m_s}$$

where

(5.6.2-3)

$$m_s = \left. \frac{d}{d\omega} (H_A(\omega)) \right|_{\omega = \omega_0}$$

is considered with $H_A(\omega)$ as the transfer function prior to transformation to the interval $-\pi$ to π .

5.6.3 "Filtering" The Transfer Function

If the transfer function is looked upon as simply a periodic function, it could be filtered as, say, a time function. To "filter the filter" is another way of putting it. The reasoning behind this method lies in the normal dealings with square waves in the time domain. Filtering a square wave with various forms of low-pass filters rounds sharp corners, and produces finite rise times.

As the frequency domain and the time domain are connected through Fourier series and transforms, and operations in one domain can equally well be performed in the other, operating upon the exact, ideal sampled data filter with time filtering is considered. That is, operate on

$$H(\omega) = \sum_{k=-\infty}^{\infty} c_k e^{-i\omega k}$$

where

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\omega') e^{i\omega' k} d\omega' \quad (5.6.3-1)$$

with a linear operation, $\mathcal{L}$, in ω (linear combinations of derivatives, summations, and integrations with real coefficients). As with linear time domain operators:

$$H(\omega) = \sum_{k=-\infty}^{\infty} c_k \mathcal{L} e^{-i\omega k} \quad (5.6.3-2)$$

The effect of a linear operator in ω on a function $e^{-i\omega t}$ is well known:

$$\mathcal{L} e^{-i\omega t} = \sigma(t) e^{-i\omega t} \quad (5.6.3-3)$$

where $\sigma(t)$ is the "transfer function" associated with the operator $\mathcal{L}$. As can be seen, the roles of ω and t have simply been reversed from that which is customary.

The problem, ultimately, will be the determination of the proper transfer function in the time domain. (Section 5.7)

Having defined a time domain transfer function, the effect of $\mathcal{L}$ can be written also as

$$\mathcal{L} H(\omega) = \sum_{k=-\infty}^{\infty} c_k \sigma_k e^{-i\omega k} \quad (5.6.3-4)$$

where

$$\sigma_k = \sigma(k) \quad (5.6.3-5)$$

has been defined to simplify notation.

Now set

$$\mathcal{L}_{H_A}(\omega) = \sum_{k=-N}^N c_k \sigma_k e^{-i\omega k} \quad (5.6.3-6a)$$

or

$$\bar{H}_A(\omega) = \sum_{k=-N}^N r_k e^{-i\omega k} \quad (5.6.3-6b)$$

with

$$r_k = c_k \sigma_k \quad (5.6.3-7)$$

and where $H_A(\omega)$ is the truncated approximate transfer function

$$H_A(\omega) = \sum_{k=-N}^N c_k e^{-i\omega k} \quad (5.6.3-8)$$

and $\bar{H}_A(\omega)$ is the filtered version. Equation (5.6.3-6) is a version of equation (5.6.3-4) with the time transfer function adjusted so that

$$\sigma(k) = \begin{cases} k; & k = 0, \pm 1, \pm 2, \dots, \pm N \\ 0; & k > N; k, \text{ an integer.} \end{cases} \quad (5.6.3-9)$$

In general, many $\sigma(t)$'s can be found for a given set of σ_k 's. Note that filtering $H(\omega)$ is the same as weighting the Fourier coefficients by weights σ_k .

Hereafter, $\sigma(t)$ will be superscripted with N; namely $\sigma^N(t)$, so that it will be recognized that the filter depends on the order of approximation of $H(\omega)$ that is desired.

To obtain another method of formulating the problem let

$$\sigma^N(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G^N(\omega) e^{i\omega t} d\omega \quad (5.6.3-10)$$

where $G^N(\omega)$ is the Fourier transform of $\sigma^N(t)$ and can be considered its "impulse response" in the frequency domain. In terms of $H(\omega)$ and $G^N(\omega)$, $r_k^N = \sigma_k^N c_k$ can be written (from equations (5.6.3-1) and (5.6.3-10)) as

$$r_k^N = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} H(\omega') e^{i\omega'k} d\omega' \int_{-\infty}^{\infty} G^N(\omega'') e^{i\omega''k} d\omega'' \quad (5.6.3-11)$$

Defining a truncating function $U(\omega)$ as

$$U(\omega) = \begin{cases} 1, & |\omega| \leq \pi \\ 0, & \text{otherwise;} \end{cases} \quad (5.6.3-12)$$

gives:

$$r_k^N = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} H(\omega') U(\omega') e^{i\omega'k} d\omega' \int_{-\infty}^{\infty} G^N(\omega'') e^{i\omega''k} d\omega'' \quad (5.6.3-13)$$

Namely, r_k^N can be looked upon as the sampled values of a function $r^N(t)$, which is the product of two time functions; the inverse transforms of $U(\omega)$ $H(\omega)$ and $G(\omega)$. That is it can be written

$$r^N(t) = c(t) \sigma^N(t) \quad (5.6.3-14)$$

where $c(t)$ is the inverse transform of the truncated transfer function (to the interval $|\omega| \leq \pi$). At this point, if desired, $H(\omega)$ can be considered as an arbitrary transfer function extending to $\pm \infty$ as $U(\omega)$ suitably restricts it.

By the convolution theorem (Section 2.7.1) the transform of $r^N(t)$ can now be written as

$$\begin{aligned} R^N(\omega) &= \int_{-\infty}^{\infty} r^N(t) e^{-i\omega t} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega' U(\omega') H(\omega') G^N(\omega - \omega') \end{aligned} \quad (5.6.3-15a)$$

and also

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega' U(\omega - \omega') H(\omega - \omega') G^N(\omega') \quad (5.6.3-15b)$$

or as $U(\omega - \omega')$ is the truncating function:

$$R^N(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\omega') G^N(\omega - \omega') d\omega' \quad (5.6.3-16)$$

If $H(\omega')$ is a low pass filter (Figure 5.6.1-1)

$$H(\omega') = \begin{cases} 1, & |\omega| \leq \omega_0 \\ 0, & \text{otherwise;} \end{cases} \quad (5.6.3-17)$$

$$= u(\omega)$$

then, for this special case

$$r_k^N(\omega) = \frac{1}{2\pi} \int_{-\omega_0}^{\omega_0} G^N(\omega - \omega') d\omega' \quad (5.6.3-18)$$

$$= \frac{1}{2\pi} \int_{\omega - \omega_0}^{\omega + \omega_0} G^N(\omega') d\omega'.$$

The coefficients of the expansion of the general filtered truncated approximation $\overline{H}_A(\omega)$ are given by:

$$r_k^N = \frac{1}{2\pi} \int_{-\pi}^{\pi} R^N(\omega') e^{i\omega'k} d\omega' \quad (5.6.3-19)$$

By insertion of the general relationship (5.6.3-15) for $R^N(\omega)$ into this equation for r_k^N , and then placing this into the expression for $\overline{H}_A(\omega)$ (equation (5.6.3-6b)) it is found that (after changes of orders of integration and summation):

$$\overline{H}_A(\omega) = \int_{-\pi}^{\pi} H(\omega') \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} G^N(\omega'' - \omega') \sum_{k=-N}^N \frac{e^{i(\omega'' - \omega)k}}{2\pi} d\omega'' \right] d\omega' \quad (5.6.3-20)$$

Now for large N , it is known that the bracket approaches $G(\omega - \omega')$ if $G^N(\omega)$ satisfies the Dirichlet conditions. See Section 5.2. Then,

$$\overline{H}_A(\omega) \approx \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\omega') G(\omega - \omega') d\omega' \quad (5.6.3-21)$$

as before. For $\bar{H}_A(\omega)$ to approach $H(\omega)$ for large N , then $G^N(\omega)/2\pi$ must also approach the Dirac delta function for large N . Namely, $G^N(\omega)/2\pi$ must become sharply peaked and have an area of unity for N large.

This also means that the time domain filter response

$$\sigma^N(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G^N(\omega') e^{i\omega' t} d\omega' \quad (5.6.3-22)$$

approaches a constant ≈ 1 which would be expected on physical grounds; as N becomes large it also seems that the lower order coefficients $r_k = c_k \sigma_k^N$ must approach c_k to adequately approximate the low frequency components of $H(\omega)$.

Finally, from equations (5.6.3-6) and (5.6.3-1) write:

$$\begin{aligned} \bar{H}_A(\omega) &= \sum_{k=-N}^N c_k \sigma_k^N e^{-i\omega k} \quad (5.6.3-23) \\ &= \int_{-\pi}^{\pi} d\omega' H(\omega') \frac{1}{2\pi} \sum_{k=-N}^N \sigma_k^N e^{-i(\omega - \omega')k} \\ &= \int_{-\infty}^{\infty} d\omega' H(\omega') K_N^{\sigma}(\omega - \omega') d\omega' \end{aligned}$$

where

$$K_N^{\sigma}(\omega) = \frac{1}{2\pi} \sum_{k=-N}^N \sigma_k^N e^{-i\omega k} \quad (5.6.3-24)$$

In effect, it can be seen that the Dirichlet kernel of Section 5.2 has been replaced by another. In order that $\bar{H}_A(\omega)$ converge to $H(\omega)$ for large N , it must be true that $\lim_{N \rightarrow \infty} K_N^{\sigma}(\omega - \omega') = \delta(\omega - \omega')$ where $\delta(\omega)$ is the Dirac delta function.

Recapitulating, this discussion has served to show that various equivalent relations can be developed which are equivalent to filtering the transfer function:

- a. Weighted Fourier Coefficients (equation 5.6.3-6): the coefficients of the filtered transfer functions are given by

$$r_k^N = c_k \sigma_k^N; \quad k = 0, \pm 1, \pm 2, \dots \pm N; \quad (5.6.3-7)$$

where

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\omega) e^{i\omega k} d\omega$$

and

$$\sigma_k = \sigma^N(k)$$

with $\sigma^N(t)$ the time transfer function of the "filter-filter" and is the function to be determined at the times $t = k$.

- b. Use of apodizing functions: this is completely equivalent to (a); determining the coefficients r_k^N from the relation

$$r_k^N = r^N(k) \quad (5.6.3-13)$$

with

$$r^N(t) = c(t) \sigma^N(t) \quad (5.6.3-14)$$

where:

$c(t)$ is the time transfer function of the ideal transfer function over $|\omega| \leq \pi$ (or the impulse response with respect to the truncated $H(\omega)$) and $\sigma^N(t)$ is the "filter filter" transfer function or the apodizing function subject to the condition $\sigma_k^N = 0, |k| > N$.

- c. Use of the Fourier convolution theorem: the coefficients of the filtered transfer function are given by

$$r_k^N = \frac{1}{2\pi} \int_{-\pi}^{\pi} R^N(\omega) e^{i\omega k} d\omega \quad (5.6.3-19)$$

where

$$R^N(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\omega') G^N(\omega - \omega') d\omega' \quad (5.6.3-16)$$

and

$$G^N(\omega) = \int_{-\pi}^{\pi} \sigma_k^N(t) e^{-i\omega t} dt \quad (5.6.3-25)$$

with $\sigma_k^N(t)$ subject to the conditions $\sigma_k^N = 0, k > N$; $G^N(\omega)$ is also subject to the condition that

$$\lim_{N \rightarrow \infty} G^N(\omega) = 2\pi \delta(\omega)$$

with $\delta(\omega)$ equal to the Dirac delta function or, equivalently, that

$$\lim_{N \rightarrow \infty} \sigma_k^N = \delta_k.$$

- d. Use of other summation techniques: equation (5.6.3-23) demonstrates that the filtered approximate transfer function can be represented in terms of an integrating kernel

$$\bar{H}_A(\omega) = \int_{-\pi}^{\pi} d\omega' H(\omega') K_N(\omega - \omega') \quad (5.6.3-23)$$

where

$$K_N(\omega) = \frac{1}{2\pi} \sum_{k=-N}^N \sigma_k^N e^{-i\omega k}$$

and

$$\lim_{N \rightarrow \infty} K_N^\sigma(\omega) = \delta(\omega).$$

It will be shown (Section 5.7) that this is equivalent to other methods of summing a Fourier series, e.g. Fejer summation.

5.6.4 "Filtering" of Non-Central Transfer Functions

If the "delayed" non-central transfer function:

$$H_Z^{(\tau)}(\omega) = \sum_{k=-\tau}^{M-\tau} w_{k+\tau} e^{-i\omega k} \quad (5.1-11)$$

is considered as the approximation to $H(\omega)$ by least-squares techniques it can be operated on with the operator (filter) of Section 5.6.3. From equation (5.6.3-4) it is found that:

$$\begin{aligned} \bar{H}_A(\omega) &= \mathcal{L} H_Z^{(\tau)}(\omega) \\ &= \sum_{k=-\tau}^{M-\tau} w_{k+\tau} \mathcal{L} e^{-i\omega k} \\ &= \sum_{k=-\tau}^{M-\tau} w_{k+\tau} \sigma_k e^{-i\omega k} \end{aligned} \quad (5.6.4-1)$$

Obviously, the coefficients for a given exponential $e^{-i\omega k}$ are

$$r_k = w_{k+\tau} \sigma_k,$$

where

$$\sigma_k = \begin{cases} \sigma_k, & -\tau \leq k \leq M-\tau, \\ 0, & \text{otherwise.} \end{cases} \quad (5.6.5-2)$$

These factors are also useful in reducing the approximation oscillations induced by the functions $e^{-i\omega k}$ being "pushed away" from being a complete set (Section 5.3) as τ becomes small (less delay, Section 5.3). Errors due to the "lop-sided" set will still remain, but at least the severe rippling will be removed.

When the various methods are used to generate the σ_k quite often they adjust the σ_k to zero at the highest frequency plus one. This highest frequency should be taken to be $M - \tau$.

5.7 PARTICULAR ("FILTER FILTER") METHODS

At this point in time it does not seem that the choice of a "filter-filter" has received an analytically rational basis. Future developments should receive attention using the unified formalism of the preceding sections.

Antenna design, where the weights correspond to the illumination, has supplied a number of techniques that work well from an empirical viewpoint and may be justified by the preceding methods. Three choices of the filter $\sigma(t)$ that give useful results are discussed in subsequent pages. Other forms have been investigated by many writers but the three presented here seem to produce the best compromise between reducing the "Gibbs phenomena" error and allowing fast fall-offs for discontinuities. "Optimal trade-offs", which is often a meaningless term in practice, must be determined in future research.

The three methods that show the most promise at present are:

1. Fejer weighting
2. Lanczos weighting
3. "cosine-squared" weighting.

All of these methods have a common property: the time transfer function (the weights) possess the property.

$$\sigma(-t) = \sigma(t), \quad (5.7-1)$$

implying that the transfer function is real. General design should include time transfer functions that are complex. Recent work on antenna design has suggested the usefulness of this possibility. Also, the fact that the linear frequency filter operator possesses a linear combination of differentiations, summations, and integrations, with real coefficients (Section 6.8.3) only restricts $\sigma(t)$ such that it possesses the property $\sigma(-t) = \sigma^*(t)$, as is true for frequency transfer functions, which suggests use of more general filters. However, as the weights of the filters must be real numbers, $\sigma(-t) = \sigma(t)$ for the filters developed thus far.

5.7.1 Fejer or Linear Weighting

Fejer's theorem on Fourier series* states that a function which satisfies the Dirichlet conditions and is magnitude integrable can be summed in the following manner. Defining the partial sum

* T. J. Bromwich: An Introduction to the Theory of Infinite Series; MacMillan and Co.

$$S_n = \sum_{k=-n}^n c_k e^{-i\omega k}; \quad n = 0, 1, 2, \dots \quad (5.7.1-1)$$

where as before

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\omega) e^{i\omega k} d\omega \quad (5.7.1-2)$$

then

$$H(\omega) = \lim_{N \rightarrow \infty} \frac{\sum_{n=0}^N S_n}{N} \quad (5.7.1-3)$$

Suppose the sum is truncated at N terms, then

$$\bar{H}_A(\omega) = \sum_{k=-N}^N \frac{(N+1 - |k|)}{N+1} c_k e^{-i\omega k} \quad (5.7.1-4)$$

Namely, summation by this method produces weights:

$$\sigma_k^N = \begin{cases} \frac{N+1 - |k|}{N+1} & ; \quad |k| \leq N, \\ 0 & ; \quad |k| > N, \end{cases} \quad (5.7.1-5)$$

with

$$\lim_{N \rightarrow \infty} \sigma_k^N = 1$$

as expected.

The equivalent "kernel" of Section 5.6.3 for the Fejer weights is given

by

$$K_N^{\sigma}(\omega) = \frac{\sin^2(N\omega/2)}{2\pi N \sin^2(\omega/2)} \quad (5.7.1-6)$$

For non-central filters $\tau \neq N$, the Fejer weights are found to be

$$\alpha_k = \frac{(M-\tau+1-|k|)}{(M-\tau+1)}, \quad -\tau \leq k \leq M-\tau.$$

5.7.2 Lanczos Weighting

If the filter $\mathcal{L}$, (Section 5.6.3) in the ω domain is defined to be a combination of truncation, or ideal bandpass, at the N th harmonic and a second operator $\mathcal{L}'$, such that it performs "local averaging," then the resultant time filter is known as Lanczos weighting*. The "local averaging" averages over an interval of length $2\pi/(N+1)$ in ω , approximately the length of the period in the highest harmonic of the truncated Fourier expansion of the ideal $H(\omega)$ (averaging over $2\pi/N$ is not chosen as it would remove the N th harmonic). The averaging operation is

$$\begin{aligned} \mathcal{L}H(\omega) &= \mathcal{L}' H_A(\omega) \\ &= \frac{\pi}{N+1} \int_{-\frac{\pi}{N+1}}^{\frac{\pi}{N+1}} H_A(\omega + \omega') d\omega' \end{aligned} \quad (5.7.2-1)$$

where

$$H_A(\omega) = \sum_{k=-N}^N c_k e^{-i\omega k}$$

and

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\omega) e^{i\omega k} d\omega$$

* Cornelius Lanczos, Applied Analysis, Prentice Hall

are the usual Fourier coefficients of the expansion of $H(\omega)$ over $|\omega| \leq \pi$.

It is found that $\bar{H}_A(\omega)$ is given by

$$\bar{H}_A(\omega) = \frac{N+1}{2\pi} \sum_{k=-N}^N c_k e^{-i\omega k} \int_{-\frac{\pi}{N+1}}^{\frac{\pi}{N+1}} e^{-i\omega' k} d\omega' \quad (5.7.2-2)$$

$$= \sum_{k=-N}^N c_k \sigma_k^N e^{-i\omega k}$$

and σ_k^N is given by

$$\sigma_k^N = \begin{cases} \frac{\sin(\frac{\pi k}{N+1})}{(\frac{\pi k}{N+1})}, & 1 \leq |k| \leq N \\ 1, & k = 0. \end{cases} \quad (5.7.2-3)$$

The filtered transfer function has a truncated form of $2N+1$ terms and the original Fourier coefficients, c_k , are weighted by the Lanczos σ_k^N factors.

For non-central filters $\tau \neq N$, the weights are found to be

$$\sigma_k = \frac{\sin(\frac{\pi k}{M-\tau+1})}{(\frac{\pi k}{M-\tau+1})}, \quad -\tau \leq k \leq M-\tau.$$

5.7.3 Cosine Squared Weighting

Results from Fejer weighting and Lanczos weighting suggest that a possible compromise between them would be a function that weights the low frequencies more than the Fejer method and the high frequencies less than the Lanczos method. Such a function is

$$\cos^2(\alpha k)$$

The first zero of $\cos^2(\alpha k)$ is chosen so that it occurs after the last Fourier coefficient. For central fitting:

$$\alpha (N + 1) = \frac{\pi}{2} \quad (5.7.3-1)$$

or

$$\alpha = \frac{\pi}{2 (N + 1)}$$

and the weights are given by:

$$\sigma_k = \cos^2 \left(\frac{\pi k}{2(N + 1)} \right), \quad 0 \leq |k| \leq N \quad (5.7.3-2)$$

α can also be adjusted to make some $\sigma_k = 0$ for $k < N$. Fitting, in general, deteriorates.

For non-central filters $\tau \neq N$, the cosine-squared weights are found to be

$$\sigma_k = \cos^2 \left(\frac{\pi k}{2 (M - \tau + 1)} \right), \quad -\tau \leq k \leq M - \tau.$$

5.7.4 Error Analysis of "Filtered" Filters

For the types of weights used (Fejer, Lanczos, and "cosine-squared"), various parameters of low-pass, bandpass, and high-pass filters are of interest.

Primarily, the concern is with the rate of "fall-off" of the filters (equation (5.6.2-3)) and the errors in the regions outside of the fall-off for a fixed number of weights.

For simplicity only "low-pass" cosine filters are considered. Considerations such as those in Section 5.9 plus Figures 5.10-1 through 5.10-6 indicate that analysis of this special case is sufficient to enable determination of the errors in bandpass filters, which are combinations of low-pass filters (Section 5.9.3) and "sine" filters.

Analysis of ripple errors has had to be performed numerically as the ripple determination by analysis is quite difficult in general. Numerical ripple computation to $N = 1000$ has been computed. N larger than 1000 is usually an impractical situation.

Fall-Off Rates

The approximate fall-offs of the various techniques are found and compared with the Fourier method.

The Fourier filter approximation coefficients of the low-pass filter:

$$H(\omega) = \begin{cases} 1, & |\omega| \leq \omega_0, \\ 0, & \omega_0 < |\omega| < \pi \end{cases} \quad (5.7.4-3)$$

are given by

$$a_k = \frac{1}{2\pi} \int_{-\omega_0}^{\omega_0} e^{i\omega k} d\omega \quad (5.7.4-4)$$

$$= \frac{\sin \omega_0 k}{\pi k}$$

The filtered filter approximation is given by

$$\bar{H}_A(\omega) = \frac{\omega_0}{\pi} + \frac{2}{\pi} \sum_{k=1}^N \sigma_k \frac{\sin \omega_0 k \cos \omega k}{k} \quad (5.7.4-5)$$

Its slope at $\omega = \omega_0$

$$\left. \frac{\partial \bar{H}_A(\omega)}{\partial \omega} \right|_{\omega = \omega_0} = -\frac{2}{\pi} \sum_{k=1}^N \sigma_k \sin^2 \omega_0 k \quad (5.7.4-6)$$

$$= -\frac{1}{\pi} \sum_{k=1}^N \sigma_k + \frac{1}{\pi} \sum_{k=1}^N \sigma_k \cos 2 \omega_0 k$$

From this relation the slopes are computed using the various σ_k 's (Fourier, Lanczos, Fejer, and cosine-squared).

For this purpose various trigonometric identities are derived in Appendix B.

Development of the Lanczos slope will be performed in detail as it is the most complicated. The other slope computations are similar and the results will merely be stated.

Lanczos Slope

$$\left. \frac{\partial \bar{H}_A(\omega)}{\partial \omega} \right|_{\omega = \omega_0} = -\frac{1}{\pi} \sum_{k=1}^N \frac{\sin \frac{\pi k}{N+1}}{\frac{\pi k}{N+1}} + \frac{1}{\pi} \sum_{k=1}^N \frac{\sin \frac{\pi k}{N+1}}{\frac{\pi k}{N+1}} \cos 2\omega_0 k, \quad (5.7.4-7)$$

Use of a trigonometric identity allows this equation to be written as

$$\left. \frac{\partial \bar{H}_A(\omega)}{\partial \omega} \right|_{\omega = \omega_0} = -\frac{N+1}{\pi^2} \sum_{k=1}^N \frac{\sin \frac{\pi k}{N+1}}{k} + \frac{N+1}{2\pi^2} \sum_{k=1}^N \sin \left(2\omega_0 + \frac{\pi}{N+1} \right). \quad (5.7.4-8)$$

From Appendix B these sums can be written in closed form:

$$\left. \frac{\partial \bar{H}_A(\omega)}{\partial \omega} \right|_{\omega = \omega_0} = -\frac{N+1}{\pi^2} \int_0^{\frac{\pi}{N+1}} E_N(\theta') d\theta' + \frac{1}{2} \int_{2\omega_0 - \frac{\pi}{N+1}}^{2\omega_0 + \frac{\pi}{N+1}} E_N(\theta') d\theta', \quad (5.7.4-9)$$

where

$$E_N(\theta) = \frac{\sin \left(N + \frac{1}{2} \right) \theta}{2 \sin (\theta/2)}$$

For $\omega_0 > \frac{\pi}{N+1}$, and N large, the second integral is essentially integration of a sine wave over a cycle, which is relatively small compared to the first integral. The first integral can be computed numerically, relatively simply for large N (greater than 10) and is found to approach

$$\frac{\pi}{N+1}$$

$$\int_0^{\omega_0} E_N(\theta') d\theta' \simeq 1.18 \times \frac{\pi}{2} . \quad (5.7.4-10)$$

0

Therefore, for Lanczos smoothing of a low-pass filter, the slope of the "fall-off" is given at $\omega = \omega_0$ approximately by

$$\left. \frac{\partial \bar{H}_A(\omega)}{\partial \omega} \right|_{\omega = \omega_0} \simeq - \frac{1.18 (N+1)}{2\pi}, \quad \omega_0 > \frac{\pi}{N+1}, \quad N \geq 10. \quad (5.7.4-11a)$$

For the other smoothing factors, the "fall-off" is found in a similar manner: Fourier ($\sigma_k = 1$)

$$\left. \frac{\partial \bar{H}_A(\omega)}{\partial \omega} \right|_{\omega = \omega_0} \simeq - \frac{N+1}{\pi}; \quad \omega_0 > \frac{\pi}{N+1}, \quad N > 10; \quad (5.7.4-11b)$$

Fejer ($\sigma_k = \frac{N+1}{N+1} |k|$)

$$\left. \frac{\partial \bar{H}_A(\omega)}{\partial \omega} \right|_{\omega = \omega_0} \simeq - \frac{N+1}{2\pi}; \quad \omega_0 > \frac{\pi}{N+1}, \quad N > 10; \quad (5.7.4-11c)$$

Cosine-squared ($\sigma_k = \cos^2 \frac{\pi k}{N+1}$)

$$\left. \frac{\partial \bar{H}_A(\omega)}{\partial \omega} \right|_{\omega = \omega_0} \simeq - \frac{N+1}{2\pi}; \quad \omega_0 > \frac{\pi}{N+1}, \quad N > 10. \quad (5.7.4-11d)$$

NOTE: The number of filter weights is $M = 2N + 1$.

It should be noted that the restrictions on the approximations so that they are valid, are also necessary for reasonable filter approximations as the terms that were neglected through use of these conditions vary wildly. Namely, rectangular filter design (and differentiators) should satisfy at least

$$N + 1 > \frac{\pi}{\omega_0} , \quad (\omega_0 \leq \frac{\pi}{2} , \quad \text{which is not really a restriction, see Section 5.9})$$

and

$$N \geq 10.$$

Smaller numbers of terms can be used in practice but the approximation must be numerically evaluated and examined carefully as to its utility.

From the slopes found above "fall-off", defined in Section 6.8.2, can be computed

$$\text{Fourier: } \frac{\Delta f}{f_s} = \frac{1}{2(N+1)}$$

NOTE:

$$\text{Fejer: } \frac{\Delta f}{f_s} = \frac{1}{N+1}$$

The number of filter weights
is $M = 2N + 1$
(5.7.4-12)

$$\text{Lanczos: } \frac{\Delta f}{f_s} = \frac{1}{1.18(N+1)}$$

Cosine-Squared:

$$\frac{\Delta f}{f_s} = \frac{1}{N}$$

It can be seen that the Fourier series provide the fastest "fall-off" with the "Lanczos" filter being the next best, but "falling-off" only about 1/2 as fast for a given N. Subsequently, it will be seen that the "ripple" error of the "Lanczos" and "cosine-squared" series (or filter) is remarkably small compared to the Fourier and Fejer series. Apparently, a factor of about two has to be paid in the number of weights to reduce the error significantly by the "filter-smoothing" technique for a fixed "fall-off".

5.7.5 Ripple Errors of "Filtered" Filters

Earlier it was found that the expansion of an ideal "cosine" low-pass filter with $2N + 1$ weights was given by

$$\bar{H}_A(\omega) = \frac{\omega_0}{\pi} + \frac{2}{\pi} \sum_{k=1}^N \sigma_k \frac{\sin \omega_0 k}{k} \cos \omega k \quad (5.7.4-5)$$

where ω_0 is the cut-off frequency.

Numerical search techniques have been utilized to find the first two "ripple" errors either side of the discontinuity, neglecting the finite "fall-off", are approximately equal for each method. In fact, they approach equality as N becomes large for each method. Various cut-off frequencies were used and the errors again approached the same limit, assuming N was large enough to obtain decent approximations ($N \omega_0 > \pi$). Finally, to ensure that errors did not increase as ω approached π , the error for the nearest ripple to π was calculated and found to be significantly smaller than the first two ripple errors. The limiting errors for the various methods are tabulated. Ripple errors for different N and ω_0 are substantially equal to those for large N and do not vary more than 10 to 20 percent (at small N) of the limiting forms for $N = 1000$. Values of ω_0 used were $\omega_0 = \pi/12$, $\pi/4$, and $\pi/2$ with half the sampling frequency considered as π . From relations in Section 5.9 it will be seen that ripple errors also are identically valid for $\omega_0 = 11\pi/12$ and $3\pi/4$.

Ripple errors normalized to percentages for the various weighting methods follow in Table 5.7.5-1.

ω_0	Fourier		Fejer		Lanczos		Cos^2	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd
$\pi/12$	8.9%	4.9%	4.9%	2.5%	1.2%	.48%	.64%	.19%
$\pi/4$	8.9	4.9	4.9	2.5	1.2	.48	.64	.19
$\pi/2$	8.9	4.9	4.9	2.5	1.2	.48	.64	.19
$3\pi/4$	8.9	4.9	4.9	2.5	1.2	.48	.64	.19
$11/12$	8.9	4.9	4.9	2.5	1.2	.48	.64	.19

TABLE 5.7.5-1 LOW-PASS FILTER RIPPLE ERRORS

It is obvious from table 5.7.5-1 that the "Lanczos" and " Cos^2 " methods produce usable errors (40 to 60 dB rejection out of the passband at the first two error ripples). The "Lanczos" method gives a slightly better fall-off though the " Cos^2 " method has better ripple deduction by about a factor of 2(6dB). It should be remarked that the first two error ripples in the passband have the same values as those out of the passband for each technique. Figures 5.10-1 through 5.10-6 demonstrate the efficacy of the various techniques graphically and show clearly the advantages of the "Lanczos" and Cos^2 techniques.

5.8 EXPANSION COEFFICIENTS AND WEIGHTS FOR THE DISCONTINUITY REMOVAL FUNCTIONS (OR "CENTRAL" BANDPASS AND DIFFERENTIATION FILTERS)

In this section the Fourier expansions of the important functions $u(\omega)$, $v(\omega)$, and, ω which can be used to remove discontinuities from transfer functions or by themselves as bandpass filters or differentiators will be given. The weighting methods of the previous section should be used to remove the Gibbs phenomena.

Furthermore, they will be given as central filters ($\tau = N$), which automatically produces a delay of NT . Both "Sine" and "Cosine" versions of the functions, or filters, will be stated except for the differentiator ($H(\omega) = i\omega$) which is given only in

"sine" form.

5.8.1 "Cosine" Low-Pass Filter ($u(\omega)$, Figure 5.6.1-1b)

The ideal filter is given by

$$H(\omega) = u(\omega),$$

$$= \begin{cases} 1, & 0 < |\omega| < \omega_0, \\ 0, & \omega_0 < |\omega| < \omega_s/2. \end{cases} \quad (5.8.1-1)$$

The Fourier approximation is

$$H_A(\omega) = \sum_{n=-N}^N c_n e^{-i\omega nT}, \quad (5.8.1-2)$$

with

$$c_n = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H(\omega) e^{i\omega nT} d\omega.$$

(5.8.1-3)

$$= \begin{cases} \frac{\omega_0 T}{\pi} \frac{\sin \omega_0 nT}{\omega_0 nT}, & n \neq 0, \\ \frac{\omega_0 T}{\pi}, & n = 0. \end{cases}$$

The corresponding non-recursive filter weights are given by

$$w_{N+n} = c_n, \quad n = -N, -(N-1), \dots, 0, \dots, N \quad (5.8.1-4)$$

When applying σ factors the weights are

$$w_{N+n} = \sigma_n c_n; \begin{cases} c_{-n} = c_n, \\ \sigma_{-n} = \sigma_n. \end{cases} \quad (5.8.1-5)$$

5.8.2 "Cosine" High Pass Filter ($v(\omega)$, Figure 5.6.1-1a)

The ideal filter is given by

$$H(\omega) = v(\omega) = \begin{cases} 0, & 0 \leq |\omega| \leq \omega_0, \\ 1, & \omega_0 < |\omega| \leq \omega_s/2. \end{cases} \quad (5.8.2-1)$$

As $v(\omega)$ is simply $1-u(\omega)$ its expansion can immediately be written using coefficients and weights from the preceding section

$$c_n = \begin{cases} \frac{-\omega_0 T}{\pi} \frac{\sin \omega_0 n T}{\omega_0 n T}, & n \neq 0 \\ 1 - \frac{\omega_0 T}{\pi}, & n = 0 \end{cases} \quad (5.8.2-2)$$

The filter weights are

$$w_{N+n} = c_n, \quad n = -N, -(N-1), \dots, 0, \dots, N; \quad (5.8.2-4)$$

5.8.3 "Sine" Low Pass Filter ($u(\omega)$, Figure 5.6.1-1b)

Again the ideal filter is

$$H(\omega) = iu(\omega) = \begin{cases} i, & 0 < \omega \leq \omega_0, \\ -i, & -\omega_0 \leq \omega < 0, \\ 0, & \text{otherwise.} \end{cases} \quad (5.8.3-1)$$

The corresponding Fourier coefficients are

$$c_n = \begin{cases} \frac{\omega_0 T}{\pi} \frac{\cos \omega_0 n T}{\omega_0 n T}, & n \neq 0, \\ 0 & , \quad n = 0. \end{cases} \quad (5.8.3-2)$$

The filter weights are

$$w_{N+n} = c_n, \quad n = -N, -(N-1), \dots, 0, \dots, N \quad (5.8.3-3)$$

and with weighting factors:

$$w_{N+n} = \sigma_n c_n; \quad \begin{cases} c_{-n} = -c_n, \\ \sigma_{-n} = \sigma_n. \end{cases} \quad (5.8.3-4)$$

5.8.4 "Sine" High Pass Filter ($v(\omega)$, Figure 5.6.1-1a)

Here the ideal filter is

$$H(\omega) = i v(\omega)$$

$$= \begin{cases} i, & \omega_0 \leq \omega \leq \frac{\omega_s}{2}, \\ -i, & \frac{-\omega_s}{2} \leq \omega \leq \omega_0, \\ 0, & \text{otherwise.} \end{cases} \quad (5.8.4-1)$$

The Fourier coefficients are

$$c_n = \begin{cases} \frac{\omega_0 T}{\pi} \frac{1 - \cos \omega_0 n T}{\omega_0 n T}, & n \neq 0, \\ 0 & , \quad n = 0. \end{cases} \quad (5.8.4-2)$$

The filter weights are given by

$$w_{N+n} = c_n, \quad n = -N, -(N-1), \dots, 0, \dots, N, \quad (5.8.4-3)$$

and with weighting factors

$$w_{N+n} = \sigma_n c_n, \quad \begin{cases} c_{-n} = -c_n, \\ \sigma_{-n} = \sigma_n. \end{cases} \quad (5.8.4-4)$$

5.8.5 "Sine" Differentiating Filters ($i\omega$)

The differentiator is defined only to some frequency $\omega_0 \leq \omega_s/2$ as differentiators to $\omega_s/2$ have Gibbs phenomena ripples of twice their usual amplitude due to the discontinuity of twice the differentiator value at $\omega_s/2$. Of course, differentiators can be designed to $\omega_s/2$, if need be, though the error increase must be tolerable. In practice, this error increase can be made less significant when Lanczos or cosine-squared σ factors are used.

Differentiators are defined as

$$H(\omega) = \begin{cases} i\omega, & 0 \leq |\omega| \leq \omega_0 \\ 0, & \text{otherwise.} \end{cases} \quad (5.8.5-1)$$

The Fourier approximation coefficients are

$$c_n = \begin{cases} -\frac{2\omega_0^2}{\omega_s} \left[\frac{\cos \omega_0 nT}{\omega_0 nT} - \frac{\sin \omega_0 nT}{(\omega_0 nT)^2} \right], & n \neq 0, \\ 0, & n = 0, \end{cases} \quad (5.8.5-2)$$

Filter weights are given by

$$w_{N+n} = c_n, \quad n = N, -(N-1), \dots, 0, \dots, N. \quad (5.8.5-3)$$

With weighting factors the filter weights are

$$w_{N+n} = \sigma_n c_n; \quad \begin{cases} c_{-n} = -c_n, \\ \sigma_{-n} = \sigma_n. \end{cases} \quad (5.8.5-4)$$

5.9 CERTAIN NON-RECURSIVE FILTER GENERALITIES

Before presenting the graphical results (Section 5.10) of certain rectangular filter designs, it is advantageous to understand a few general relations among non-recursive filter types. These relations allow consideration of the designs of a few cases and generalization to many other cases. This reduces error analysis of filters substantially to the knowledge of the behavior of typical cases. Also, if need be, the design of "related" filters can be obtained directly.

5.9.1 Cosine Filters (Low and High Pass)

Low-and high-pass filters considered as even "delayed" non-recursive transfer functions ($\tau = N$) in the frequency domain involve only the real part of the transfer function. A typical, even ideal low-pass filter is depicted in Figure including the periodic extension to $\omega = 3\pi$.

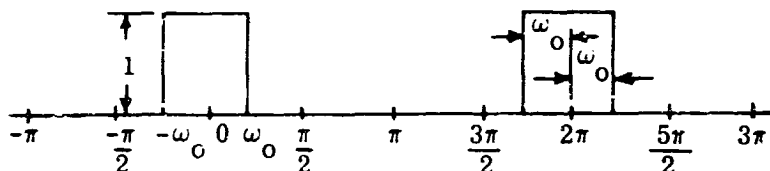


Figure 5.9.1-1. An "Even" Ideal Low-Pass Filter

From analysis of the Figure 5.9.1-1 and recognition that the transfer function is periodic and real, it can be seen that translation of the Fourier expansion (either truncated or filtered as in Section (5.6)) of the low-pass filter by π gives a high-pass filter whose cut-off frequency is at $\pi - \omega_0$.

That is, if a cosine approximation to $H(\omega)$ exists, namely $\bar{H}_A(\omega)$ (or $H_A(\omega)$) of the form

$$\begin{aligned} \bar{H}_A(\omega) |_{LP} &= \sum_{k=-N}^N r_k e^{-i\omega k} \\ &= r_0 + 2 \sum_{k=1}^N r_k \cos \omega k \end{aligned} \quad (5.9.1-1)$$

for a low-pass filter whose cut-off is at ω_0 ; then,

$$\begin{aligned} \overline{H}_A(\omega - \pi) \Big|_{LP} &= \sum_{k=-N}^N r_k e^{-i\omega k} e^{i\pi k} \\ &= \sum_{k=-N}^N (-1)^k r_k e^{-i\omega k} \end{aligned} \quad (5.9.1-2)$$

or

$$\overline{H}_A(\omega) \Big|_{HP} = r_0 + 2 \sum_{k=1}^N (-1)^k r_k \cos k\omega \quad (5.9.1-3)$$

is the approximation to the high-pass filter whose cut-off is at $\pi - \omega_0$.

Error analysis in the pass-band of the low-pass filter gives identical results with the error analysis in the pass-band of the high pass. In fact the two filters are simply reflections about the line $\omega = \frac{\pi}{2}$. Stop-band analyses are also identical and reflected.

From the analytical point of view, analysis and synthesis of high-pass filters is equivalent to that of low-pass filters.

Similarly, it can be shown that analysis of the low-pass filter in the region $0 \leq \omega \leq \frac{\pi}{2}$ (ω_0 = cut-off frequency in radians) is sufficient to determine the behavior of a filter (low-pass) whose cut-off lies in $\frac{\pi}{2} < \omega_0 \leq \pi$. From Figure 5.9.1-2 it can be that the low-pass filter whose cut-off

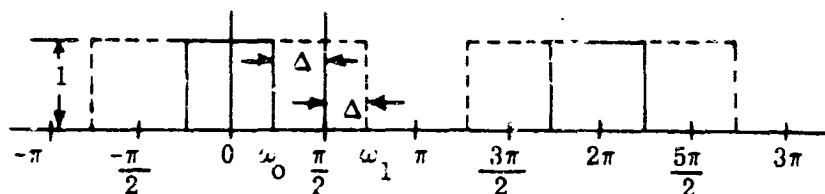


Figure 5.9.1-2. An "Image" Low-Pass Filter

at $\omega = \omega_0 = \frac{\pi}{2} - \Delta$ can be related to the low-pass filter whose cut-off is at $\omega_1 = \frac{\pi}{2} + \Delta$. The relation is relatively simple:

1. Translate the approximation for the low pass to π giving a high-pass filter (as was done in preceding paragraphs) whose cutoff is at $\pi - \omega_0 = \frac{\pi}{2} + \Delta$, namely, the cutoff is at

$$\omega_1 = \frac{\pi}{2} + \Delta$$

2. Subtract this high pass from an "all-pass" filter ($H(\omega) = 1$ for all ω).

The result is the low-pass filter with a cut-off at $\omega_1 = \frac{\pi}{2} + \Delta$. Therefore, analysis of low-pass filters for $\omega \leq \frac{\pi}{2}$ is sufficient to determine the behavior of filters for $\frac{\pi}{2} \leq \omega_1 < \pi$. A bit of thought will show that ripples in the pass-band of the low pass for $\omega_0 = \frac{\pi}{2} - \Delta$ become the ripples in the stop-band for the low-pass filter with cut-off at $\omega_0 = \frac{\pi}{2} + \Delta$ behave similarly. In other words the ω_1 filter is a reflection about $\omega = \frac{\pi}{2}$ plus a reflection about $H(\omega) = \frac{1}{2}$.

Mathematically, if $\overline{H}_A(\omega)$ is the transfer function for the low pass whose cut-off is at $\omega_0 = \frac{\pi}{2} - \Delta$, the transfer function of the low pass at $\omega_1 = \frac{\pi}{2} + \Delta$ is

$$\begin{aligned} \overline{H}_A(\omega) \Big|_{\omega_1} &= 1 - \overline{H}_A(\omega) \Big|_{\omega_0} \\ &= (1 - r_0) + 2 \sum_{k=1}^N (-1)^{k+1} r_k \cos k\omega \quad (5.9.1-4) \end{aligned}$$

where

$$\overline{H}_A(\omega) \Big|_{\omega_0} = r_0 + 2 \sum_{k=1}^N r_k \cos k\omega. \quad (5.9.1-5)$$

5.9.2 Sine Filters (Low and High) Pass

Low-and high-pass filters considered as odd functions in the frequency domain involve only the imaginary part of the standard transfer function. A typical odd ideal low-pass filter is depicted in Figure 5.9.2-1 including the periodic extension to $\omega = \pi$.

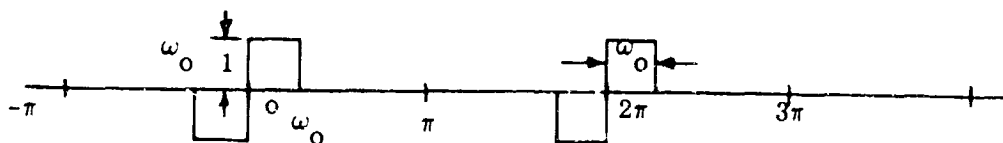


Figure 5.9.2-1. An "Odd" Low-Pass Filter

As the low-pass filter is odd, it cannot pass DC; namely $H(0) = 0$. Also, a discontinuity of 2 units exists at $\omega = 0$, which means that a straight truncated Fourier expansion would have approximately an 18 percent overshoot at $\omega = 0+$ where $0+$ is the first maximum. (See Section 5.2.)

As was done for cosine filters in the previous section high-pass filters can be generated from low-pass filters. That is using the notation of the previous section, it can be seen by examining Figure 5.9.2-1 that a high-pass filter of cut-off at $\pi - \omega_0$ can be generated from the low-pass at ω_0 by translating the low pass to π and reflecting about $H(\omega) = 0$. That is,

$$\bar{H}_A(\omega) \Big|_{\text{HP}} = -\bar{H}(\omega - \pi) \Big|_{\text{LP}}, \quad (5.9.2-6)$$

where the notation is obvious.

As before, if a useful approximate expansion exists for the odd low-pass filter with a cut-off at ω_0 :

$$\begin{aligned} \bar{H}_A(\omega) \Big|_{\text{LP}} &= \sum_{k=-N}^N r_k e^{-i\omega k} \\ &= -2i \sum_{k=1}^N r_k \sin \omega k; \end{aligned} \quad (5.9.2-7)$$

then, the transfer function of the equivalent high pass at $\omega = \pi - \omega_0$ is

$$\bar{H}_A(\omega) \Big|_{\text{HP}} = -2i \sum_{k=1}^N (-1)^{k+1} r_k \sin \omega k \quad (5.9.2-8)$$

Ripple behavior analysis in the pass bands of both the low-pass and equivalent high pass are identical, as are those in the stop bands. It can be seen, with a bit of thought, that the two filters are really reflections of each other about the line $\omega = \frac{\pi}{2}$.

5.9.3 Bandpass Filters (Sine and Cosine)

Both sine and cosine bandpass and band stop filters can be generated by simply taking high-pass filters and subtracting their expansions (or weights). That is, from a useful high pass filter at frequency (angular) ω_0 and another at ω_1 , a useful bandpass filter whose pass band extends from ω_0 to ω_1 can be obtained.

Of course, two sine high pass filters must be used if an odd band-pass filter is wanted and two cosine high pass filters if an even bandpass filter is desired. It is not necessary that there be the same number of terms in each filter, but error ripples are usually reduced if both high pass filters have the same number of terms, which is that of the filter with the highest number of weights.

The error analysis of the band-pass filter is related to that of the two high-pass filters. The error ripples interfere and can at worst be the sum of the errors of the two high pass filters at a given frequency.

As the errors usually drop-off rapidly away from the points of discontinuity the errors of the bandpass filters are usually not significantly worse than that of the high-pass filters from which they are generated. The errors are only affected by 10 to 20 percent. The exceptional case occurs when ω_0 is approximately equal to ω_1 , a narrow-band filter, then the maximum errors add. Examination of the figures of the next section above will convince the reader of these facts more easily than any laborious mathematics.

The foregoing means that wide band-pass filter errors are equivalent to the errors of the high pass used with the worst errors. Narrow band filter errors are approximately twice those of the two high pass filters used to synthesize it, and occur in the passband.

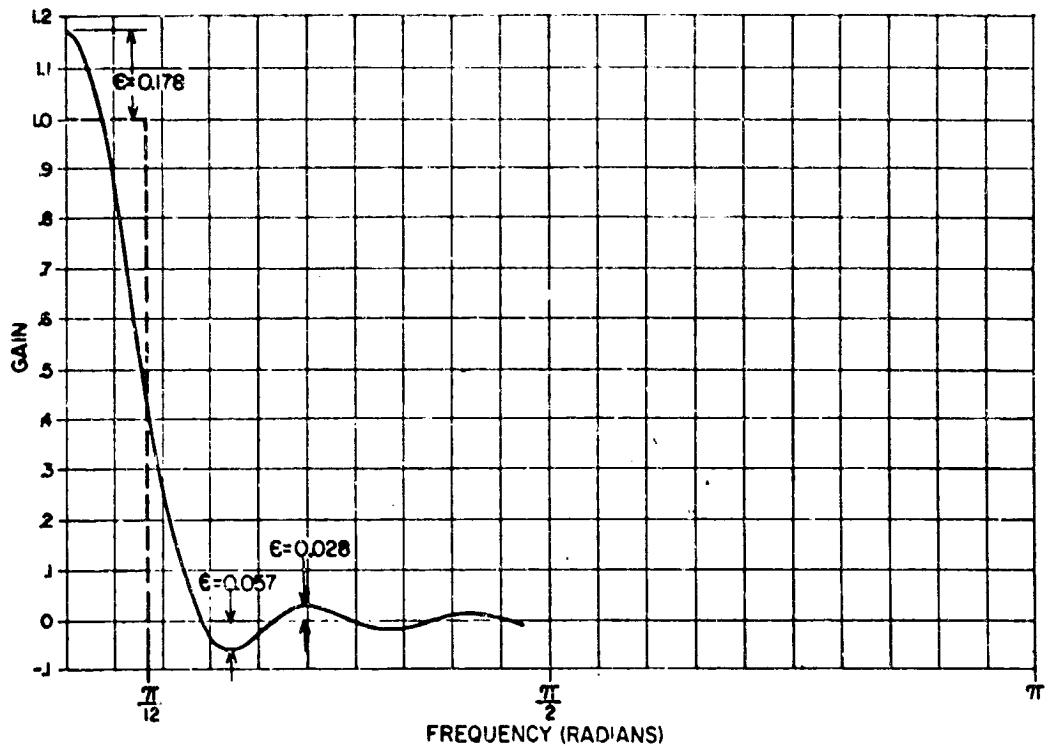
5.10 GRAPHICAL PRESENTATION OF NON-RECURSIVE FILTER DESIGNS

In this Section the results of designs using the various smoothing techniques are presented.

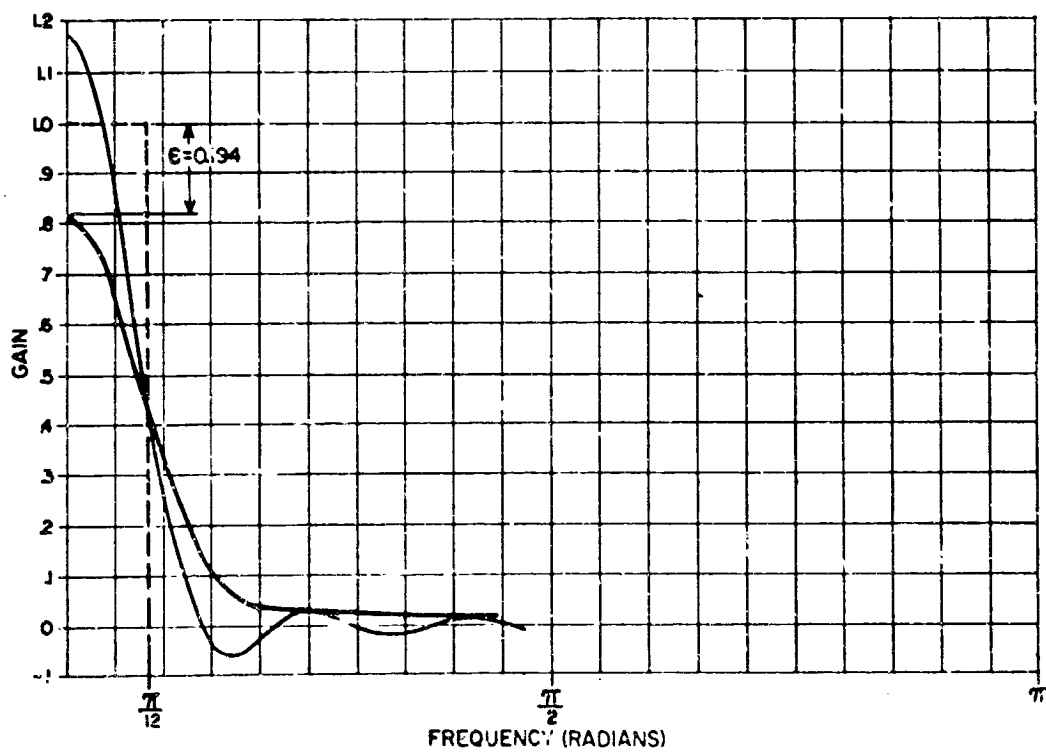
Figures 5.10-1 through 5.10-6 illustrate various central cosine and sine filters designed using the weights of Section 5.8 and the smoothing ("filter-filter") factors of Section 5.7. The number of filter weights of these filters is 25 (cosine) or 24 (sine) as $N = 12$ or $M = 25$ was chosen for illustrative purposes.

The "cosine" and "sine" filters demonstrate a number of pertinent points made in the previous sections. These points are

1. The number of filter weights must be such that a decent approximation can be made (see low-pass approximations for $\omega_0 = \pi/12$).
2. Significant improvement in error reduction can be made at the price of halving slopes by the various "filter-filter" methods.
3. Sine filters that extend to $\omega = \omega_s/2$ have a large error there (about twice) due to the "double" discontinuity caused by the "oddness" of sine filters.

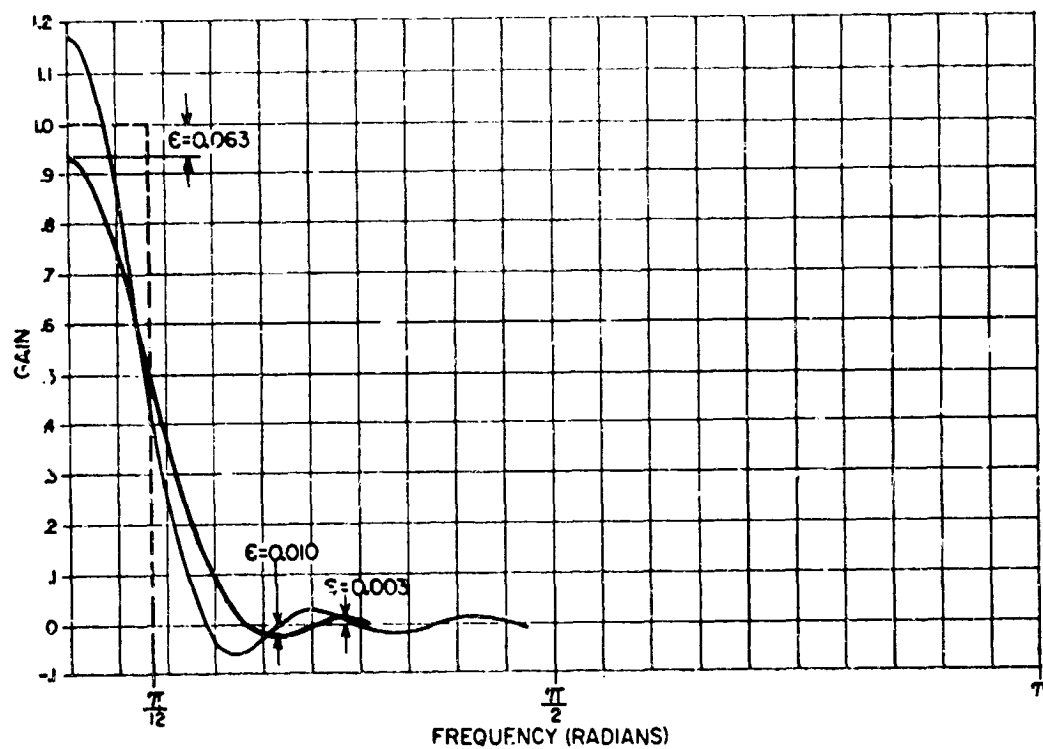


(a) Fourier Weights

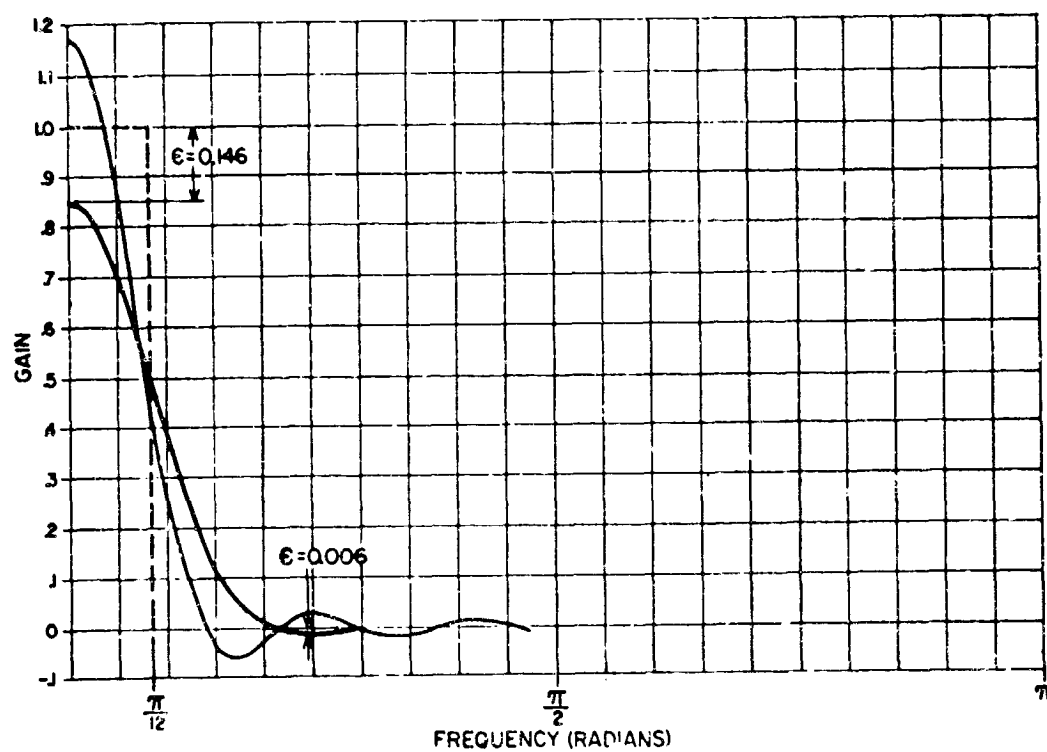


(b) Fejer Weights

Figure 5.10-1. Cosine Filter: Low-Pass ($\omega_c = \pi/12$)

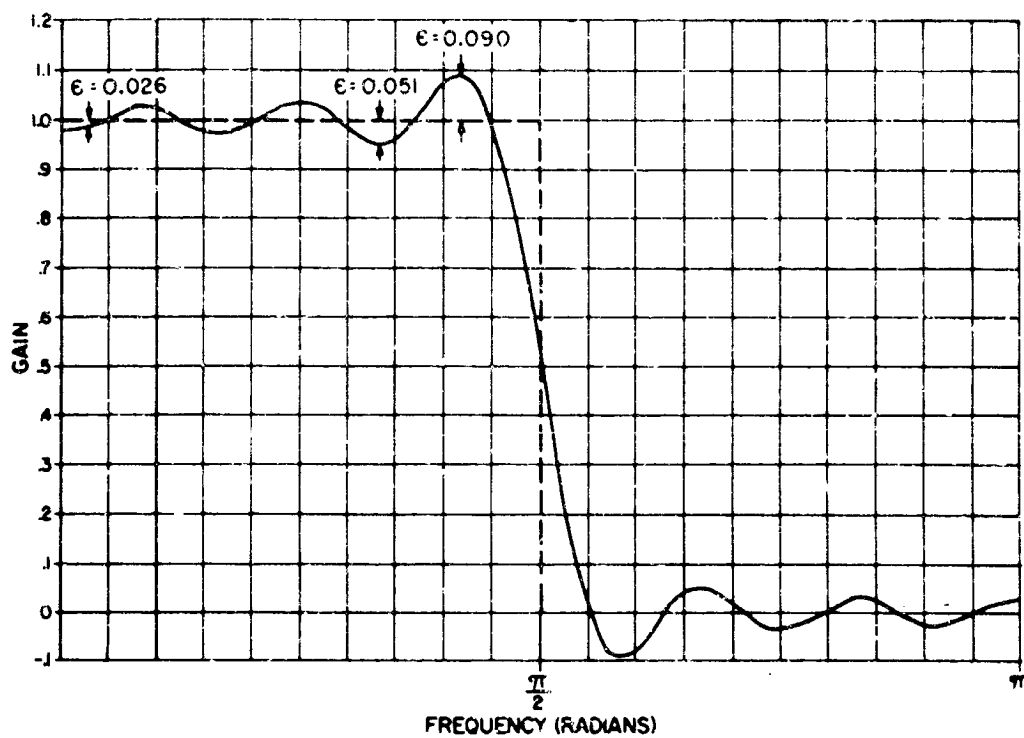


(c) Lanczos Weights

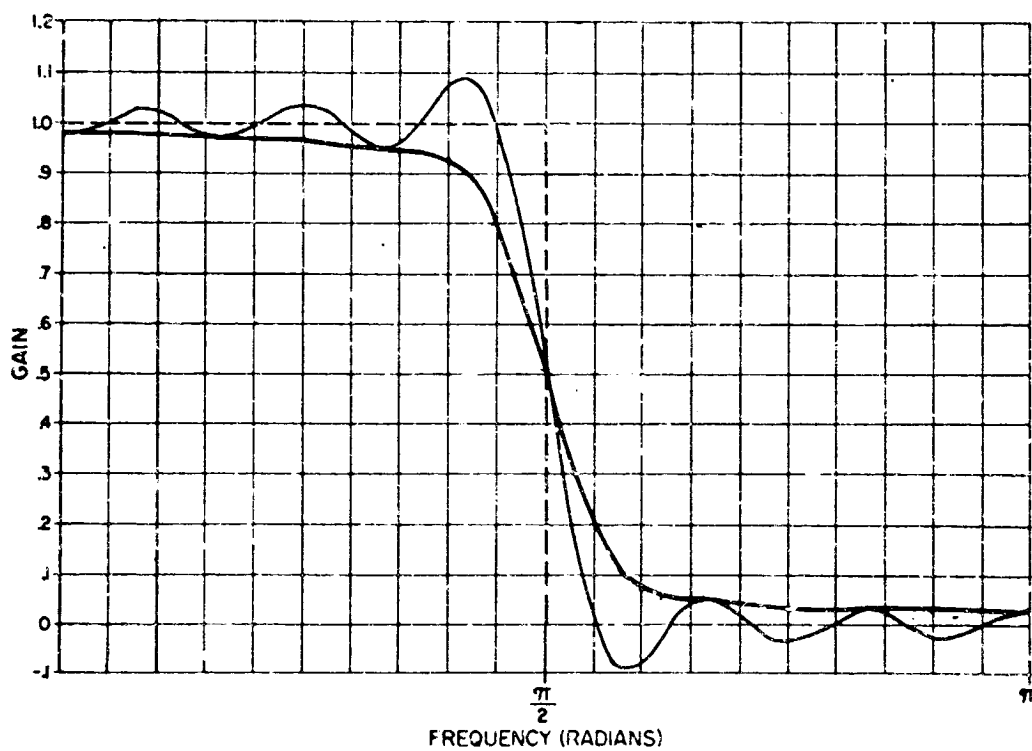


(d) Cosine-Squared Weights

Figure 5.10-1 (Continued)

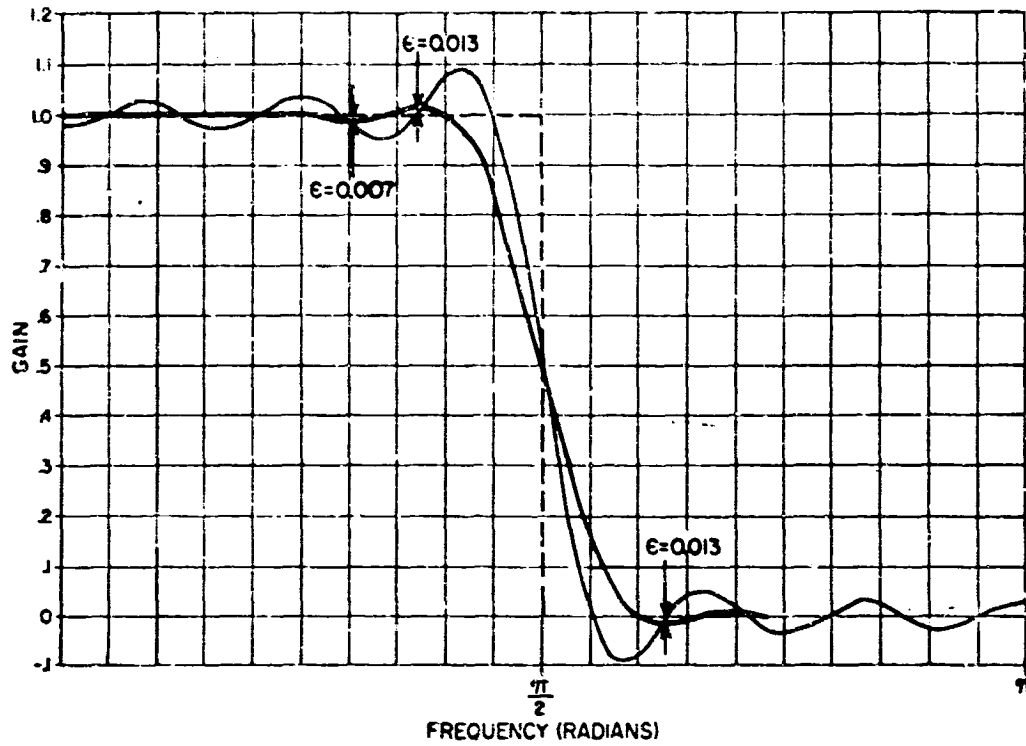


(a) Fourier Weights

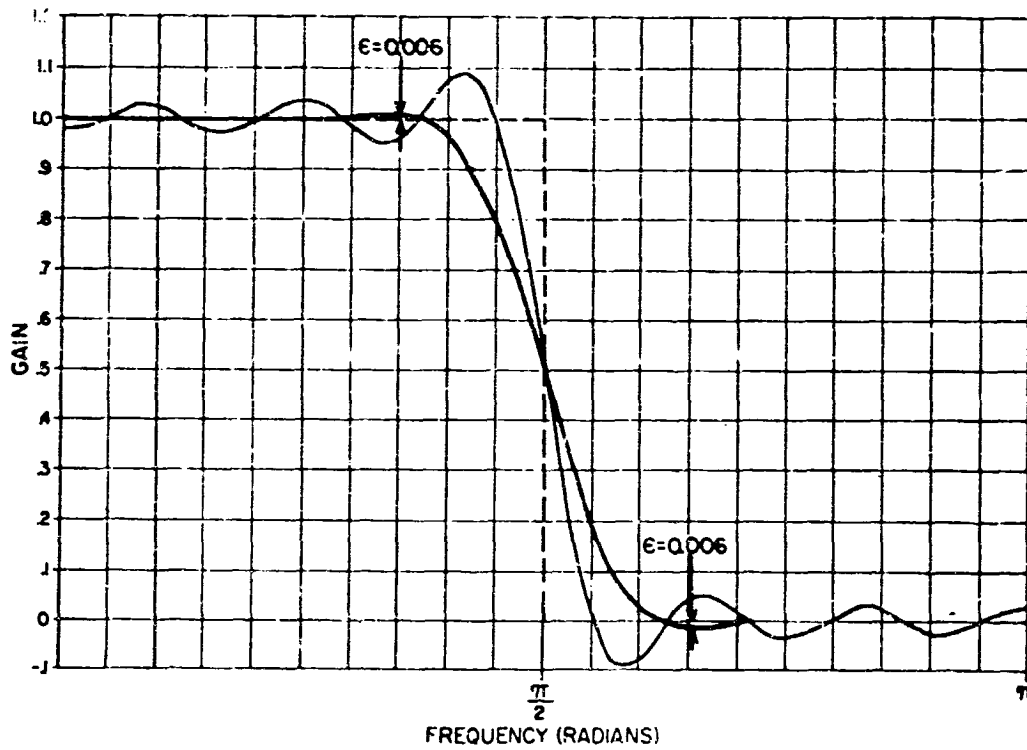


(b) Fejer Weights

Figure 5.10-2 Cosine Filter: Low-Pass ($\omega_0 = \pi/2$) $N = 12$

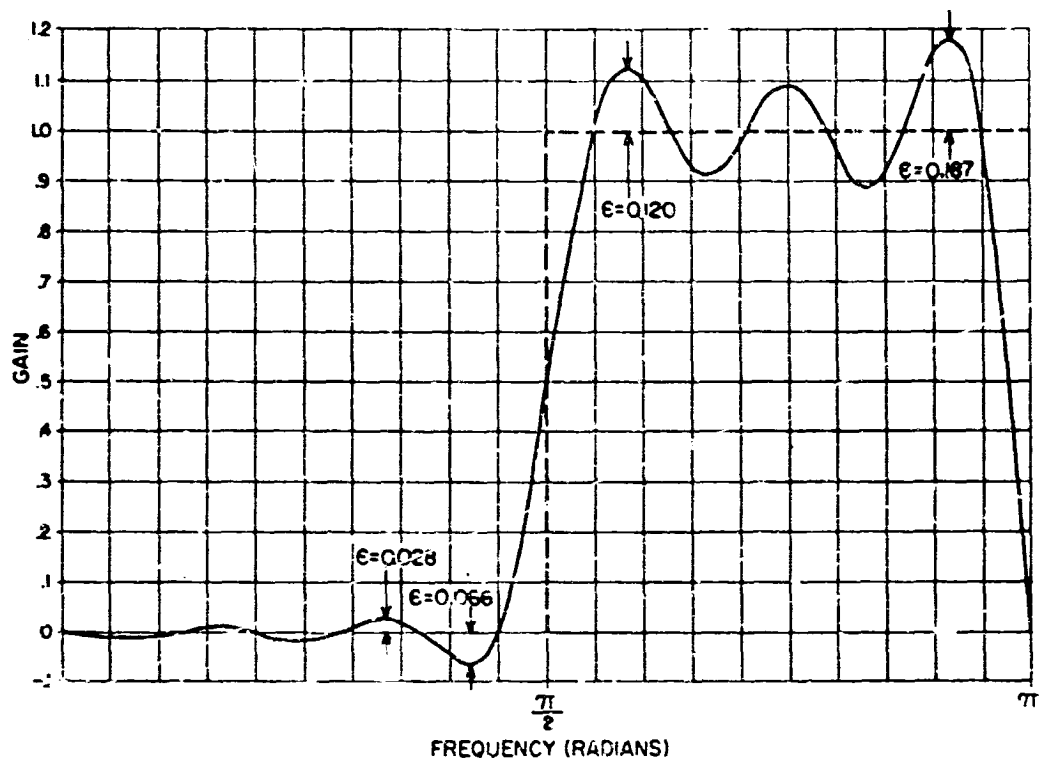


(c) Lanczos Weights

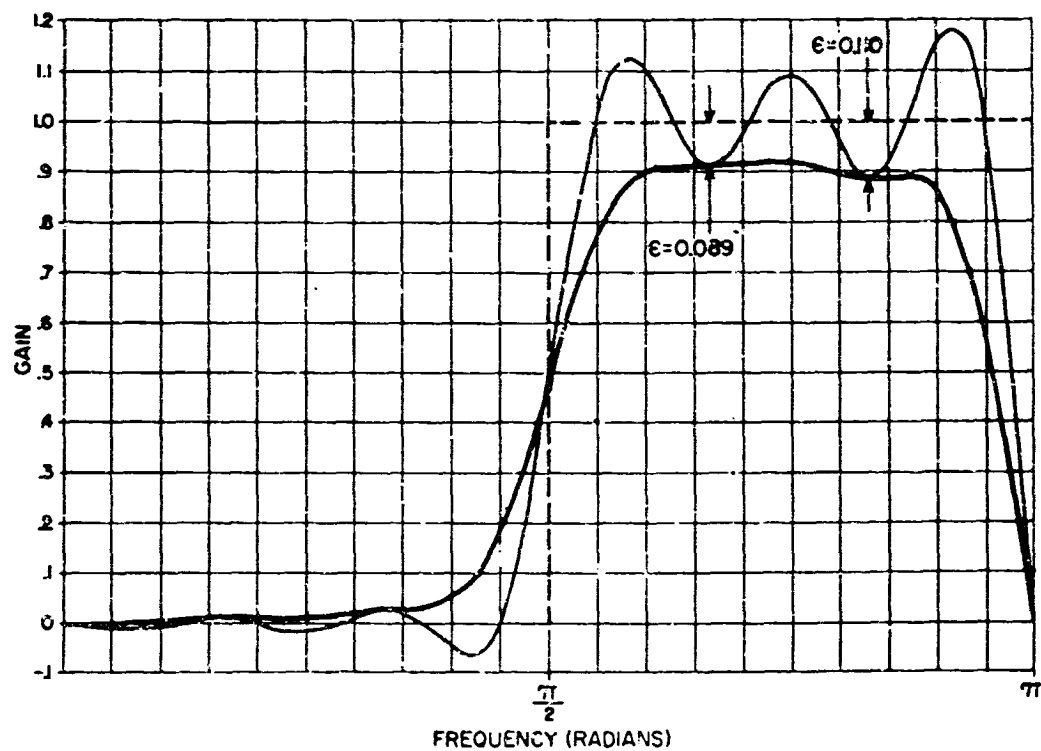


(d) Cosine-Squared Weights

Figure 5.10-2. (Continued)

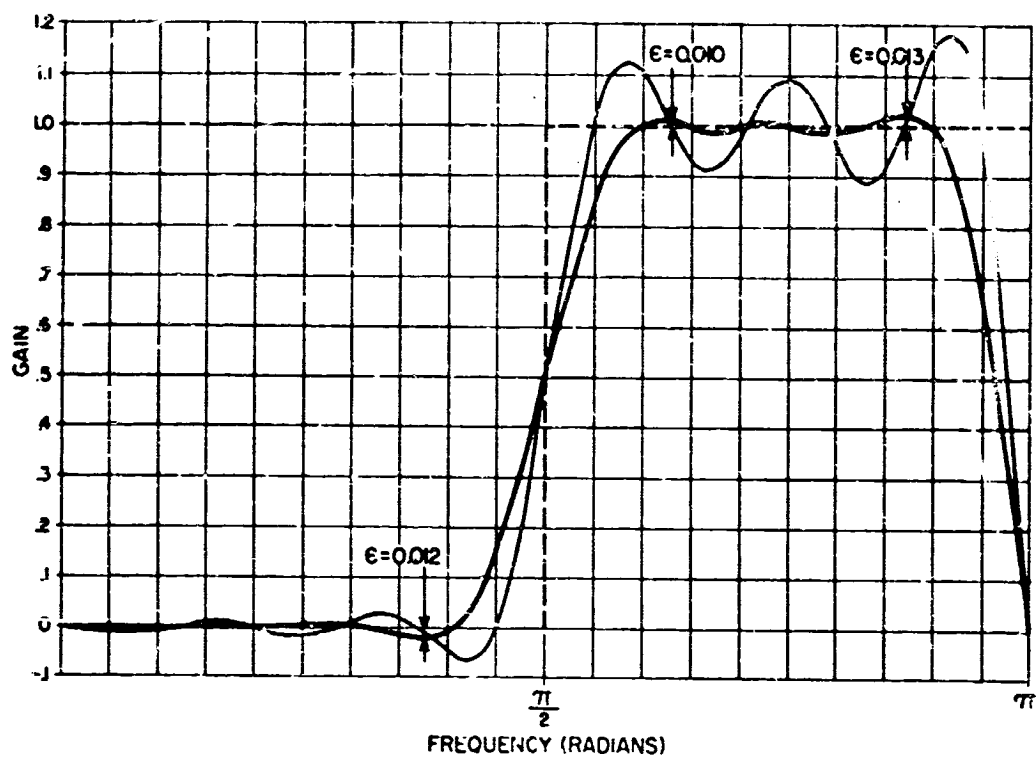


(a) Fourier Weights

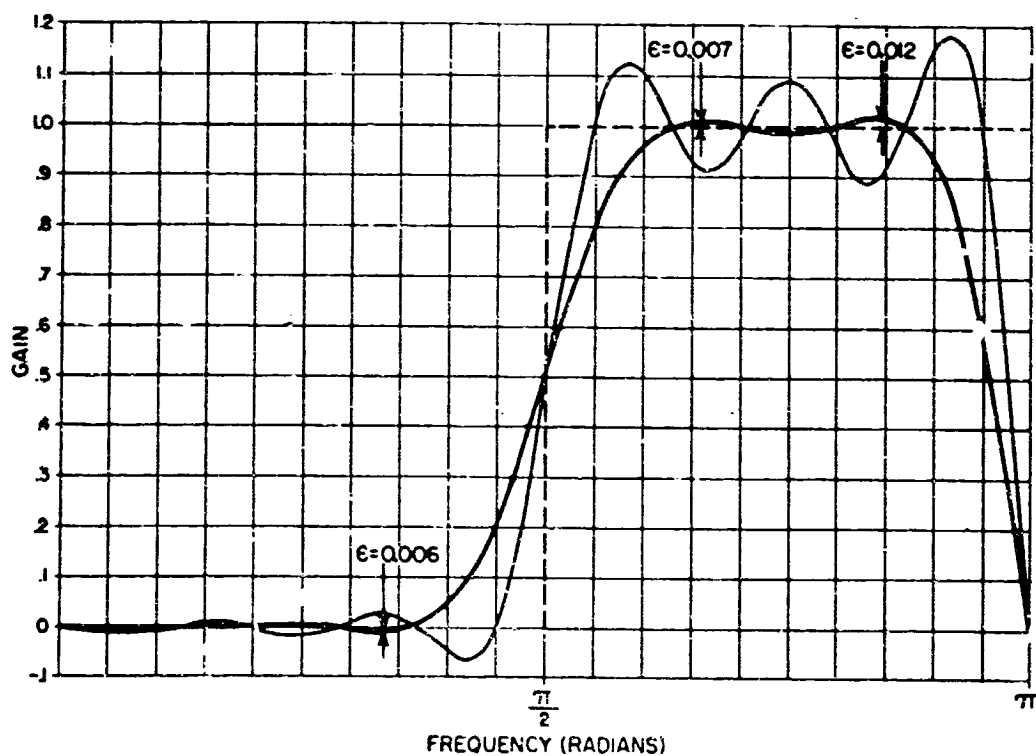


(b) Fejer Weights

Figure 5.10-3. Sine Filter: High Pass ($\omega_0 = \pi/2$) $N = 12$

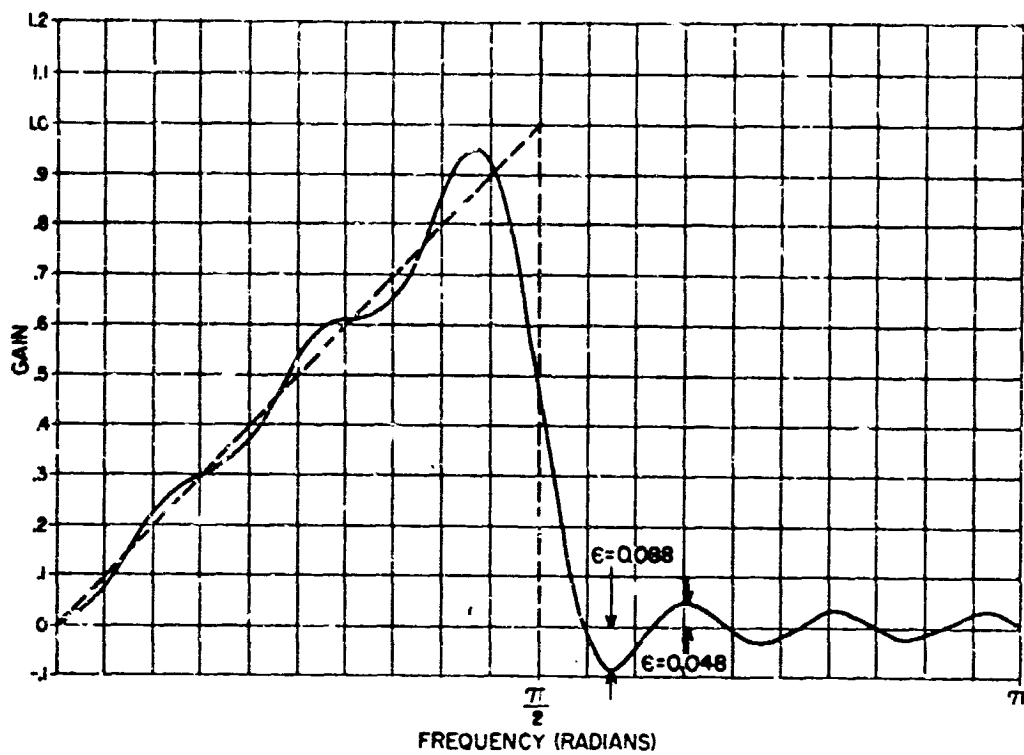


(c) Lanczos Weights

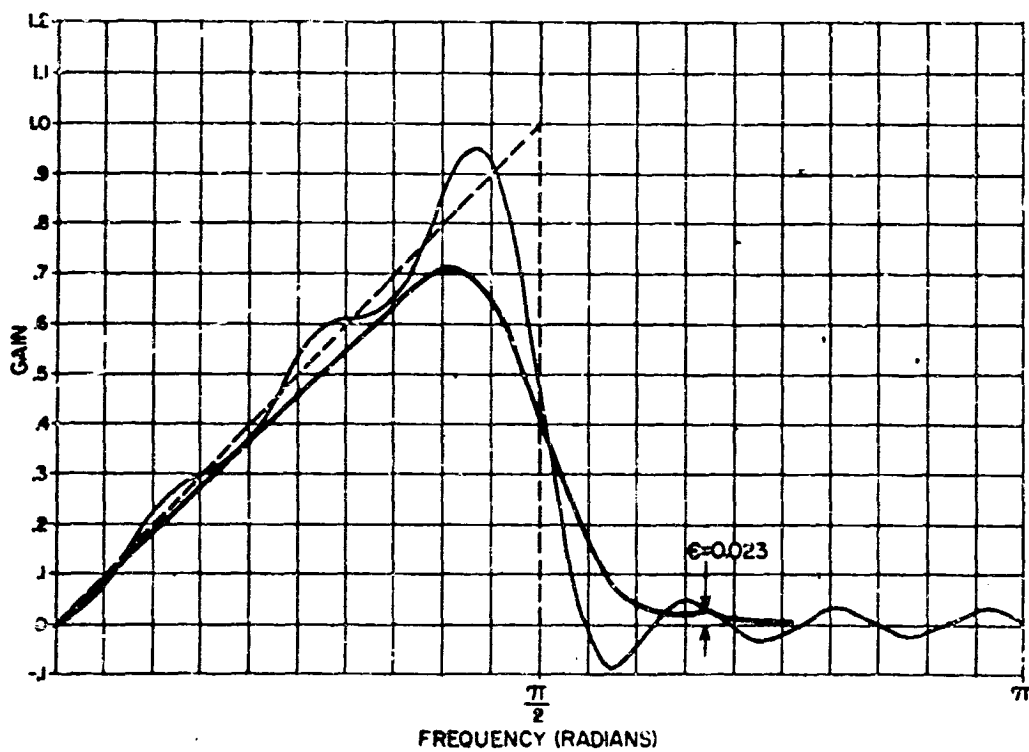


(d) Cosine-Squared Weights

Figure 5.10-3. (Continued)

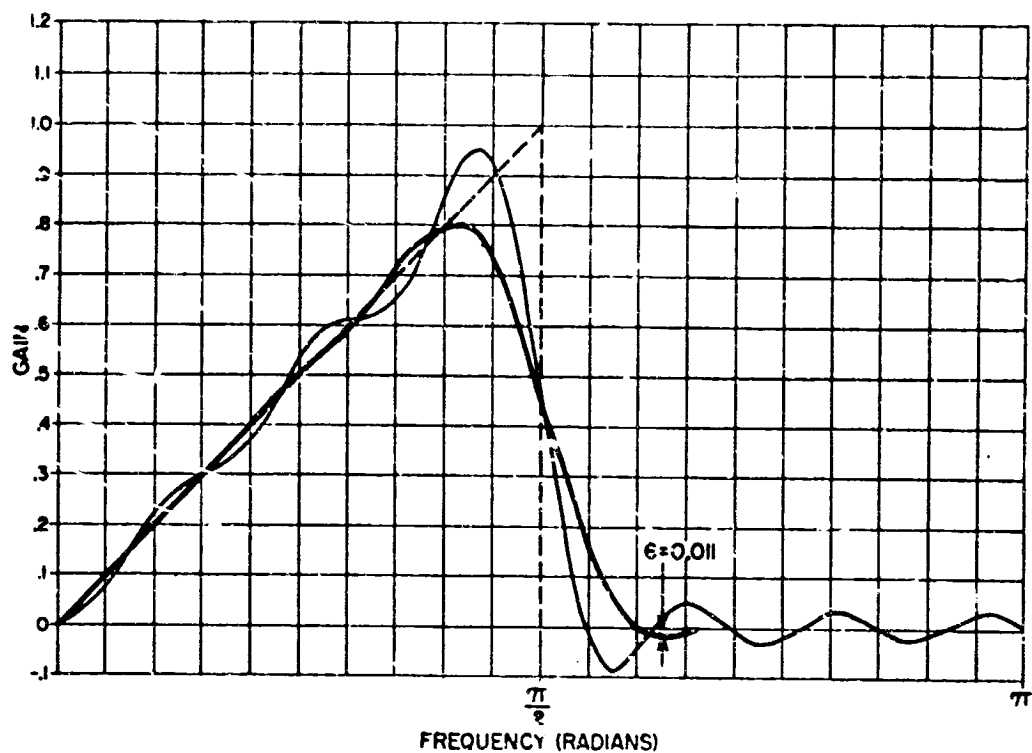


(a) Fourier Weights

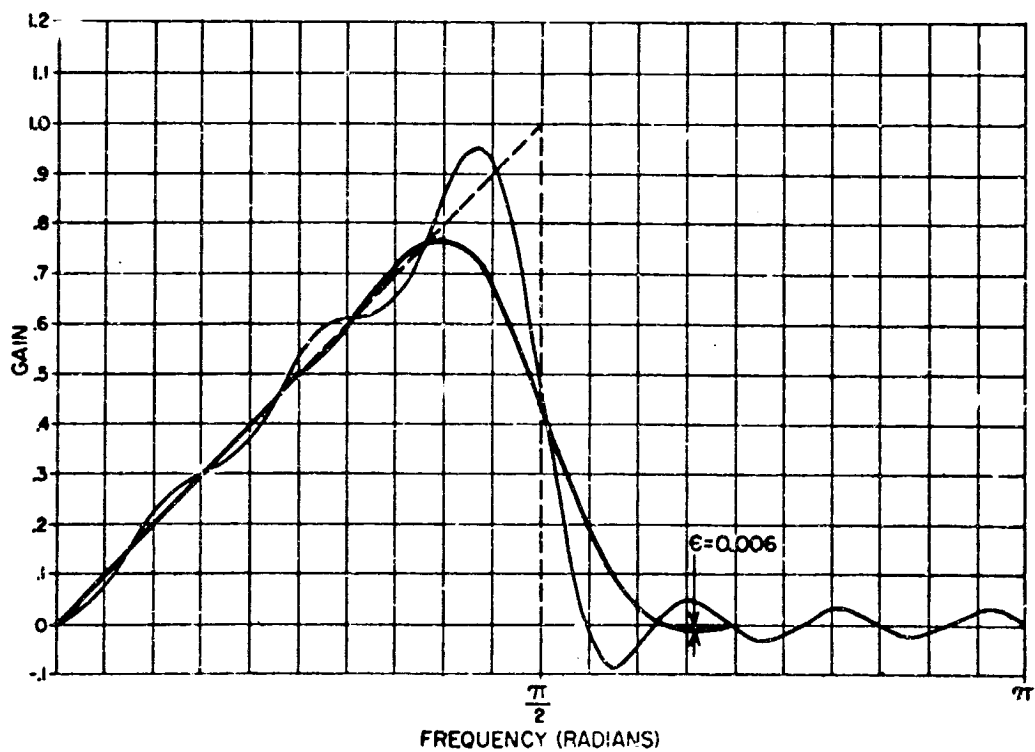


(b) Fejer Weights

Figure 5.10-4. Sine Filter: Differentiator ($\omega_0 = \pi/2$) $N = 12$

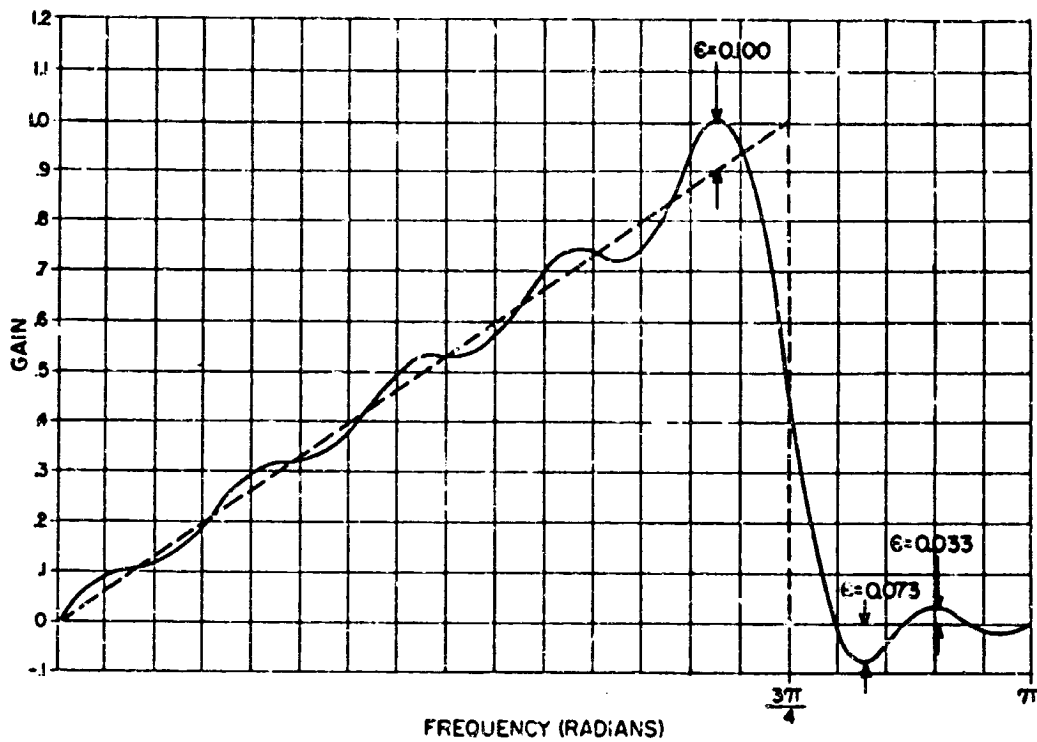


(c) Lanczos Weights

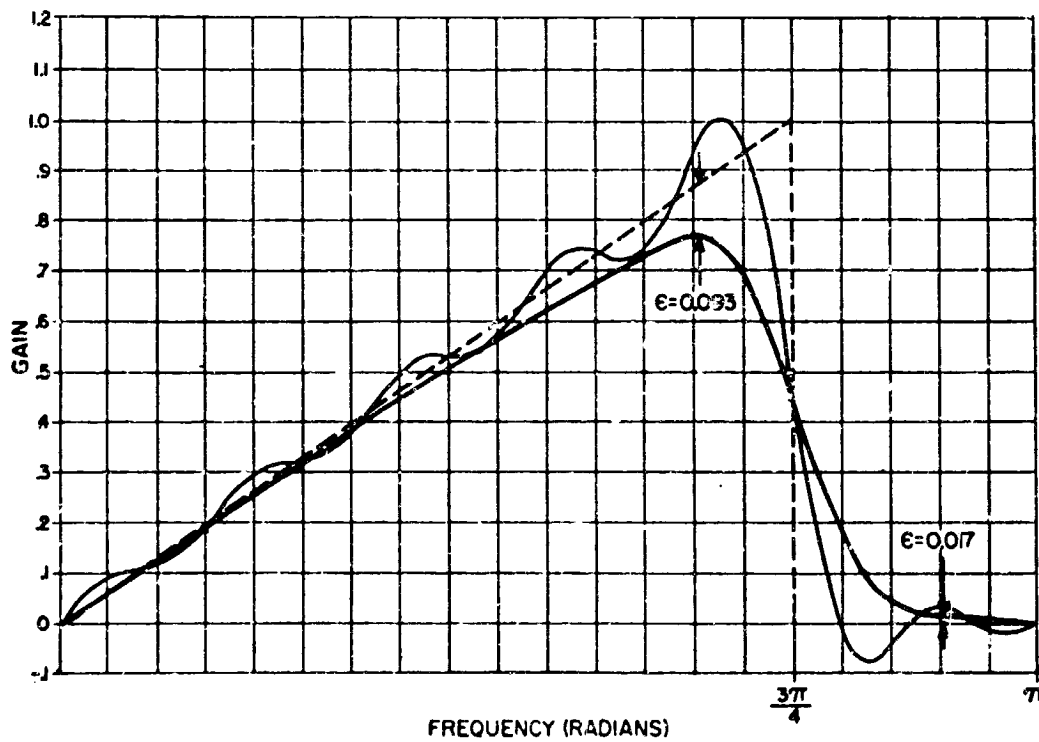


(d) Cosine-Squared Weights

Figure 5.10-4. (Continued)

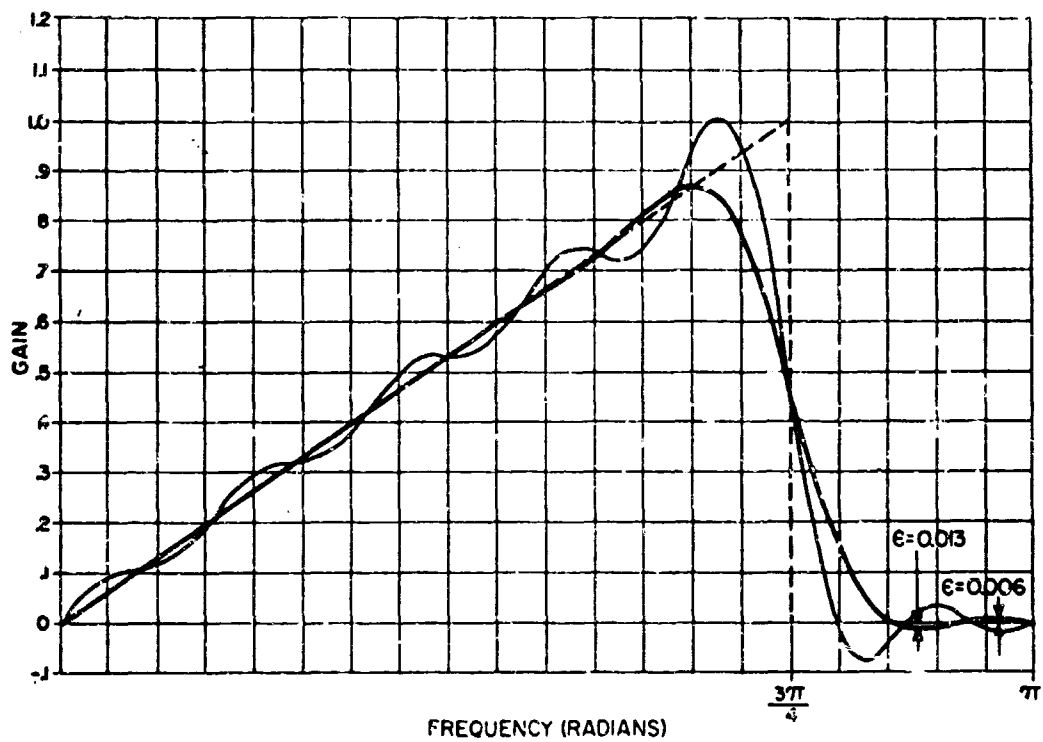


(a) Fourier Weights

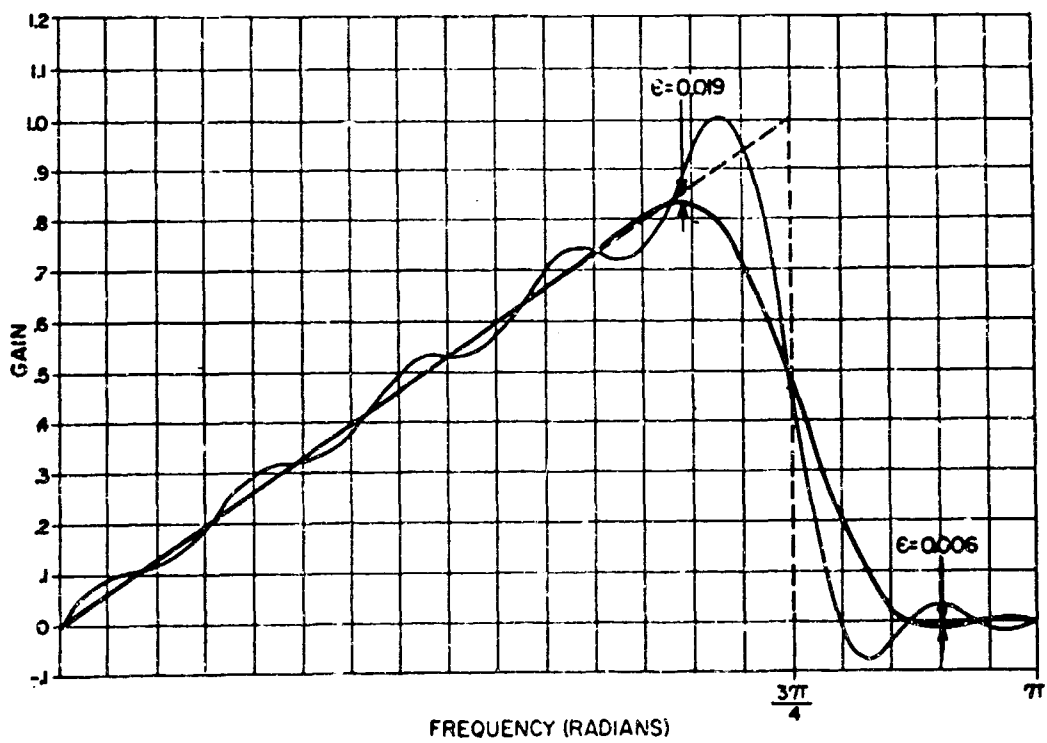


(b) Fejer Weights

Figure 5.10-5. Sine Filter: Differentiator ($\omega_0 = 3\pi/4$) $N = 12$

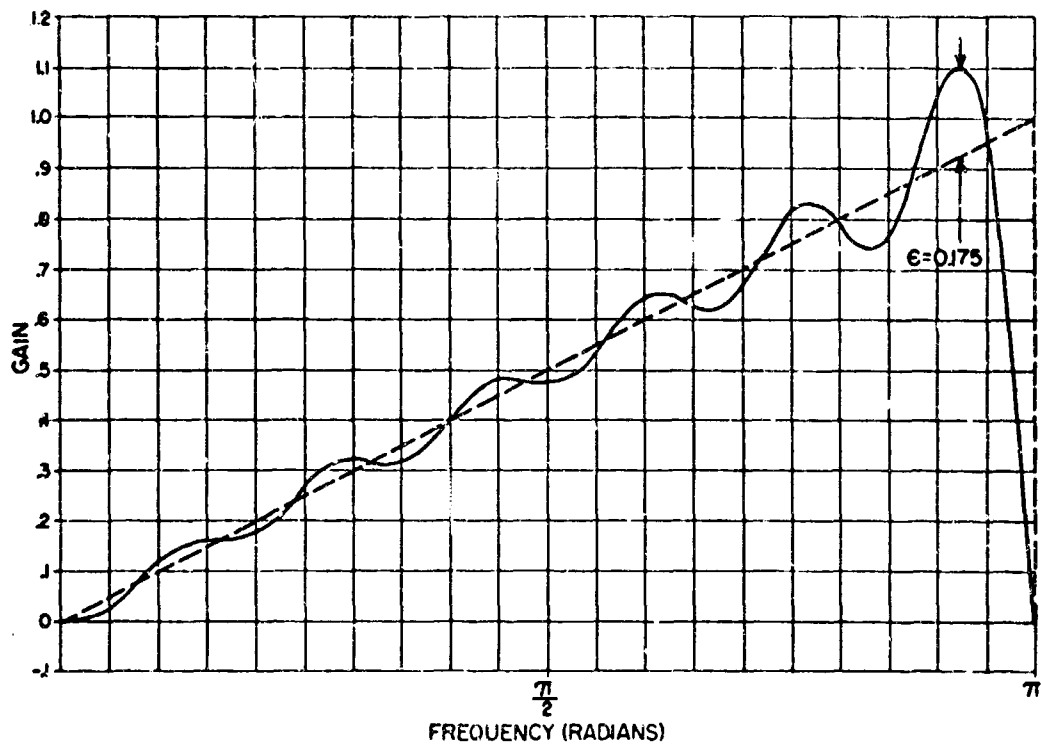


(c) Lanczos Weights

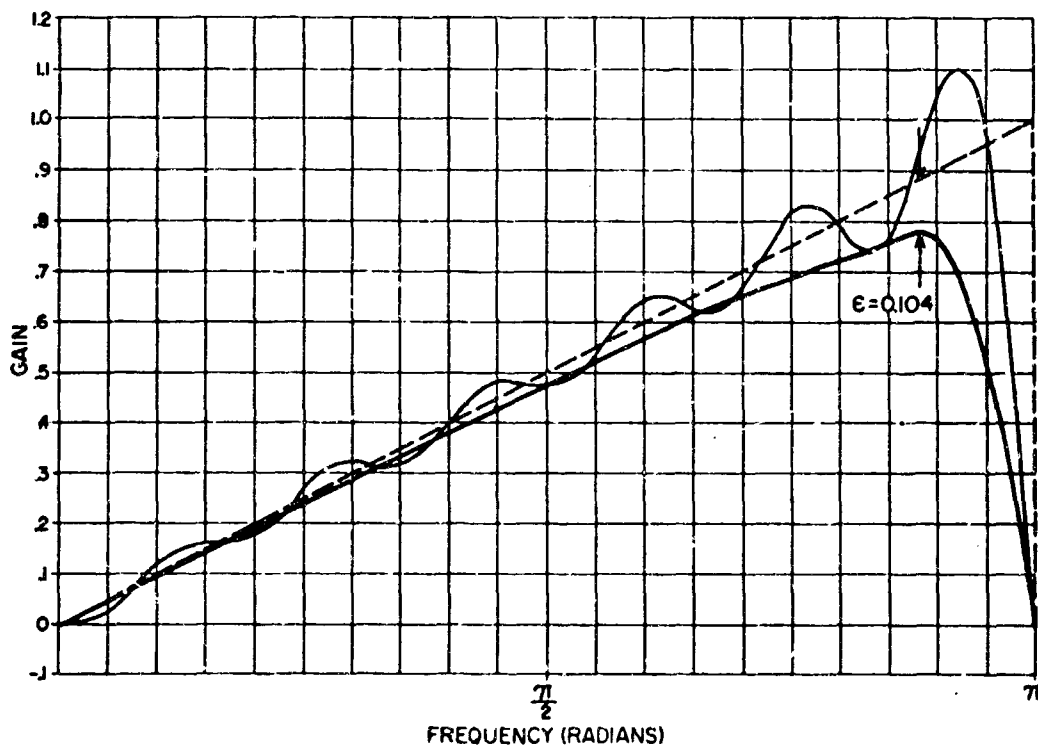


(d) Cosine-Squared Weights

Figure 5.10-5. (Continued)

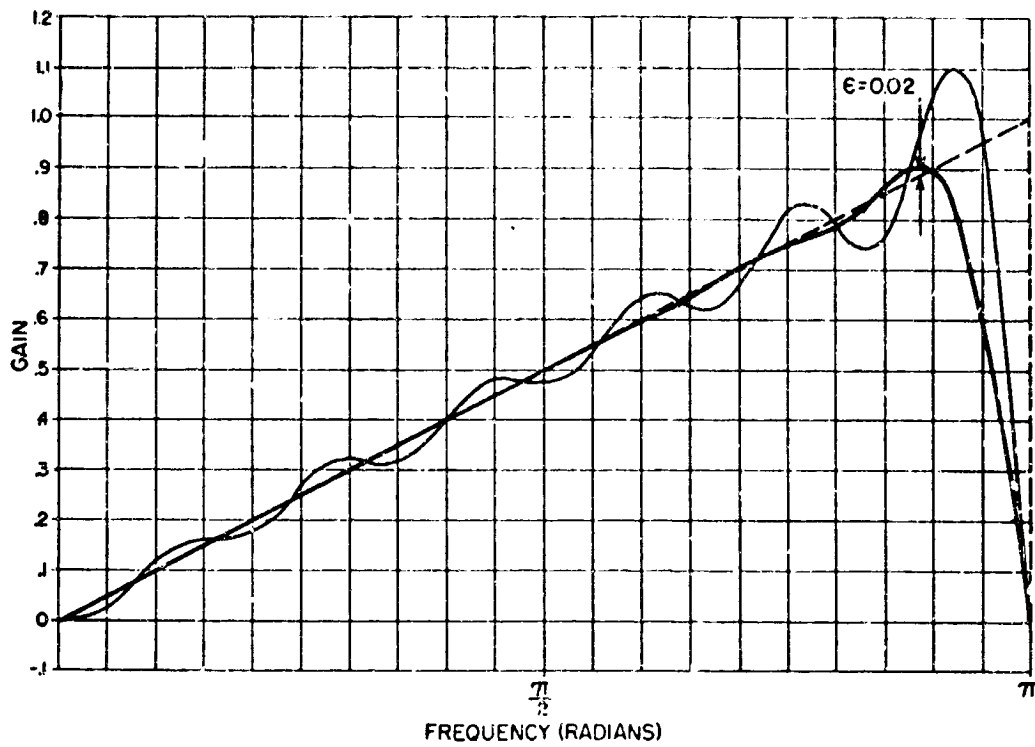


(a) Fourier Weights

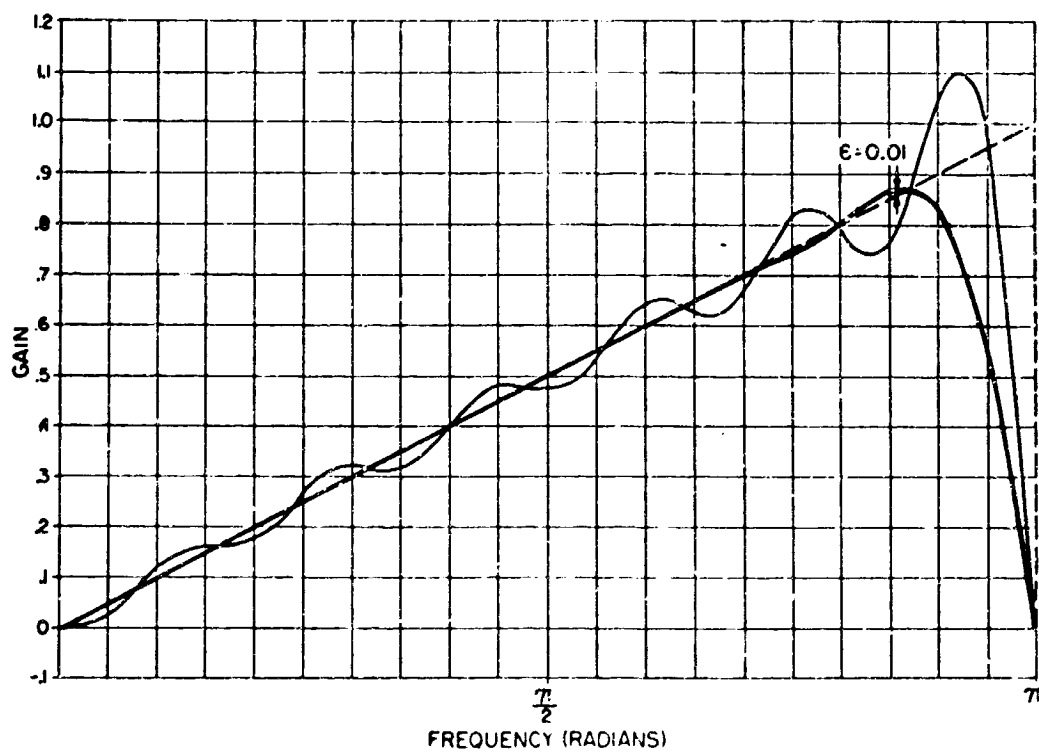


(b) Fejer Weights

Figure 5.10-6. Sine Filter: Differentiator ($\omega_0 = \pi$) $N = 12$



(c) Lanczos Weights



(d) Cosine-Squared Weights

Figure 5.10-6 (Continued)

5.11 RELATIONSHIP OF AMPLITUDE AND PHASE ACCURACIES TO REAL AND IMAGINARY PART ACCURACIES

Assuming that the "delayed" non-recursive transfer function approximation is given in the form

$$H(\omega) = H(\omega)e^{i\theta(\omega)}$$

it can be set equal to (5.11-1)

$$f(\omega) + ig(\omega).$$

Investigation of the necessary accuracy of the two expansions for $f(\omega)$ and $g(\omega)$ should be made where

$$f(\omega) = H(\omega) \cos \theta(\omega)$$

and (5.11-2)

$$g(\omega) = H(\omega) \sin \theta(\omega)$$

If approximations are determined, $f_A(\omega)$ and $g_A(\omega)$, such that errors

$$\epsilon_f(\omega) = f(\omega) - f_A(\omega)$$
(5.11-3)

$$\epsilon_g(\omega) = g(\omega) - g_A(\omega),$$

are made, then the approximation is given by:

$$\begin{aligned} H_A(\omega) &= f_A(\omega) + ig_A(\omega) \\ &= H(\omega) - (\epsilon_f(\omega) + i\epsilon_g(\omega)) \end{aligned}$$
(5.11-4)

or in terms of magnitudes

$$H_A(\omega) = H(\omega)e^{i\theta(\omega)} - B(\omega)e^{i\phi(\omega)}$$
(5.11-5)

where

$$B(\omega) = \sqrt{G_f^2(\omega) + \epsilon_g^2(\omega)} \quad (5.11-6)$$

and

$$\tan \phi(\omega) = \frac{\epsilon_g(\omega)}{\epsilon_f(\omega)} \quad (5.11-7)$$

The magnitude of $H_A(\omega)$ is given by

$$H_A(\omega) = \sqrt{H(\omega)^2 + B^2(\omega) - 2 H(\omega) B(\omega) \cos(\Theta(\omega) - \phi(\omega))} \quad (5.11-8)$$

It can be seen that $H_A(\omega)$ must lie between two bounds

$$H(\omega) - B(\omega) \leq H_A(\omega) \leq H(\omega) + B(\omega) \quad (5.11-9)$$

Therefore, if $\epsilon_f(\omega)$ and $\epsilon_g(\omega)$ are determined such that

$$B(\omega) = \epsilon_H = \sqrt{\epsilon_f^2(\omega) + \epsilon_g^2(\omega)} \quad (5.11-10)$$

throughout $0 \leq \omega \leq \pi$, then the error in approximating the magnitude of $H(\omega)$ is known to be less than, or equal to $B(\omega) = \epsilon_H$.

The phase of $H_A(\omega)$ is given by

$$(5.11-11)$$

$$\tan \Theta_A = \frac{H(\omega) \sin \Theta(\omega) - B(\omega) \sin \phi(\omega)}{H(\omega) \cos \Theta(\omega) - B(\omega) \cos \phi(\omega)}$$

or alternatively

$$\tan \Theta_A(\omega) = \frac{\tan \Theta(\omega) - \frac{B(\omega)}{H(\omega)} \frac{\sin \phi(\omega)}{\cos \Theta(\omega)}}{1 - \frac{B(\omega)}{H(\omega)} \frac{\cos \phi(\omega)}{\cos \Theta(\omega)}} \quad (5.11-12)$$

Let

$$Y(\omega) = \frac{B(\omega)}{H(\omega) \cos \Theta(\omega)} \quad (5.11-13)$$

and have

$$\tan \Theta_A(\omega) = \frac{\tan \Theta(\omega) - \gamma(\omega) \sin \phi(\omega)}{1 - \gamma(\omega) \cos \phi(\omega)} \quad (5.11-14)$$

where $\phi(\omega)$ is usually a rapidly varying function of ω .

Obviously the size of $\gamma(\omega)$ determines the validity of the phase approximation.

For $\Theta(\omega)$ not approximately equal to the tolerable error, ϵ_Θ ; but less than 45 degrees it is found that the maximum error in $\theta(\omega)$ is of the order of

$$\epsilon_\Theta \approx \frac{B(\omega)}{H(\omega)}; \quad \epsilon_\Theta \ll \Theta < 45^\circ. \quad (5.11-15)$$

When Θ approaches 90 degrees information can be found about the error by first assuming $B(\omega) / H(\omega) \ll 1$ (which must be true for "reasonable" approximations) in the numerator of equation (5.11-14). Then

$$\tan \Theta_A = \frac{\tan \Theta}{1 - \frac{B(\omega)}{H(\omega)} \frac{\cos \phi}{\cos \Theta}} \quad (5.11-16)$$

which implies that the fractional error in $\tan \Theta$, ϵ_Θ^f , is

$$\begin{aligned} \epsilon_\Theta^f &= \frac{\tan \Theta - \tan \Theta_A}{\tan \Theta} \\ &= 1 - \frac{1}{1 - \gamma \cos \phi} = \frac{-\gamma \cos \phi}{1 - \gamma \cos \phi} \end{aligned} \quad (5.11-17)$$

Use of the trigonometrical identity

$$\tan (\Theta - \Theta_A) = \frac{\tan \Theta - \tan \Theta_A}{1 + \tan \Theta_A \tan \Theta} \quad (5.11-18)$$

and a bit of manipulation and it can be shown that the error in angle, ϵ_Θ , up to about 90 degrees, also is of the order of

$$\tan \epsilon_\Theta \approx \frac{B(\omega)}{H(\omega)} \quad (5.11-19)$$

and as $(B(\omega)/H(\omega)) \ll 1$ for "reasonable" approximations

$$e_{\Theta} \approx \frac{B(\omega)}{H(\omega)} \quad (5.11-20)$$

5.12 SYNTHESIS OF NON-RECURSIVE FILTERS BY POLYNOMIAL TECHNIQUES

To this point in the discussion Fourier techniques have been used, both continuous and discrete, to approximate the transfer function in the form (5.1-5). The practical utility of these techniques depends on a number of factors (Sections 5 to 5.10).

Polynomial methods have been used historically and the literature is quite extensive. Polynomial methods allow only a few advantages that would compel the use of these techniques rather than remain content with Fourier techniques:

- a. The transfer function is given in sampled form where the samples were not taken equidistantly.
- b. The transfer function is given in terms of powers of $\cos \omega T$ and/or $\sin \omega T$.

As will be shown, Fourier series are equivalent to expansions in Tchebychev polynomials in $\cos \omega T$ which, as is well-known, minimize the maximum error bound versus other polynomial expansions.*

Subsequently, the relation between Tchebychev polynomials and general polynomials will be found. If the designer is given a transfer function in polynomial form in $x = \cos \omega T$, then the conversion to the Fourier form may be made using the relation between a general polynomial and Tchebychev polynomials.

In Section 5.12.5, a technique for fitting a transfer function exactly at points whose values are given unequally spaced in frequency is developed using the Tchebychev expansion relations along with polynomial interpolation techniques. Only real coefficient filters are considered ($H(-\omega) = H^*(\omega)$).

5.12.1 Conversion of the Non-Recursive Filter Transfer Function To a Polynomial Form

Examination of subsequent developments shows that the transformation $x = \cos \omega T$ transforms the transfer function into a linear combination of Tchebychev polynomials (with the exception of a factor $\sqrt{1-x^2}$ in the imaginary part). Further, the transfer function in terms of Tchebychev polynomials can be rearranged into powers of x .

* Hamming, R.; Numerical Methods For Scientists and Engineers

For a large class of functions it can be shown* that an approximation to a function $f(x)$, say $f_A(x)$ in the two forms

$$f_A^{(1)}(x) = \sum_{k=0}^N a_k T_k(x),$$

and

(5.12.1-1)

$$f_A^{(2)}(x) = \sum_{k=0}^N \alpha_k x^k,$$

where $T_k(x)$ is the k^{th} Tchebychev polynomial, can be obtained for two quite different values of N for the same accuracy over a region $-1 \leq x \leq 1$. The expansion in the $T_k(x)$ is often orders of magnitude shorter than the power series. It should be recognized that $f_A^{(1)}(x)$ is nothing more than a Fourier cosine series in disguise ($x = \cos \omega T$).

To convert the transfer function to polynomial form it must be expressed in an appropriate sine-cosine form. In future exposition only the forms for an odd number of weights will be used. The "delayed" non-recursive filter form can be written, upon changing indices

$$H_z^{(\tau)}(\omega) = e^{-i\omega \bar{\tau} T} \sum_{n=-N}^N w_{N+n} e^{-i\omega n T}, \quad \begin{cases} \bar{\tau} = N - \tau \\ M = 2N + 1 \end{cases} \quad (5.12.1-2)$$

Defining, for convenience

$$\bar{H}(\omega) = H_z^{(\tau)}(\omega) e^{i\omega \bar{\tau} T} \quad (5.12.1-3)$$

$$= \sum_{n=-N}^N w_{N+n} e^{-i\omega n T}$$

and expressing the summation on the right in sines and cosines:

$$\bar{H}(\omega) = \sum_{k=0}^N a_k \cos \omega k T - i \sum_{k=1}^N b_k \sin \omega k T, \quad (5.12.1-4)$$

*IBID

where

$$\begin{aligned} a_0 &= w_N \\ a_k &= w_{N+k} + w_{N-k}; \quad k = 1, 2, \dots, N \\ b_k &= w_{N+k} - w_{N-k}; \quad k = 1, 2, \dots, N \end{aligned} \quad (5.12.1-5a)$$

or

$$\begin{aligned} w_N &= a_0; \\ w_{N+k} &= (1/2) (a_k + b_k); \quad k = 1/2, \dots, N, \\ w_{N-k} &= (1/2) (a_k - b_k); \quad k = 1, 2, \dots, N. \end{aligned} \quad (5.12.1-5b)$$

Recognizing that the a_k 's and b_k 's total $2N + 1$ variables and are related, through a non-singular linear transformation, to the $2N + 1$ w_k 's; then they may be considered as independent variables. That is, if the a_k 's can be determined such that equation (5.12.1-4) is a sufficient approximation, then the w_k 's computed from (5.12.1-5) give the same approximation.

To obtain a polynomial form for the "delayed" transfer function use the Tchebychev polynomials.*

The Tchebychev polynomial of the "first kind" of the k^{th} degree is defined by

$$T_k(x) = \cos(k \cos^{-1} x); \quad (5.12.1-6a)$$

letting $x = \cos \theta$; then

$$T_k(x) = \cos k \theta \quad (5.12.1-6b)$$

From trigonometric identities a recursion relation that exists among the $T_k(x)$ can be found: namely

$$T_{k+1}(x) = 2x T_k(x) - T_{k-1}(x) \quad (5.12.1-7)$$

with

$$T_0(x) = \cos(0) = 1$$

and

$$T_1(x) = \cos \theta = x.$$

*Korn and Korn: Mathematical Handbook for Scientists and Engineers McGraw-Hill; pp. 723-724.

Use of this recursion relation gives

$$\begin{aligned}T_2(x) &= 2x^2 - 1 \\T_3(x) &= 4x^3 - 3x\end{aligned}\tag{5.12.1-8}$$

etc:

which are, indeed, polynomials in x . Therefore $\cos k\theta$ can be expressed as a polynomial in $\cos \theta$.

Similarly, define a Tchebychev polynomial of the "second kind" of the $k-1^{\text{st}}$ degree by

$$\sqrt{1-x^2} U_k(x) = \sin(k \cos^{-1} x)\tag{5.12.1-9a}$$

or letting $x = \cos \theta$; then

$$\sqrt{1-x^2} U_k(x) = \sin k\theta.\tag{5.12.1-9b}$$

From trigonometric identities it can be seen that a recursion relation exists among the $U_k(x)$, also, namely

$$U_{k+1}(x) = 2x U_k(x) - U_{k-1}(x)\tag{5.12.1-10}$$

identical to the recursion relation for the $T_k(x)$. This time the first two values of the recursion relation are defined by

$$U_1(x) = 1$$

$$U_2(x) = 2x.$$

Using the recursion relation:

(5.12.1-11)

$$U_3(x) = 4x^2 - 1$$

$$U_4(x) = 8x^3 - 4x$$

etc;

which are polynomials.

Orthogonality relations can be found among the polynomials $T_k(x)$ and $U_k(x)$ similar to those for sines and cosines since they are really sines and cosines in disguise. Both integral orthogonality and orthogonality of discrete sets of points exist, as for sines cosines. As this development is not needed at present it will not be performed here.

In terms of the Tchebychev polynomials of the first and second kinds the transfer function (5.12.1-4) becomes:

$$\bar{H}(\omega) = \sum_{k=0}^N a_k T_k(x) + i \sqrt{1-x^2} \sum_{k=1}^N b_k U_k(x) \quad (5.12.1-12)$$

where $x = \cos \omega T$.

Now $\bar{H}(\omega)$ is a given function (5.12.1-3) determined by an $H(\omega)$ to be approximated and a phase shift that is a known function of ω (a linear phase shift). $\bar{H}(\omega)$ also possesses the property

$$\bar{H}(-\omega) = \bar{H}^*(\omega) \quad (5.12.1-13)$$

which means it can be written as:

$$\bar{H}(\omega) = f(\omega) + ig(\omega) \quad (5.12.1-14)$$

where $f(\omega)$ and $g(\omega)$ are real functions and $f(-\omega) = f(\omega)$, an even function; and $g(-\omega) = -g(\omega)$, an odd function.

Therefore, there are two approximation problems for an arbitrary transfer function $H(\omega)$:

$$f(\omega) = \sum_{k=0}^N a_k T_k(x); \quad x = \cos \omega T \quad (5.12.1-15a)$$

and

$$g(\omega) = \sqrt{1 - x^2} \sum_{k=1}^N b_k U_k(x); \quad 0 \leq \omega T \leq \pi. \quad (5.12.1-15b)$$

The approximation procedure by polynomials can be clearly outlined. Polynomial approximations are found, e.g. Lagrange interpolations, in the variable x :

$$\begin{aligned} f(\omega) &= f\left(\frac{1}{T} \cos^{-1} x\right); \quad x = \cos \omega T \\ &\simeq \sum_{k=0}^N \alpha_k x^k; \end{aligned} \quad (5.12.1-16a)$$

$$\begin{aligned} g(\omega) &= g\left(\frac{1}{T} \cos^{-1} x\right), \\ &\simeq \sqrt{1 - x^2} \sum_{k=0}^{N-1} \beta_k x^k; \end{aligned} \quad (5.12.1-16b)$$

where as ω ranges over $0 \leq \omega T \leq \pi$, x ranges over $1 \geq x \geq -1$.

The coefficients α_k and β_k are determined by the method of approximation to be used. Once the α_k and β_k are determined, then the forms (5.12.1-16) must be converted back to the forms (5.12.1-15) using Tchebychev polynomials of the first and second kinds. From the a_k and b_k the w_k are then computed and the approximation problem solved.

As the $T_k(x)$ are linearly independent among themselves (as are the $U_k(x)$), it is possible to express powers of x in terms of them (or the $U_k(x)$). The $U_k(x)$ are not linearly independent of the $T_k(x)$ and can be expressed in terms of them (or the $T_k(x)$ in terms of the $U_k(x)$).

5.12.2 Reconversion From Polynomial Form to Fourier Form by Determination of the Transforming Matrix Elements

Using the work of the previous section; set

$$\sum_{k=0}^N \alpha_k x^k = \sum_{k=0}^N a_k T_k(x) \quad (5.12.2-1)$$

Using the properties of Tchebychev polynomials the a_k 's in terms of the α_k 's in an exact form can be found. As the calculation is tedious and complicated it is relegated to Appendix A.

In Appendix A it is found that

$$a_0 = \sum_{k=0}^{\left\{\frac{N+1}{2}\right\}} \frac{1}{2^k} C(2k, k) \alpha_{2k}. \quad (5.12.2-2)$$

m even:

$$a_m = \sum_{k=\frac{m}{2}}^{\left\{\frac{N+1}{2}\right\}} \frac{2}{2^k} C(2k, (m+2k)/2) \alpha_{2k}$$

m odd:

$$a_m = \sum_{k=\frac{m+1}{2}}^{\left\{\frac{N+1}{2}\right\}} \frac{2}{2^{2k-1}} C(2k-1, (m+2k-1)/2) \alpha_{2k-1}$$

where

$$\left\{\frac{N+1}{2}\right\}$$

is the truncated integer; for example

$$\left\{\frac{5+1}{2}\right\} = 3$$

and

$$\left\{\frac{4+1}{2}\right\} = 2.$$

Also,

$$C(k, m) = \frac{k!}{m! (k-m)!}$$

The b_k 's can also be found from the imaginary part expansion:

$$\sum_{k=0}^{N-1} \beta_k x^k = \sum_{k=1}^N b_k U_k(x) \quad (5.12.2-3)$$

assuming that the polynomial expansion has been found:

$$\sum_{k=0}^{N-1} \beta_k x^k \simeq \frac{g(\omega)}{\sqrt{1-x^2}}. \quad (5.12.2-4)$$

Appendix A gives the derivation and it is found that:

$$\begin{aligned} \text{m even:} \quad b_m &= \sum_{k=\frac{m}{2}}^{\left\{\frac{N+1}{2}\right\}} \frac{2m}{(2k+m)} \frac{C(2k-1, (2(k-1)+m)/2)}{2^{2k-1}} \beta_{2k-1} \end{aligned} \quad (5.12.2-5)$$

$$\begin{aligned} \text{m odd:} \quad b_m &= \sum_{k=\frac{m-1}{2}}^{\left\{\frac{N}{2}\right\}} \frac{2m}{2k+m+1} \frac{C(2k, (2k+m-1)/2)}{2^{2k}} \beta_{2k} \end{aligned}$$

where $\{x\}$ is the truncated integer part of x , and

$$C(k, m) = \frac{k!}{m! (k-m)!}$$

5.12.3 Reconversion From Polynomial Form to Fourier Form in an Easily Machine Programmable Manner

If a digital computer is available it is often useful to have a method that allows simpler programming. Such a method follows.

To find the a_k 's in the relation

$$\sum_{k=0}^N \alpha_k x^k = \sum_{k=0}^N a_k T_k(x) \quad (5.12.2-1)$$

or as $x = \cos \omega$ (use $\cos \omega$ rather than $\cos \omega T$ for simplicity)

$$\sum_{k=0}^N \alpha_k (\cos \omega)^k = \sum_{k=0}^N a_k \cos k\omega \quad (5.12.3-1)$$

use this time, Newton's method of evaluating a polynomial plus an identity

$$2xT_k(x) = T_{k-1}(x) + T_{k+1}(x), \quad (5.12.3-2a)$$

$$\text{or } \cos \omega \cos k\omega = (\cos (k-1)\omega + \cos (k+1)\omega) / 2. \quad (5.12.3-2b)$$

Newton's method of writing a polynomial is:

$$\sum_{k=0}^N \alpha_k x^k = \alpha_0 + (\alpha_1 + \dots (\alpha_{N-2} + (\alpha_{N-1} + \alpha_1 \cos \omega) \cos \omega) \dots) \cos \omega \quad (5.12.3-3)$$

Starting with the innermost bracket:

$$\begin{aligned} & \alpha_{N-2} (\alpha_{N-1} + \alpha_N \cos \omega) \cos \omega \\ &= \alpha_{N-2} + \alpha_{N-1} \cos \omega + \alpha_N \cos \omega \cos \omega \end{aligned} \quad (5.12.3-4)$$

which upon using the identity becomes:

$$\alpha_{N-2} + \frac{\alpha_N}{2} + \alpha_{N-1} \cos \omega + \frac{\alpha_N}{2} \cos 2\omega.$$

Calculating the next bracket gives:

$$\begin{aligned} & \alpha_{N-3} + (\alpha_{N-2} + \frac{\alpha_N}{2} + \alpha_{N-1} \cos \omega + \frac{\alpha_N}{2} \cos \omega) \cos \omega \\ &= \alpha_{N-3} + (\alpha_{N-2} + \frac{\alpha_N}{2}) \cos \omega + \alpha_{N-1} (\frac{1}{2} + \frac{\cos 2\omega}{2}) \\ &+ \frac{\alpha_N}{2} (\frac{\cos \omega}{2} + \frac{\cos 3\omega}{2}) \\ &= \alpha_{N-3} + \frac{\alpha_{N-1}}{2} + (\alpha_{N-2} + \frac{\alpha_N}{2} + \frac{\alpha_N}{4}) \cos \omega + \frac{\alpha_{N-1}}{2} \cos 2\omega \\ &+ \frac{\alpha_N}{4} \cos 3\omega. \end{aligned} \quad (5.12.3-5)$$

Continuing this process until the α_0 term has been added, gives the form on the right side of equation (5.12.3-1). A particular value of ω is not used when programming the procedure for a machine; simply form the coefficients of the Chebyshev expansion at each stage. At the end of the procedure the a_k 's have been produced.

For the similar expression

$$\sum_{k=0}^{N-1} \beta_k x^k = \sum_{k=1}^N b_k U_k(x); \quad (5.12.2-3)$$

or

$$\sin \omega \sum_{k=0}^{N-1} \beta_k (\cos \omega)^k = \sum_{k=1}^N b_k \sin k\omega \quad (5.12.3-6)$$

use the preceding procedure on the sum

$$\sum_{k=0}^{N-1} \beta_k (\cos \omega)^k \quad (5.12.3-7)$$

and arrive at, say,

$$\sum_{k=0}^{N-1} d_k \cos k\omega \quad (5.12.3-8)$$

By multiplying this intermediate expression by $\sin \omega$:

$$\begin{aligned} \sin \omega \sum_{k=0}^{N-1} \beta_k (\cos \omega)^k &= \sum_{k=0}^{N-1} d_k \sin \omega \cos k\omega \\ &= \sum_{k=0}^{N-1} \frac{d_k}{2} \sin [\sin (k+1) \omega - \sin (k-1) \omega] \end{aligned} \quad (5.12.3-9)$$

This becomes, after adjusting indices,

$$\begin{aligned}
 \sin \omega \sum_{k=0}^{N-1} \beta_k (\cos \omega)^k \\
 &= \sum_{k=1}^N \frac{d_{k-1}}{2} \sin k\omega - \sum_{k=-1}^{N-2} \frac{d_{k+1}}{2} \sin k\omega \\
 &= \sum_{k=1}^N b_k \sin k\omega
 \end{aligned} \tag{5.12.3-10}$$

where

$$\begin{aligned}
 b_N &= d_{N-1}/2 \\
 b_{N-1} &= d_{N-2}/2 \\
 b_k &= (d_{k-1} - d_{k+1})/2; \quad k = 2, 3, \dots, N-2 \\
 b_1 &= d_0 - d_2/2
 \end{aligned} \tag{5.12.3-11}$$

This procedure is only slightly more complex than that for the $\cos k\omega$ expansion.

5.12.4 Exact Transfer Function Fitting by Lagrange Interpolation

Interpolation formulae proliferate varying in complexity of computation or speed of computation. As weight computation is not usually performed except to generate a filter, which is then used many times, only the Lagrange method is discussed which does not necessarily compute the weights in the fastest or most accurate manner. Once the weights are obtained, by any interpolation method, the speed by which they are computed, as long as it is not ridiculously slow, is usually unimportant. The weights are unique, excepting round-off, regardless of the method used.

The interpolation problem is restated in a standard form. The desired transfer function was given in the form

$$\bar{H}(\omega) = f(\omega) + ig(\omega) \tag{5.12.1-14}$$

where $f(\omega) = f(-\omega)$, (even)

and

$$g(-\omega) = -g(\omega), \text{ (odd).}$$

An exact fit at $N + 1$ points is desired by polynomials:

$$f_A(\omega) = \sum_{k=0}^N \alpha_k x^k \quad (5.12.1-16a)$$

$$\simeq f(\omega)$$

and

$$\frac{g_A(\omega)}{\sin \omega T} = \sum_{k=0}^{N-1} \beta_k x^k \quad (5.12.1-16b)$$

$$\simeq \frac{g(\omega)}{\sin nT}$$

where $\cos \omega T = x$.

Reconversion to the transfer function form $\bar{H}(\omega)$, finding the filter weights, is performed in Sections 5.12.1, 5.12.2, and 5.12.3.

Only the region $0 \leq \omega T \leq \pi$, or $1 \geq x \geq -1$ is considered as the fact that $x = \cos \omega T$ guarantees proper behavior for $-\pi \leq \omega T < 0$ for "real" filters.

Because $f_A(\omega)$ involves $N + 1$ coefficients α_k , there must be $N + 1$ values of $f(\omega)$ to determine them. Similarly, N values of $g(\omega)$ are needed to determine the β_k of $g_A(\omega)$. The sampling of the real and imaginary parts can be independent; that is, a different set of values of ω may be used for sampling them. The fit will be exact for each part separately at their sampling points. Normally, the sampling coincides.

A difficulty arises if the point $\omega T = \pi$ is included and an attempt is made to fit with $2N + 1$ weights. This occurs because the term $b_N \sin N\omega T = 0$ at $\omega T = \pi$ which makes it impossible to determine the coefficient b_N . In terms of the filter weights (5.12.1-5) the w_n 's are underdetermined, which can be resolved most easily by setting $w_{2N} = 0$ implying $w_0 = a_N$.

Another problem also arises if $N + 1$ transfer function samples are taken and a sample at $\omega = 0$ is not taken and both the real and imaginary samples are used to determine the α_k and β_k . It will be found that there are $2N + 2$ real and imaginary part samples and only $2N + 1$ coefficients. There is one superfluous imaginary part sample.

The difficulty can be resolved in two ways. The easiest, of course, is to ignore one of the imaginary part samples, in which case an exact fit will not be obtained at this point in the imaginary part. Alternatively, it can be assumed that a filter is being design with $2N+3$ weights, which would mean that $N+2$ α_k and $N+1$ β_k must be determined. The $N+1$ imaginary samples can now be used to determine the β_k . Setting α_{N+1} equal to zero resolves the problem of having too many α_k 's and in turn sets a β_{N+1} to zero. Again there are too many weights, $2N+3$, and only $2N+2$ α_k and β_k (or a_k and b_k). This problem is resolved by setting w_{2N+2} to zero which results in $w_0 = b_{N+1}$. In the end, a filter with $2N+2$ weights will be obtained.

With these qualifications in mind, the real part the samples are:

$$f(\omega_0), f(\omega_1), \dots f(\omega_N),$$

and the imaginary part samples:

$$\frac{g(\omega'_1)}{\sin \omega'_1 T}, \quad \frac{g(\omega'_2)}{\sin \omega'_2 T}, \dots \frac{g(\omega'_N)}{\sin \omega'_N T};$$

where the sampling frequencies for the real and imaginary parts are not necessarily the same (denoted by the prime).

The "fitting" problem for both real and imaginary parts is identical.

Fit a function $h(x)$ by a polynomial

$$\begin{aligned} z(x) &= z(\omega) \\ &= \sum_{k=0}^M \gamma_k x^k \end{aligned} \tag{5.12.4-1}$$

exactly at the points

$$x_k = \cos \omega_k T,$$

given the values $h(x_k)$ at

$$\omega_k = \omega_0, \omega_1, \dots, \omega_M$$

where the ω_k are arbitrarily placed in the interval $0 \leq \omega T \leq \pi$. It can be seen that x_k varies from 1 to -1 as $\omega_k T$ varies from 0 to π .

At this point Lagrange interpolate the function $z(x)$.

The Lagrange interpolation formula is given by*

$$z_L(x) = \sum_{k=0}^M \frac{\pi(x) h(x_k)}{(x-x_k) \pi'(x_k)} \quad (5.12.4-2)$$

where $\pi(x) = (x-x_0)(x-x_1) \dots (x-x_M)$,

$$-1 \leq x \leq 1$$

and

$$\pi'(x_k) = \left. \frac{d\pi(x)}{dx} \right|_{x=x_k}$$

It fits exactly at the points $x_k = \cos \omega_k T$.

It can be shown* that any other method of producing a polynomial of M^{th} order that fits exactly at the $M+1$ points x_k will have exactly the same coefficients as $z_L(x)$, excepting round-off.

Convergence at the points between samples and the magnitude of the errors are other matters.

The error, $\epsilon(x)$, of polynomial interpolation that fits exactly at the $M+1$ points is given by *

$$G(x) = h(x) - z(x) = \frac{\pi(x)}{(M+1)!} h^{(M+1)}(\bar{x}) \quad (5.12.4-3)$$

where

$$h^{(M+1)}(\bar{x}) = \left. \frac{d^{(M+1)}h(x)}{dx^{(M+1)}} \right|_{x=\bar{x}}$$

and x is some value in the interval

$$-1 < \bar{x} < 1.$$

As $\bar{x}$ is usually unknown and extremely difficult to find analytically, an error bound of an M^{th} order approximation can be found if the maximum value of $h^{(M+1)}(x)$ is known in the interval. Seldom is $h^{(M+1)}(x)$ of an experimental curve known with any accuracy, as finding a high order derivative numerically with accuracy is quite difficult.

In general, it may be seen that the maximum error is not only a function of the value of $h^{(M+1)}(x)$ but of the polynomial.

$$\pi(x) = (x-x_0)(x-x_1) \dots (x-x_M) \quad (5.12.4-4)$$

*Hamming, R. : Numerical Methods For Scientists and Engineers

As the point $\bar{x}$ is a function of the sampling points x_k , it is difficult to choose the points x_k so that error is minimum throughout the region $-1 < x < 1$.

While the error is not known exactly, as $\bar{x}$ is not known, the error can be bounded by considering the maximum value of $h^{M+1}(x)$. Namely $D = h^{M+1}(x)_{\max}$ $-1 < x < 1$.

Then the maximum error is certainly

$$\epsilon_M = \frac{\pi(x) D}{(M+1)!} \quad (5.12.4-5)$$

Assuming effects of the x_k on $h^{M+1}(x)$ are not known, the next best way of controlling $\epsilon_{\max}$ is to minimize the maximum value of $\pi(x)$.

The solution to this problem is found by using the Tchebychev polynomials appropriately. To quote the needed property*: "Of all polynomials $P_{M+1}(x)$ with a leading coefficient (x^{M+1}) equal to 1, the polynomial

$$\frac{T_{M+1}(x)}{2^M}$$

has the smallest least upper bound for its absolute value in the interval $-1 < x < 1$."

Since $T_{M+1}(x)$ has a maximum value of 1 in the interval then this upper bound is $1/2^M$. In fact, all maximums of $T_{M+1}(x)$ are equal to one.

Control of the factor $\pi(x)$ in $\epsilon_{\max}$, the maximum interpolation error, can be obtained by letting

$$\pi(x) = \frac{T_{M+1}(x)}{2^M} \quad (5.12.4-6)$$

as $\pi(x)$ has a leading coefficient of 1. This, in turn, implies that the sampling points x_k are the zeros of the Tchebychev polynomial $T_{M+1}(x)$ which are given by

$$x_{M-k} = \cos \frac{(2k+1)\pi}{2(M+1)} \quad (5.12.4-7)$$

$$k = 0, 1, \dots, M.$$

The maximum error, then, becomes

$$\begin{aligned} \epsilon_{\max} &= \frac{D}{2^M (M+1)!} \\ &= \frac{h^{M+1}(x)_{\max}}{2^M (M+1)!} \end{aligned} \quad (5.12.4-8)$$

*IBID

Recapitulating, the minimum maximum possible error is obtained by sampling at the zeros of the $M + 1^{\text{st}}$ Tchebychev polynomial if x is known. In actuality, this is equi-spaced sampling in the ω domain of the transfer function. Namely, equi-spaced sampling in the original transfer function variable ω is more desirable than any other to minimize the maximum possible error. This suggests, in a very strong way, that transfer function approximation using polynomial methods, unless unavoidable for compelling reasons, should be avoided. Fourier methods are generally more efficient and more accurate.

5.12.5 Convergence of Polynomial Interpolation

While it is customary to think that finer and finer interpolation gives smaller errors, as for example in numerical integration or in the very definition of an integral, it is not true, in general, that exact fitting of a polynomial of $M + 1^{\text{st}}$ order at $M + 1$ points give a smaller error with increasing M .

Excellent discussions of the reasons for this unfortunate state of affairs are given in "Applied Analysis", by C. Lanczos. Basically the reason is that the derivatives $h^{M+1}(x)$ grow too rapidly for "non-analytic" functions. "Non-analytic" functions are those that have singularities somewhere in the finite complex plane. For these functions, which comprise a large fraction of those found in practice a finite radius of convergence, R , exists in their Taylor series expansions. This implies for an infinite number of values of M , that

$$\frac{(R + \zeta)^M}{M!} |h^M(x)| \geq 1, \zeta > 0$$

or

(5.12.5-1)

$$|h^M(x)| \geq \frac{M!}{(R + \zeta)^M}.$$

The upper bound on the M^{th} derivative grows as M . Therefore, the upper bound on the error grows with M .

Where the sample values are taken also has a profound effect on the convergence. Equi-spaced sampling in x suffers all the difficulties mentioned above. Sampling at the zeros of Tchebychev polynomials allows convergence. A discussion of these phenomena are outside the scope of this work but can be found in a reference "Applied Analysis" by Cornelius Lanczos.

As an example of a "bad" function in the interval $-1 \leq x \leq 1$, the function

$$\frac{1}{1 + x^2}$$

is a well known example. In fact, as M grows, the approximation diverges between equi-spaced sampling points as M increases.

Polynomial interpolation curve fitting is a dangerous game!!

Important examples of "non-analytic" transfer functions can be ordinary networks which have "poles".

6. RECURSIVE FILTERING TECHNIQUES

Recursive filters, those that use past output sequence values to develop the present output sequence member, appear frequently in the analysis of linear systems. It is the intended purpose of this section to present basic methods of generating recursive filters of various kinds along with limitations imposed. Section 7 will incorporate techniques to alleviate some of the limitations.

Recursive filters possess certain distinct advantages over non-recursive filters and should be used whenever possible. In many applications they have significantly fewer coefficients allowing shorter computing time and less data storage. Approximation ripples in and out of pass-bands are usually not present with recursive filters while great care must be taken with non-recursive filters to avoid the phenomena. Finally, they are often more tractable analytically allowing pertinent characteristics to be discovered more easily and quickly.

Three general classes of techniques will be discussed.

A given analog transfer function is converted via a slightly modified z transform to a digital transfer function which ideally would be equal for all frequencies up to half the sampling frequency. As this requires bandlimiting of the analog transfer function the techniques of Section 7 are needed to extend the useful approximation region toward half the sampling frequency. A numerical integration form of the bi-linear z transform is also discussed but allows approximation to quite limited frequencies.

Another form of the bi-linear z transform is given which allows synthesis of an important class of band-pass and band-reject filters from an analog prototype. As much work has been done in development of analog filters of this type the technique makes available the considerable literature on analog band-pass and band-reject filters to generate comparable digital filters.

A third technique for exact fitting a recursive transfer function to a fixed number of points of a given transfer function is presented. Unfortunately the resultant filter may be unstable. A technique for guaranteeing stability at the price of phase modification is developed.

Miscellaneous topics in this section include the use of the delayed z transform to compute transient responses of analog systems among other things and a set of transformations for converting a digital low-pass filter to band-pass and band-reject filters directly in the digital domain.

6.1 DIGITAL FILTERS THROUGH APPROPRIATE Z TRANSFORMATION OF ANALOG FILTERS

In Section 4.4 the formula

$$y(kT) = Ax(kT) + \sum_{n=-\infty}^{\infty} w_1^c((k-n)T) x(nT) \quad (4.4.1-1)$$

was found for numerical integration where $A = p_N/q_N$. For high sampling frequencies it was found that for a constant coefficient, linear system

$$w_j^c(jT) = T \bar{w}_j(jT) \quad (4.4.1-3)$$

where $\bar{w}_1(jT)$ is the impulse response of the low pass portion of $H(s)$; namely $H_1(s)$ (Section 4.3). If $H(s)$ is low pass $M < N$, then $A = 0$ and $H_1(s) = H(s)$.

Assume now that $w_1^c(jT)$ is equal to $T\bar{w}_1(jT)$ in equation

Subsequently, it will be found that this relation can then be replaced by a realizable recursive digital filter, under a certain condition the filter will produce approximately the same results as the analog filter $H(s)$ and various techniques (Section 7) can be used to modify the analog transfer function so that both the analog and digital filter will have the same responses, within a delay for "low" frequencies.

Replacing $w_1^c(jT)$ by $T\bar{w}_1(jT)$ in

$$y(kT) = Ax(kT) + T \sum_{n=-\infty}^{\infty} \bar{w}_1((k-n)T) x(nT) \quad (6.1-1)$$

and taking the z transform:

$$Y(z) = AX(z) + T \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \bar{w}_1((k-n)T) x(nT) z^{-k} \quad (6.1-2)$$

where

$$X(z) = \sum_{k=0}^{\infty} x(kT) z^{-k}$$

and

$$Y(z) = \sum_{k=0}^{\infty} y(kT)z^{-k},$$

with $y(kT)$ and $x(kT)$ assumed to be equal to zero for $k < 0$.

Adjusting indices in the summation by letting $k-j = m$ ($k = j + m$)

$$Y(z) = \left[A + T \sum_{m=0}^{\infty} \bar{w}_1(mT)z^{-m} \right] X(z). \quad (6.1-3)$$

The digital transfer function is then

$$\begin{aligned} \hat{H}_z(z) &= \frac{Y(z)}{X(z)}, \\ &= A + T \sum_{m=0}^{\infty} \bar{w}_1(mT)z^{-m} \end{aligned} \quad (6.1-4)$$

For constant coefficient, linear systems the function

$$H_z(z) = \sum_{m=0}^{\infty} w_1(mT)z^{-m}, \quad (6.1-5)$$

which is commonly called the z transform of the system, can be summed in closed form as the ratio of two polynomials in z^{-1} (Section 6.2) and realized as a recursive digital filter (Section 3.1).

However the frequency response of $H_z(z)$, found by setting $z = e^{i\omega T}$ is not, in general, equal to that of $H_1(\omega)$. From section 2.1.2 equation (2.1.7-10) a relation can immediately be written between $H_z(e^{i\omega T})$, or $H_z(\omega)$ for convenience, and $H_1(\omega)$. This relation is

$$H_z(\omega) = \frac{\bar{w}_1(0^+)}{2} + \frac{1}{T} \sum_{k=-\infty}^{\infty} H_1(\omega - k\omega_s), \quad \omega_s = \frac{2\pi}{T}; \quad (6.1-6)$$

where $\bar{w}_1(0^+)$ is the limit of $w_1(t)$ as t approaches zero from positive values. The summation on the right consists of terms which are simply translations of the transfer function $H_1(\omega)$ by positive and negative multiples of the sampling frequency. The term $\bar{w}_1(0^+)/2$ is produced by the discontinuity in $\bar{w}_1(t)$ at $t = 0$ if one exists (See Section 2.1.2).

From the initial value theorem of Laplace transforms:

$$\bar{w}_1(0^+) = \lim_{s \rightarrow \infty} [s H_1(s)] \quad (6.1-7)$$

it can be seen that $\bar{w}_1(0^+)$ is only non-zero for lumped parameter filters whose denominator order, N , is equal or 1 greater than the order of the numerator, M :

$$M = N$$

or

$$M = N-1$$

As $H_1(s)$ has a numerator order less than that of the denominator only $M = N-1$ can produce the constant $\bar{w}_1(0^+)$ not equal to zero.

To remove the effect of this constant consider the quantity

$$\begin{aligned} \hat{\hat{H}}_z(z) &= H_z(\omega) - \frac{\bar{w}_1(0^+)}{2} \\ &= \frac{1}{T} \sum_{k=-\infty}^{\infty} H_1(\omega - k\omega_s). \end{aligned} \quad (6.1-8)$$

If the transfer function $H_1(\omega)$ is such that it is essentially zero for $|\omega| > \omega_s/2$; namely, "bandlimited" by half the sampling frequency, then

$$\hat{\hat{H}}_z(\omega) \approx \frac{1}{T} H_1(\omega), \quad |\omega| \leq \omega_s/2. \quad (6.1-9)$$

If the transfer function is not bandlimited, then $\hat{\hat{H}}_z(\omega)$ is "aliased" and not equal, in general, to $H_1(\omega)$ over the entire range $0 < |\omega| \leq \omega_s/2$. The amount of interference of the terms for $k \neq 0$ allow error estimation prior to synthesis.

However, if $H_1(\omega)$ is such that it is approximately zero for $|\omega| > \omega_s - \omega_a$, $\omega_a \leq \omega_s/2$, then

$$\hat{H}_2(\omega) \approx \frac{1}{T} H_1(\omega) \text{ for } |\omega| \leq \omega_a \quad (6.1-10)$$

This occurs because the "tails" of higher order terms of the summation (6.1-8) ($k = \pm 1, \pm 2, \dots$) do not overlap in this region. Use of this fact will allow approximations to $H_1(\omega)$ by a recursive digital filter up to a frequency ω_a in a subsequent section, albeit with a time delay (Section 7).

In summary, this discussion has served to indicate that the function

$$\begin{aligned} T \hat{H}_2(\omega) &= T H_2(\omega) - \frac{T \bar{w}(0^+)}{2} \\ &= \left[T \sum_{m=0}^{\infty} \bar{w}_1(mT) z^{-m} - T \frac{\bar{w}_1(0^+)}{2} \right] z = e^{i\omega T} \end{aligned} \quad (6.1-11)$$

can approximate an analog filter $H_1(\omega)$ under certain definable conditions.

Therefore the digital transfer function in equation (6.1-4) should be slightly modified to read

$$\begin{aligned} \frac{Y(z)}{X(z)} &\equiv \bar{H}_2(z), \\ &= \left[A - \frac{T \bar{w}(0^+)}{2} + T \sum_{m=0}^{\infty} \bar{w}_1(mT) z^{-m} \right]. \end{aligned} \quad (6.1-12)$$

As $\bar{H}_2(z)$ can be summed in closed form, as the ratio of two polynomials in z^{-1} for linear constant coefficient differential equations (Section 6.2), it represents a recursive digital filter of Section 3.1.

It should be noted that the digital impulse response (Section 3.4) of $H_2(z)$ is the actual sampled impulse response of $H(s)$, the continuous network, which is somewhat obvious from the manner in which $H_2(z)$ was defined.

6.2 DIRECT Z TRANSFORMATION

The z transform of the impulse response of a linear, constant coefficient system can be found relatively easily from the appropriate expansion of the transfer function $H(s)$. If the output is desired delayed by an amount of T seconds then the "delayed" z transform should be used (Section 6.3). Usually the transfer function is given as the ratio of two polynomials (See Section 4.1).

$$H(s) = \frac{P(s)}{Q(s)} = \frac{\sum_{k=0}^M p_k s^k}{\sum_{k=0}^N q_k s^k} \quad M \leq N \quad (4.2-3)$$

$H(s)$ can be expanded in "partial fractions" namely

$$H(s) = A + \sum_{j=1}^L \sum_{\mu=1}^{\nu(j)} \frac{A_j^{(\mu)}}{(s + \beta_j)^\mu} \quad (6.2-1)$$

with the sum of the "multiplicities," $\nu(j)$ of the roots given by

$$\sum_{j=1}^L \nu(j) = N \quad (6.2-2)$$

and $(-\beta_j)$ are the L distinct roots of the denominator $Q(s)$. The constant A is given by

$$A = \begin{cases} 0, & M \neq N, \\ p_N/q_N, & M = N \end{cases} \quad (6.2-3)$$

and the subscripted constants A_j^μ , by

$$A_j^{(\nu(j)-m)} = \frac{1}{m!} \frac{d^m}{ds^m} \left[H(s)(s + \beta_j)^{\nu(j)} \right]_{s = -\beta_j} \quad (6.2-4)$$

with $m = 0, 1, 2, \dots, \nu(j)-1$.

As the inverse Laplace transform and z transformation are linear operators the impulse response of each term of (6.2-1) can be computed and the z transform taken. The resultant transforms are summed to give total z transform.

To compute the impulse response of a typical term:

$$\frac{A_j^{(\mu)}}{(s + \beta_j)^\mu} \quad (6.2-5)$$

take its inverse Laplace transform by contour integration. It also may be found in any text on continuous systems.

$$\left[\frac{A_j^{(\mu)}}{(s + \beta_j)^\mu} \right] = \begin{cases} \frac{A_j^{(\mu)}}{(\mu-1)!} t^{\mu-1} e^{-\beta_j t}, & t > 0 \\ 0, & t < 0. \end{cases} \quad (6.2.6)$$

Finding the z transform of $t^{(\mu-1)} e^{-\beta_j t}$ is simplified if it is recognize that

$$(-1)^{(\mu-1)} \frac{\partial^{(\mu-1)}}{\partial \beta_j^{(\mu-1)}} e^{-\beta_j t} = t^{(\mu-1)} e^{-\beta_j t} \quad (6.2-7)$$

$\mu = 1, 2, \dots$

This relation allows the computation of the z transform of multipoles by simple differentiation of the z transform of $e^{-\beta_j t}$. The z transform of $e^{-\beta_j t}$ is

$$\mathcal{Z}(e^{-\beta_j t}) = \sum_{n=0}^{\infty} e^{-\beta_j nT} z^{-n} = \frac{1}{1 - e^{-\beta_j T} z^{-1}} \quad (6.2-8)$$

As parametric differentiation and z transformation are linear, commutative operations, the z transform:

$$z^{(\mu-1)} e^{-\beta_j t} \quad (6.2-9)$$

is equal to

$$(-1)^{(\mu-1)} \frac{\partial^{(\mu-1)}}{\partial \beta_j^{(\mu-1)}} \left[\frac{z}{z - e^{-\beta_j T}} \right] \quad (6.2-10)$$

Combining the previous results it is found that $H_z(z)$, the z transform transfer function which is the z transform of the Laplace transfer function $H(s)$ impulse response, is given by

$$H_z(z) = \sum_{j=1}^L \sum_{\mu=1}^{\nu(j)} \frac{(-1)^{(\mu-1)} A_j^{(\mu)}}{(\mu-1)!} \frac{\partial^{(\mu-1)}}{\partial \beta_j^{(\mu-1)}} \left[\frac{z}{z - e^{-\beta_j T}} \right] + A \quad (6.2-11)$$

where A is simply added.

All of the operations necessary to compute $H_z(z)$ from $H(s)$ can be performed by a computer program. It is not necessary to do the differentiations and summations analytically. When $H_z(z)$ is computed as above all the individual terms are combined into the ratio of two polynomials in z. In general, the final form of $H_z(z)$ will be

$$H_z(z) = \frac{\sum_{k=0}^N a_k z^{-k}}{\sum_{k=0}^N b_k z^{-k}}, \quad (6.2-12)$$

where the a_k and b_k are determined by converting equation (6.2-11) to a least common denominator and normalizing such that $b_0 = 1$. N is the degree of the original denominator. Again this computation can be performed by machine rather than by hand when N is large. Subsequently, a recursive technique will be developed for computing the z transform of a high order pole.

For computational simplicity some useful relations should be used.

1. For $H(s)$ with real coefficients, roots occur in complex conjugate pairs. Partial fraction expansion coefficients for one complex root are conjugates of these for the conjugate root. Therefore, only coefficients for half of the complex poles must be computed.
2. It was seen that the impulse response term $\bar{w}(0+)$ is needed when $M + N - 1$. By examination of the impulse response as the term by term inverse Laplace transform of the partial fraction expansion of $H(s)$ it can be seen that

$$\bar{w}(0+) = \text{Real part} \left[\sum_{j=1}^L A_j^{(1)} \right] \quad (6.2-13)$$

Alternatively, as

$$\bar{w}(0+) = \lim_{s \rightarrow \infty} [sH(s)]$$

then

(6.2-14)

$$\bar{w}(0+) = \frac{p_{N-1}}{q_N}$$

Either relation may be used according to the manner in which $H(s)$ is given

6.2.1 Mechanization of the Computation of the z Transform for m^{th} Order Poles

The z transform for partial fraction terms corresponding to a j^{th} multipole of order $\nu(j)$:

$$\sum_{\mu=1}^{(j)} \frac{A_j^{(\mu)}}{(s + \beta_j)^\mu} \quad (6.2.1-1)$$

is given by (See equation 6.2-10)).

$$\sum_{\mu=1}^{\nu(j)} \frac{A_i^{(\mu)} (-1)^{\mu-1}}{(\mu-1)!} \frac{\partial^{(\mu-1)}}{\partial \beta_j^{(\mu-1)}} \left[\frac{z}{z - e^{-\beta_j T}} \right] \quad (6.2.1-2)$$

Computation of this expression for a high order pole is difficult to perform accurately by hand. However, machine computation is simple if certain relations are recognized. The procedure to be developed obtains results in powers of z rather than z^{-1} . Conversion to inverse powers is best performed after all poles have been combined to form the transfer function. Then only the ordering of coefficients need be changed.

Computing the terms of equation (6.2.1-2) starting with the lowest order is obviously the preferable procedure.

The partial derivatives can be expressed as

$$\frac{\partial^{(\mu-1)}}{\partial \beta_j^{(\mu-1)}} \left[\frac{z}{z - e^{-\beta_j T}} \right] \quad (6.2.1-3)$$

$$= z (-T)^{(\mu-1)} \frac{\partial^{(\mu-1)}}{\partial v^{(\mu-1)}} \left[\frac{1}{z - e^v} \right] v = -\beta_j T$$

Defining the derivatives as

$$D_m = \frac{\partial^m}{\partial v^m} \left[\frac{1}{z - e^v} \right] \quad (6.2.1-4)$$

the first few are given by

$$\begin{aligned} D_0 &= \frac{1}{z - v} , \\ D_1 &= \frac{e^v}{(z - e^v)^2} , \\ D_2 &= \frac{2! e^{2v}}{(z - e^v)^3} + \frac{1! e^v}{(z - e^v)^2} , \end{aligned} \quad (6.2.1-5)$$

$$D_3 = \frac{3! e^{3v}}{(z - e^v)^4} + \frac{2 \times 2! e^{2v}}{(z - e^v)^3} + \frac{2! e^{2v}}{(z - e^v)^3} + \frac{e^v}{(z - e^v)^2}$$

With this simplification the z transform of the multipole expansion (6.2.1-2) above can be written as

$$\sum_{\mu=1}^{\nu(j)} \frac{A_i^{(\mu)}}{(\mu-1)!} T^{(\mu-1)} z D_{(m-1)} \quad (6.2.1-6)$$

Placing the first few D_m 's over common denominators:

$$D_0 = \frac{1}{z - e^v}$$

$$D_1 = \frac{e^v}{(z - e^v)^2} \quad (6.2.1-7)$$

$$D_2 = \frac{e^{2v} + e^v z}{(z - e^v)^3},$$

$$D_3 = \frac{e^{3v} + 4e^{2v} z + e^v z^2}{(z - e^v)^4},$$

and

$$D_4 = \frac{e^{4v} + 11e^{3v} z + 11e^{2v} z^2 + e^v z^3}{(z - e^v)^5}$$

It can be seen that the general form of D_m appears to be

$$D_m = \frac{a_0^m e^{mv} + a_1^m e^{(m-1)v} z + \dots + a_{m-1}^m e^v z^{m-1}}{(z - e^v)^{m+1}} \quad (6.2.1-8)$$

where the a_j^m are the coefficients which are presently unknown. The superscript m on the a_j^m is not an exponent but serves to indicate the set of m coefficients for the m^{th} D_m . The subscripts indicate the appropriate power of z .

By using relation (6.2.1-4)

$$D_{m+1} = \frac{\partial}{\partial v} D_m \quad (6.2.1-9)$$

where D_m is used in the form (6.2.1-8) and equating it, after placing it in least common denominator form, to the general form for D_{m+1} :

$$D_{m+1} = \frac{a_0^{m+1} e^{(m+1)v} + a_1^{m+1} e^{mv} z + \dots + a_m^{m+1} e^v z^m}{(z - e^v)^{m+2}} \quad (6.2.1-10)$$

a recursive relation is found expressing the coefficients for D_{m+1} in terms of those for D_m ; namely

$$a_j^{m+1} = (j+1) a_j^m + (m - (j-1)) a_{j-1}^m, \quad (6.2.1-11)$$

with end conditions

$$a_0^m = a_{m-1}^m = 1; \quad m = 2, 3, \dots$$

Obviously the relation is only needed for $m = 3$ and higher as $m = 0$, and 1 are special cases. For $m = 2$ the coefficients

$$a_1^2 = 1,$$

(6.2.1-12)

$$a_0^2 = 1,$$

are known and used as the starting point for the recursive formula (eq (6.2.1-11)) to compute higher order derivative coefficients.

As an example the coefficients for D_3 and D_4 are calculated using the recursive relation.

$$\text{Now } a_0^3 = a_2^3 = 1 \text{ from the end conditions.}$$

(6.2.1-13a)

Only a_1^3 must be computed:

$$a_1^3 = 2a_1^2 + (2 - (1-1)) a_0^2$$

$$= 2 + 2$$

(6.2.1-13b)

$$= 4, \text{ as } a_1^2 = a_0^2 = 1.$$

Therefore

$$D_3 = \frac{e^{3v} + 4e^{2v}z + e^v z^2}{(z - e^v)^4}$$

(6.2.1-14)

as was obtained by direct calculation (eq 6.2.1-7).

Similarly, for D_4 , $a_0^4 = a_3^4 = 1$. a_1^4 and a_2^4 must be computed from the coefficients for D_3 :

$$\begin{aligned}
 a_1^4 &= 2 a_1^3 + (3 - (1-1)) a_0^3, \\
 &= 11;
 \end{aligned}
 \tag{6.2.1-15}$$

and

$$\begin{aligned}
 a_2^4 &= 3 a_2^3 + (3 - (2-1)) a_1^3, \\
 &= 3 + 2 \times 4, \\
 &= 11.
 \end{aligned}$$

D_4 is then given by

$$D_4 = \frac{e^{4v} + 11 e^{3v} z + 11 e^{2v} z^2 + e^v z^3}{(z - e^v)^5}
 \tag{6.2.1-16}$$

as calculated previously (eq 6.2.1-7).

The computation for the coefficients of successive derivatives is easily organized into a computer program along with the power of e^v to give the coefficients of the polynomial in z which is the numerator of D_m . Computation of the denominator of D_m , $(z - e^v)^{M+1}$, is most easily performed using the binomial theorem

$$(z - e^v)^n = \sum_{j=0}^n \frac{n!}{j! (n-j)!} (-e^v)^{n-j} z^j
 \tag{6.2.1-17}$$

A recursive relation can also be found, if desired, among successive coefficients of the denominators for a fixed n , though this is not as necessary as the recursive relation for the numerator.

6.2.2 An Example of Digital Filter Design by the z Transform

Synthesis of a bandpass filter of a Q of 20 (center frequency divided by bandwidth) follows as an example. The response before and after correction for the impulse response term at $t = 0$ ($\bar{w}(0^+)/2$) is plotted in Figure 6.2.2-1.

The Laplace transfer function of a bandpass network (a single stage L-R-C network) is given by

$$H(s) = \frac{2 \zeta \omega_o A s}{s^2 + 2 \zeta \omega_o s + \omega_o^2} \quad (6.2.2-1)$$

where

ω_o = center frequency in radians,

ζ = $1 / 2 Q$,

Q = ω_o / β ,

β = bandwidth in radians,

A = gain at ω_o .

Actually the above relationship between ζ and Q is an approximation which holds for large Q . Typically of interest is large Q , so this approximation will be used. Q 's greater than 5 are usually sufficiently accurate.

The denominator of $H(s)$ is factored into two roots:

$$s^2 + 2 \zeta \omega_o s + \omega_o^2 = (s + \zeta \omega_o + i \omega_o \sqrt{1 - \zeta^2}) (s + \zeta \omega_o - i \omega_o \sqrt{1 - \zeta^2}), \quad (6.2.2-2)$$

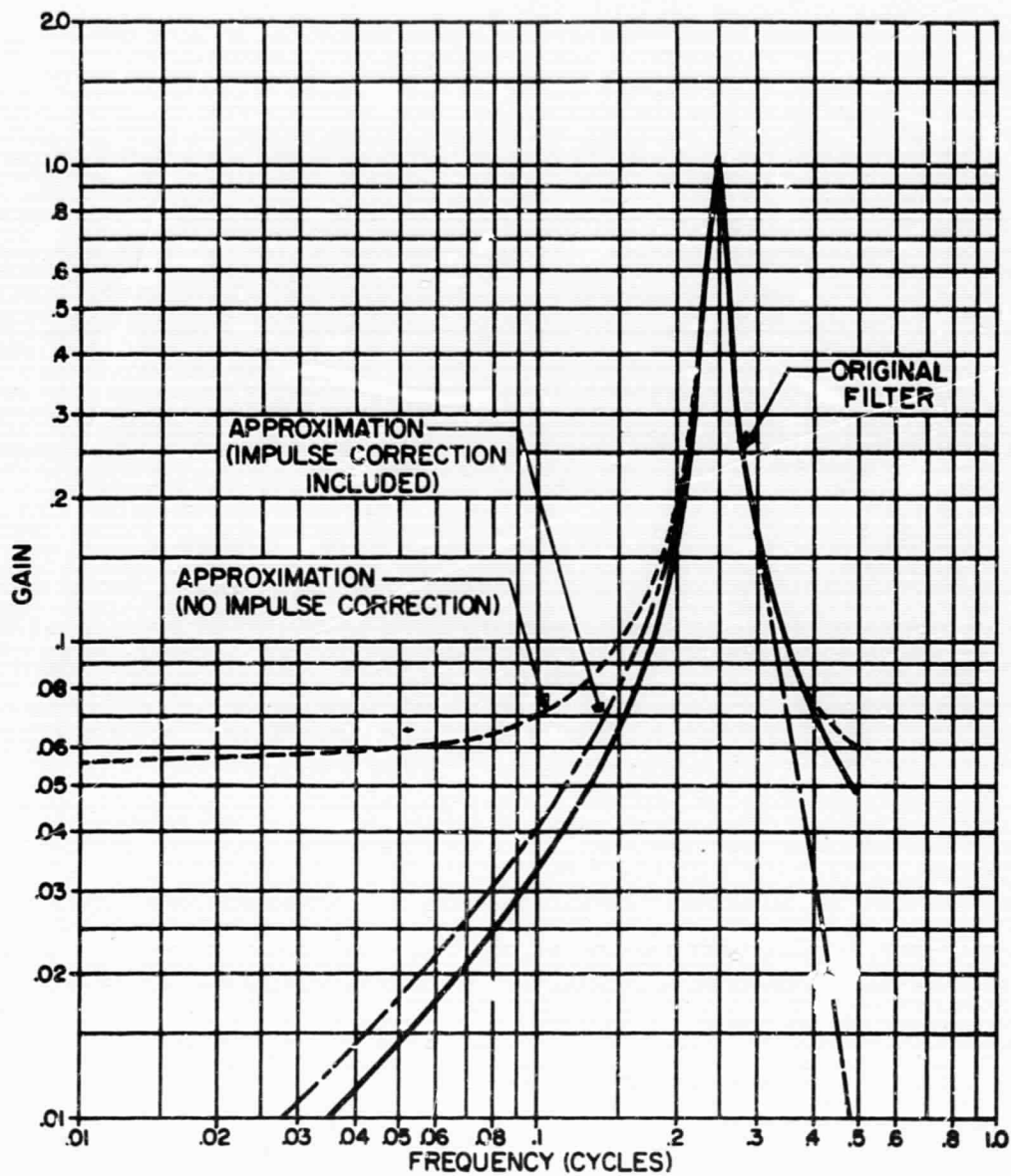


Figure 6.2.2-1. "Modified Impulse Invariant" Design of a Single Pole Bandpass Filter

which allows

$$H(s) = \frac{2 \zeta \omega_0 A s}{(s + \zeta \omega_0 + i \omega_0 \sqrt{1 - \zeta^2})(s + \omega_0 - i \omega_0 \sqrt{1 - \zeta^2})} \quad (6.2.2-3)$$

to be written. The roots of the denominator are complex conjugates.

Next the partial fraction expansion of $H(s)$ is determined:

$$\frac{2 \zeta \omega_0 A s}{(s+a)(s+a)^*} = \frac{A_1}{s+a} + \frac{A_2}{s+a^*} \quad (6.2.2-4)$$

where

$$a = \zeta \omega_0 + i \omega_0 \sqrt{1 - \zeta^2}$$

and a^* is the complex conjugate of a . To determine A_1 , multiply both sides of (6.2.2-4) by $(s+a)$, and set $s = -a$. This gives

$$\begin{aligned} A_1 &= H(s)(s+a) \Big|_{s=-a}, \\ &= \frac{2 \zeta \omega_0 A s}{s+a^*} \Big|_{s=-a}, \\ &= \frac{2 \zeta \omega_0 A a}{a - a^*} \end{aligned} \quad (6.2.2-5)$$

or

$$A_1 = A \omega_0 \left(\zeta - i \frac{\zeta^2}{\sqrt{1 - \zeta^2}} \right)$$

In general it can be shown that the partial fraction expansion coefficients for complex conjugate roots are complex conjugates. Therefore, $A_2 = A_1^*$, though A_2 could be computed as was A_1 , if desired. The resultant expansion is

$$H(s) = A \omega_0 \left(\frac{1 - i \frac{\zeta}{\sqrt{1-\zeta^2}}}{s + a} + \frac{1 + i \frac{\zeta}{\sqrt{1-\zeta^2}}}{s + a^*} \right) \quad (6.2.2-6)$$

Furthermore, the impulse response at $t = 0^+$, the sum of the real parts of the partial fraction expansion coefficients of single pole terms is

$$\bar{w}(0^+) = 2 A \zeta \omega_0 \quad (6.2.2-7)$$

The corresponding z transform of $H(s)$ above is

$$H_z(z) = A \omega_0 \zeta \left(\frac{1 - i \frac{\zeta}{\sqrt{1-\zeta^2}}}{1 - e^{-aT} z^{-1}} + \frac{1 + i \frac{\zeta}{\sqrt{1-\zeta^2}}}{1 - e^{-a^*T} z^{-1}} \right) \quad (6.2.2-8)$$

Combining as the ratio of two polynomials:

$$H_z(z) = \frac{2 A \omega_0 \zeta \left(1 - e^{-\zeta \omega_0 T} \left(\cos(\omega_0 T \sqrt{1-\zeta^2}) + \frac{\zeta}{\sqrt{1-\zeta^2}} \sin(\omega_0 T \sqrt{1-\zeta^2}) \right) \right)}{1 - 2e^{-\zeta \omega_0 T} \cos(\omega_0 T \sqrt{1-\zeta^2}) z^{-1} + e^{-2\zeta \omega_0 T} z^{-2}} \quad (6.2.2-9)$$

Next, correct for the discontinuity in the impulse response, and the scaling factor T :

$$\begin{aligned} \bar{H}_z(z) &= T \left(H_z(z) - \frac{w(0^+)}{2} \right) \\ &= \frac{A \omega_0 \zeta T \left(1 - 2e^{-\zeta \omega_0 T} \frac{\zeta}{\sqrt{1-\zeta^2}} \sin(\omega_0 T \sqrt{1-\zeta^2}) z^{-1} - e^{-2\zeta \omega_0 T} z^{-2} \right)}{1 - 2e^{-\zeta \omega_0 T} \cos(\omega_0 T \sqrt{1-\zeta^2}) z^{-1} + e^{-2\zeta \omega_0 T} z^{-2}} \end{aligned} \quad (6.2.2-10)$$

This is the form necessary for the digital filter.

$$\alpha = A\omega_0 \zeta T,$$

$$a_0 = \alpha,$$

$$a_1 = -2\alpha e^{-\zeta\omega_0 T} \frac{\zeta}{\sqrt{1-\zeta^2}} \sin(\omega_0 T \sqrt{1-\zeta^2}),$$

$$a_2 = -\alpha e^{-2\zeta\omega_0 T} \quad (6.2.2-11)$$

$$b_0 = 1,$$

$$b_1 = 2e^{-\zeta\omega_0 T} \cos(\omega_0 T \sqrt{1-\zeta^2}),$$

$$b_2 = e^{-2\zeta\omega_0 T};$$

then

$$\bar{H}_z(z) = \frac{\sum_{m=0}^2 a_m z^{-m}}{\sum_{n=0}^2 b_n z^{-n}} \quad (6.2.2-12)$$

The corresponding difference equation that implements the filter is

$$y_n = -b_1 y_{n-1} - b_2 y_{n-2} + a_0 x_n + a_1 x_{n-1} + a_2 x_{n-2} \quad (6.2.2-13)$$

where, y_n is the output sequence, and x_n , the input sequence.

The figure plotted of $H_z(z)$ (Figure 6.2.2-1) before compensation, and $\bar{H}_z(z)$ after compensation clearly shows the validity of the technique. An $\omega_0 = 2\pi f_0$ of 1/4 the sampling frequency was used as center frequency and a Q of 20 was chosen. $H_z(z)$ before compensation was scaled by 1/T so that plots could be compared.

Examination of the corrected digital filter indicates that a better "fall-off" is obtained for $\omega > \pi/2$ than the continuous filter. This occurs because of the image tail of the first aliasing harmonic (equation (6.1-6)) happens to cancel the "fall-off" as it is shifted by 180 degrees in phase. At frequencies less than the center frequency the image tail from the aliasing harmonic is in phase.

6.3 THE DELAYED z TRANSFORMATION

If the output sequence of a reconstruction procedure, (Section 7.1) a numerical integration procedure or a digital filter is desired at a time not equal to an integral multiple of the sampling time ($t = nT$), but at a time $t = nT - mT - \tau$ where τ is not an integer, then the "delayed" z transform should be used

The rationale is quickly seen if the derivations of numerical integration procedures are examined (Sections 4.4 and 6.1). The outputs are evaluated at an input sampling instant. Had $t = nT - \tau$ been used slightly different formula for the z transform of the analog filter impulse response would have been found:

$$H_z(z) = \sum_{n=0}^{\infty} \bar{w}(nT - \tau) z^{-n} \quad (6.3-1)$$

τ may be any number but it is convenient and not restrictive to consider it as less than T. If τ is, for example, equal to $\tau' + mT$ ($\tau' \leq T$), then mT is simply a delay of m units of the sampling time T

Another application of the delayed z transform is in computation of transient solutions (Section 6.4) where the transient must be delayed as the numerical integration procedure for the homogeneous solution usually has delay (Section 7). Again the delayed z transform of the transient solution has the form (6.2-1) with the inverse transform given, this time, by samples of the continuous system transient solution.

In all cases discussed to this point, if a function $f(t - \tau)$ is the sampled inverse Laplace transform of the transfer function (impulse response) or the transient solution of a constant coefficient system, the "delayed" z transform may be found in closed form as the ratio of two polynomials in z^{-1} similar to the plain z transform.

If the transfer function or transient solution is given as the ratio of two polynomials it may be expanded in partial fractions as in Section 6.5.

$$F(s) = A + \sum_{j=1}^L \sum_{n=1}^{\nu(j)} \frac{A_j^{(n)}}{(s + \beta_j)^n} \quad (6.3-2)$$

$$\mathcal{L}^{-1} \left[\frac{A_j^{(\mu)}}{(s + \beta_j)^\mu} \right] = \begin{cases} \frac{A_j^{(\mu)}}{(\mu - 1)!} t^{\mu-1} e^{-\beta_j t}, & t > 0, \\ 0, & t < 0. \end{cases} \quad (6.3-3)$$

$$\equiv f_j^\mu(t)$$

The delayed z transform to be computed is

$$F_z^D(z) = A + \sum_{n=0}^{\infty} f(nT - \tau) z^{-n}, \quad 0 < \tau \leq T, \quad (6.3-4)$$

with

$$f(t) = \sum_{j=1}^L \sum_{m=1}^{v(j)} f_j^\mu(t)$$

As $\tau > 0$, and $f(t) = 0$ for $t < 0$, it may be noted that the term for $n = 0$ is not included in the delayed z transform.

To compute the z transformation for the "delayed" terms consider the time function

$$e^{-\beta_j(t-\tau)}$$

common in form to all terms. The general terms are given by

$$(-1)^{(\mu-1)} \frac{\partial}{\partial \beta_j} \left[e^{-\beta_j(t-\tau)} \right] = (t-\tau)^{(\mu-1)} e^{-\beta_j(t-\tau)}. \quad (6.3-5)$$

As discussed previously the z transform and differentiation by β_j are commutative operations. Therefore, the z transform of the individual terms is given by

$$\mathcal{Z}(f_j^\mu(t-\tau)) = (-1)^{(\mu-1)} \frac{\partial}{\partial \beta_j} \left[\sum_{k=1}^{\infty} e^{-\beta_j(kT-\tau)} z^{-k} \right] \quad (6.3-6)$$

Adjusting the indices on the summation:

$$\sum_{k=1}^{\infty} e^{-\beta_j(kT - \tau)} z^{-k} = \sum_{k=0}^{\infty} e^{-\beta_j kT} z^{-k} e^{-\beta_j(T-\tau)} z^{-1} \quad (6.3-7)$$

$$= \frac{z^{-1} e^{-\beta_j(T-\tau)}}{1 - e^{-\beta_j T} z^{-1}}$$

The z transform of the general "delayed term" is

$$\mathcal{Z} \left(\frac{A_j(\mu)}{(\mu-1)!} (t-\tau)^{(\mu-1)} e^{-\beta_j t} \right) = (-1)^{(\mu-1)} \frac{A_j(\mu)}{(\mu-1)!} \frac{\partial^{(\mu-1)}}{\partial \beta_j^{(\mu-1)}} \left[\frac{z^{-1} e^{-\beta_j(t-\tau)}}{1 - e^{-\beta_j T} z^{-1}} \right] \quad (6.3-8)$$

It can be seen that the z transform of delayed terms are ratios of polynomials in z^{-1} as in the case of integer delay. The terms for the whole transform can be combined into the ratio of two polynomials in z^{-1} which in turn produces a difference equation as before.

Recursive relations can be developed, as for the plain z transform, to mechanize the computation of the transform for higher order poles.

6.4 MECHANIZATION OF INITIAL CONDITION OR "TRANSIENT" COMPUTATIONS

The transient solution $f(t)$, which is the inverse Laplace transform of $F(s)$ (4.2-4) can be readily computed by expanding $F(s)$ in partial fractions (see Section 6.2) and obtaining the inverse transform term by term

This function $f(t)$ can then be evaluated at $t = kT$

$$f_k = f(kT)$$

and added to the solution for zero initial conditions. As will be seen in section 7 numerical solutions for zero initial conditions usually have a time delay of a known amount. Before $f(t)$ is used it should be delayed by the amount of time (or simulated time) found in the numerical integration procedure.

A simple procedure for computing the "transient", via the impulse response of a digital filter can be found. At first assume that the delay is an integral multiple of T plus an amount τ ($0 < \tau \leq T$) in which case the delay is a simple deferment of calculation for the requisite number of sampling periods plus a delay by τ . By taking a delayed z transform of $f(t)$ it is found that the resultant transform is the ratio of two polynomials in z as the transform of the individual terms of the time function, found by partial fraction expansion of $F(s)$, are ratios of polynomials in z (Section 6.3).

In fact $F(s)$ may be looked upon as a transfer function as far as the z transform calculation is concerned. The method of Section 6.3 is directly applicable to $F(s)$ using the delayed z transform. The time function $f(t)$ may be considered as the impulse response of $F(s)$.

Writing the delayed z transform of $f(t)$, which can be computed via $F(s)$ from the method of Section 6.3 as a ratio of two polynomials in z^{-1} ,

$$F_z^D(z) = \frac{\sum_{m=0}^M c_m z^{-m}}{\sum_{n=0}^N d_n z^{-n}} \quad (6.4-1)$$

where the coefficients have been normalized so that $d_0 = 1$, an equivalent difference equation or digital filter can be written;

$$h_j = -\sum_{n=1}^N d_n h_{j-n} + \sum_{m=0}^M c_m g_{j-m} \quad (6.4-2)$$

See Section 3.2.

Now the response of this difference equation to an impulsive input in the digital domain:

$$\delta_j = \begin{cases} 1, & j = 0, \\ 0, & j \neq 0; \end{cases}$$

is the impulse response of $F_z^D(z)$ in the digital domain or the sampled time function:

$$h_j = f(jT - \tau).$$

In summary, application of δ_j to the recursive relation produces the "transient" time function at the sampling time:

$$h_j = -\sum_{n=1}^N d_n h_{j-n} + \sum_{m=0}^M c_m \delta_{j-m} \quad (6.4-3)$$

Prior to application of δ_0 the initial values of h_j are set to zero; $j = -1, -2, \dots, -N$.

6.4.1 Example of Delayed z Transform Utilization to Compute Transient Responses

This example is a simple case of a more general discussion of initial condition computation given in Section 4.2. Consider the simple linear, constant coefficient differential equation.

$$\frac{dy}{dt} + ay = a \frac{dx}{dt} \quad (6.4.1-1)$$

Taking the Laplace transform:

$$sY(s) - y(0+) + aY(s) = a[sX(s) - ax(0+)] \quad (6.4.1-2)$$

or

$$Y(s) = \frac{as}{s+a} X(s) + \frac{y(0+) - ax(0+)}{s+a}$$

where

$Y(s)$ = Laplace transform of $y(t)$,

$X(s)$ = Laplace transform of $x(t)$,

and $y(0+)$ and $x(0+)$ are initial conditions.

The transfer function $H(s)$ is obviously given by

$$H(s) = \frac{as}{s+a} \quad (6.4.1-3)$$

and the transient transform by (See Section 4.2)

$$F(s) = \frac{y(0+) - ax(0+)}{(s+a)} \quad (6.4.1-4)$$

The inhomogeneous solution is determined by the methods of Sections 6.1 and 7 applied to $H(s)$. The transient, or homogeneous solution, can be found by taking the inverse Laplace transform of $F(s)$ directly:

$$f(t) = \left[y(0+) - ax(0+) \right] e^{-at} \quad (6.4.1-5)$$

Delayed by τ ($0 < \tau \leq T$) and evaluated at $t = kT$, if necessary, to compensate for delay in the inhomogeneous solution.

$$f(kT - \tau) = \left[y(0+) - ax(0+) \right] e^{-a(kT - \tau)} \quad (6.4.1-6)$$

By using the delayed z transform at the point where $F(s)$ was determined (eq (4.1-4)).

$$\begin{aligned} \mathcal{Z}(F(s)) &= F_z^D(z) \\ &= \left[y(0+) - ax(0+) \right] \frac{e^{-a(T-\tau)} z^{-1}}{1 - e^{-aT} z^{-1}} \end{aligned} \quad (6.4.1-7)$$

which has the corresponding difference equation (Section 6.4).

$$h_j^\tau = e^{-aT} h_{j-1}^\tau + C e^{-a(T-\tau)} g_{j-1} \quad (6.4.1-8)$$

where

$$C = y(0+) - ax(0+)$$

Applying an impulse

$$g_j = \begin{cases} 1, & j = 0 \\ 0, & j > 0 \end{cases}$$

and letting

$$h_{-j} = 0, \quad -j = -1, -2, \dots,$$

the impulse response is found to be

$$h_0^\tau = 0,$$

$$h_1^\tau = e^{-a(T-\tau)},$$

(6.4.1-9)

$$h_2^\tau = e^{-a(2T-\tau)},$$

etc,

or

$$h_k^\tau = f(kT-\tau)$$

as expected.

While in this example, it was a simple matter to determine the transient response directly rather than via z transform the recursion formula allows computation by simple multiplication and addition rather than using an exponentiation computation. For more complicated problems, the z transform technique is usually easier to use as the number of exponentials and sines any cosines increases with the complexity of $F(s)$ producing increased computation time.

6.5 DIGITAL INPUT - OUTPUT RELATION FOR LINEAR, CONSTANT COEFFICIENT SYSTEMS WITH "IMPULSE SIMPLIFIED" INPUTS

For "impulse sampled" (Section 2.1.4) inputs a simple relation between the input sample values and the output sample values of a constant coefficient, linear system at the same, or displaced, sampling instants may be found. Subsequent z transformation of this relation produces a recursion relation, or digital filter, relating input and output samples in a manner quite similar to that of the previous section.

The impulse sampled input function is given by

$$x_s(t) = x(t) \sum_{j=0}^{\infty} \delta(t-jT).$$

$x(t)$ = the analog input prior to impulse sampling, assumed

to be zero for $t < 0$,

$\delta(t)$ = Dirac "delta" function.

Equation (4.1-8) allows the output, $y(t)$, of an analog filter $H(s)$, for zero initial conditions, to be written in terms of the impulse response $\bar{w}(t)$:

$$\begin{aligned}
 y(t) &= \int_0^{\infty} \bar{w}(t-\tau) x_s(\tau) d\tau \\
 &= \int_0^{\infty} \bar{w}(t-\tau) x(\tau) \sum_{j=0}^{\infty} \delta(\tau-jT) d\tau \\
 &= \sum_{j=0}^{\infty} \bar{w}(t-jT) x(jT)
 \end{aligned} \tag{6.5.1}$$

if $H(s)$ is such that the numerator is of lower order than the denominator. If the numerator order of $H(s)$ is equal to the denominator order, $M = N$, then the impulse sampled input will also appear in the output modified by the factor p_N/q_N of Section 6.1.

The value at a sampling instant, $t = kT$, is given by

$$y(kT) = \sum_{j=0}^{\infty} \bar{w}((k-j)T) x(jT) \tag{6.5.2}$$

or

$$y_k = \sum_{j=0}^{\infty} \bar{w}_{k-j} x_j$$

upon suppressing functional notation.

Taking the z transform of this equation

$$\begin{aligned}
 Y(z) &= \sum_{k=0}^{\infty} y_k z^{-k} \\
 &= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \bar{w}_{k-j} x_j z^{-k}
 \end{aligned}
 \tag{6.5-3}$$

Letting

$$\begin{aligned}
 m &= k-j \\
 Y(z) &= \sum_{j=0}^{\infty} \sum_{m=j}^{\infty} \bar{w}_m x_j z^{-m} z^{-j} \\
 &= H_z(z) X(z)
 \end{aligned}
 \tag{6.5-4}$$

where

$$H_z(z) = \sum_{m=0}^{\infty} w(mT) z^{-m}$$

as

$$\bar{w}_{-j} = 0 \text{ for } j = 1, 2, \dots, \text{ and}$$

$$X(z) = \sum_{j=0}^{\infty} x_j z^{-j},$$

Fortunately, for linear, constant coefficient systems $H_z(z)$ can be expressed in "closed form" as the ratio of two polynomials in z^{-1} (See Section 6.2). As such it can be implemented as a recursive digital filter relating the input and output sample sequences:

$$y_j = - \sum_{n=1}^N b_n y_{j-n} + \sum_{m=0}^M a_m x_{j-m} \quad (6.5-5)$$

All the considerations relating the frequency responses of the given transfer function $H(s)$ and the digital transfer function $H_z(z)$ in section 6.1 apply as the definitions of $H_z(z)$ are identical in both sections.

It should be noticed that $H_z(z)$ is different from $H(\omega)$ by a factor of $1/T$ for band-limited transfer functions over $0 \leq |\omega| \leq \omega_z/2$. In the sample and hold reconstruction methods of Section 7, this is unimportant as the various holds have a factor of T to spare in their transfer functions. In fact the utility of the impulse sampling model is found primarily in sample and hold systems.

When the output is desired delayed then the transform method of Section 6.4 should be used with $t = kT - \tau$ in (6.5-2).

6.6 BILINEAR z TRANSFORMATION (DATA PROCESSING FORM)

Previously considered was the application of the z transform to a linear continuous network impulse response to obtain approximately the same frequency response with a digital filter. The output sampled sequence was given in terms of the inverse z transform of the product of $\bar{H}_z(z) X(z)$. The frequency response to a sampled sinusoidal input sequence of frequency $\omega = 2\pi f$ was found to be $\bar{H}_z(z)$ with $z = e^{i\omega T}$. Obtaining the proper frequency response required sophisticated, but simple, modification. Difference equations were analyzed and found to be describable by a z transform transfer function providing a convenient method for mechanizing the continuous network by means of a digital computer.

On a more abstract basis any rational function (a ratio of two polynomials) of z^{-1} can be considered as a candidate for a z transfer function. Setting $z = e^{i\Omega T}$ gives the frequency response as a function of Ω . Merely equate coefficients of the numerator and denominator of this rational function of z^{-1} to those in equation 3.1-1 which is its implementation as a linear difference equation, a recursive linear filter.

With these thoughts in mind a search may be made for transformations of s in the transfer function of a continuous system, $H(s)$, such that the imaginary axis of s , namely $s = i\omega$ is mapped onto the unit circle of another complex variable, say z , and the new function of z remains rational. Furthermore, this transformation would preferably be one-to-one. The z transform used previously was not one-to-one, many points of the $i\omega$ axis mapped into the unit circle; in fact all points modulo $2\pi/T = \omega_s$ in ω , become the same point in z .

A transformation which is one-to-one in frequency (and in fact everywhere) and maps $0 \leq \omega < \infty$, onto the unit circle in z , $0 \leq \Omega T \leq \pi$, and $-\infty \leq \omega \leq 0$ onto the lower half of the unit circle $-\pi \leq \Omega T \leq 0$, is

$$\begin{aligned} s &= \frac{z-1}{z+1} \\ &= \frac{1-z^{-1}}{1+z^{-1}} \end{aligned} \quad (6.6-1)$$

or

$$z = \frac{1+s}{1-s} \quad (6.6-2)$$

This is called the "bilinear" z transform.

It is seen that when $s = i\omega$

$$|z| = \frac{1+i\omega}{1-i\omega} = 1 \quad (6.6-3)$$

as stated. A diagram of the mapping follows (Figure 6.6-1)

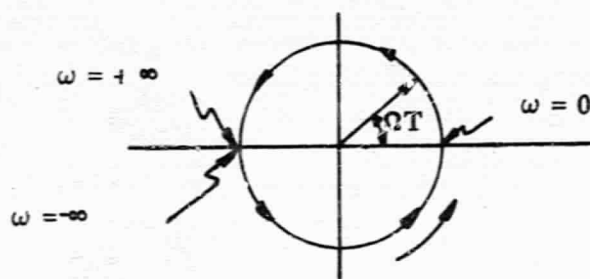


Figure 6.6-1. Mapping of "s" Plane to "z" Plane

The correspondence between ω and ΩT is found by letting $s = i\omega$ and $z = e^{i\Omega T}$.

$$\begin{aligned}
 i\omega &= \frac{e^{i\Omega T} - 1}{e^{i\Omega T} + 1} \\
 &= \frac{e^{\frac{i\Omega T}{2}} - e^{-\frac{i\Omega T}{2}}}{e^{\frac{i\Omega T}{2}} + e^{-\frac{i\Omega T}{2}}}
 \end{aligned}
 \tag{6.6-4}$$

and finally,

$$\omega = \tan \frac{\Omega T}{2}, \quad 0 \leq |\Omega T| \leq \pi
 \tag{6.6-5}$$

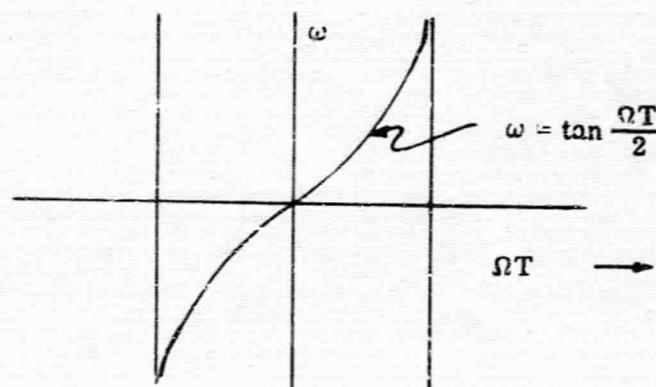


Figure 6.6-2. Frequency Relation Between ω and Ω

The transform is one-to-one in frequency as stated. It also is one-to-one elsewhere.

Consider the original continuous network transfer function $H(s)$ which is the frequency response when $s = i\omega$. As the variable ΩT traces out 0 to π , the frequency response takes on all its values though there is a non-linear warping of frequency axes given by (6.6-5) (see Figure 6.6-2). The shape of $H(i\omega)$ will be distorted in the Ω variable though amplitudes and phases for values of Ω and ω connected by relation (6.6-5) will be identical.

This means that certain analog filters such as Tchebychev, Butterworth and elliptical filters may be synthesized in the variable Ω . The frequencies of cut-off or stop-band must be specified in the variable Ω (of course T must be such that $|\Omega| \leq \pi/T$). By means of the transformation (6.6-5) the corresponding values of ω in the s plane are found. Construct $H(s)$ by normal s plane techniques; replace s by $(z^{-1} - 1)/(1 + z^{-1})$, and perform the algebra necessary to obtain this transfer function in z^{-1} as a ratio of two polynomials in z^{-1} . A recursive filter is obtained when $H(s)$ is rational.

A simple example best illustrates the method. Assume a low-pass RC filter which has a transfer function

$$H(s) = \frac{\omega_0}{s + \omega_0} \quad (6.6-6)$$

The magnitude of this function when $s = i\omega$ is

$$|H(i\omega)| = \frac{\omega_0}{\sqrt{\omega^2 + \omega_0^2}} = \frac{1}{\sqrt{1 + \left(\frac{\omega}{\omega_0}\right)^2}} \quad (6.6-7)$$

It has unity gain at $\omega = 0$, and is 3 dB down (1/2 power) at $\omega = \omega_0$.

For a low pass digital filter Ω is equated to some cut-off frequency Ω_0 . Compute the equivalent frequency in s by

$$\omega_0 = \tan \frac{\Omega_0 T}{2} \quad (6.6-8)$$

Then replace s :

$$\begin{aligned} H_z(z) &= H\left(\frac{1 - z^{-1}}{1 + z^{-1}}\right) \\ &= \frac{\omega_0}{\frac{1 - z^{-1}}{1 + z^{-1}} + \omega_0} \\ &= \frac{\omega_0 (1 + z^{-1})}{1 - z^{-1} + \omega_0 (1 + z^{-1})} \\ &= \frac{\frac{\omega_0}{1 + \omega_0} + \frac{\omega_0}{1 + \omega_0} z^{-1}}{1 - \left(\frac{1 - \omega_0}{1 + \omega_0}\right) z^{-1}} \end{aligned} \quad (6.6-9)$$

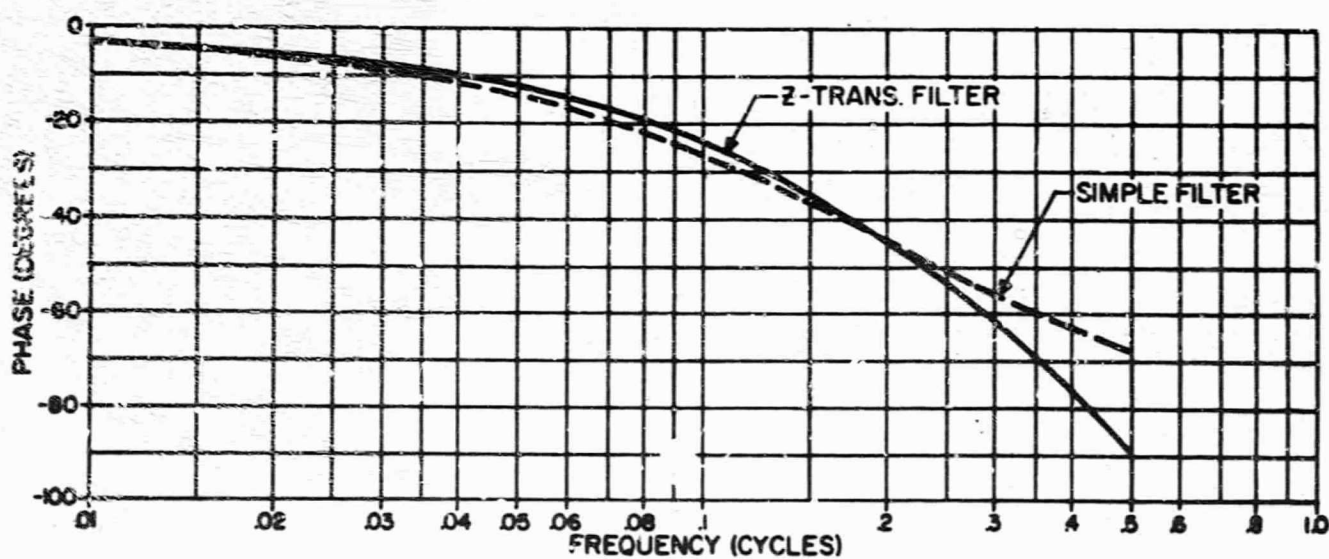
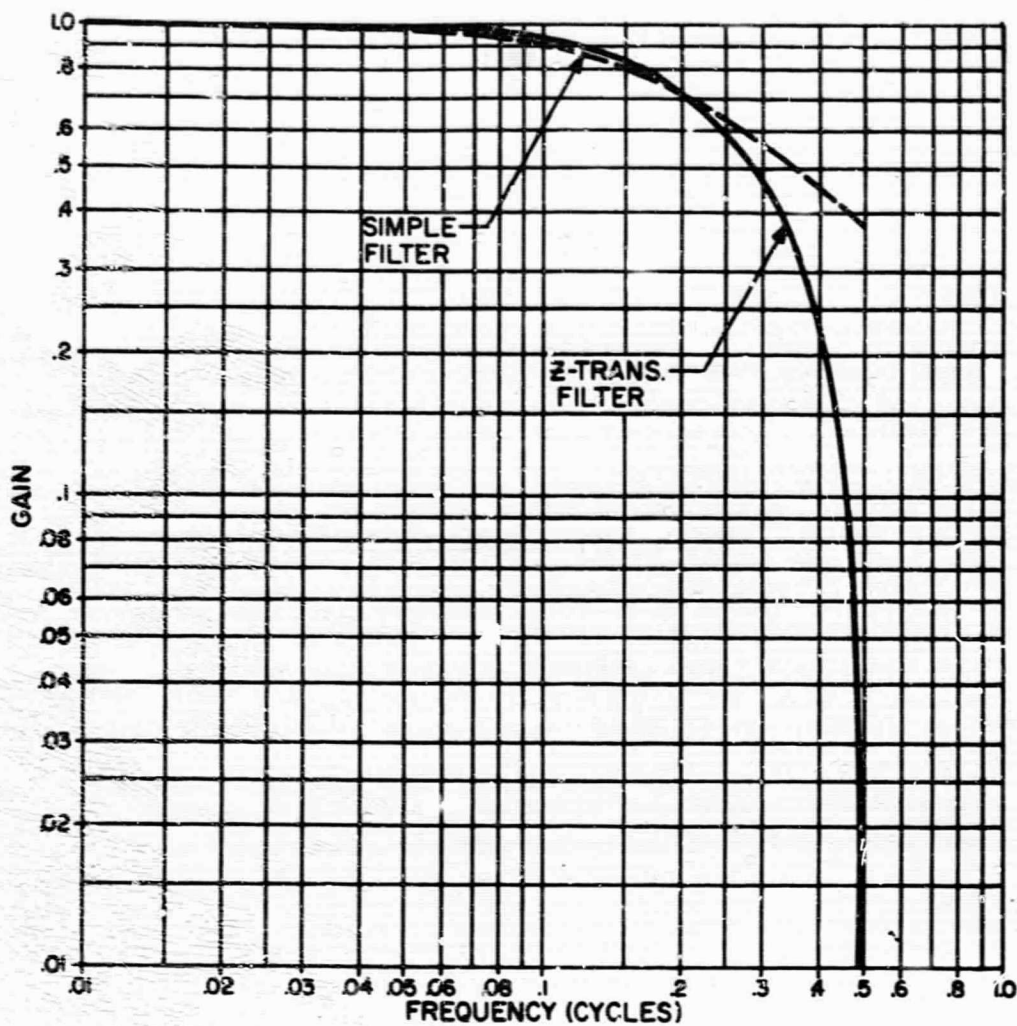


Figure 6.6-3. Bilinear z Transform Low-Pass Filter Design

The difference equation to implement the filter is

$$y_n - \left(\frac{1 - \omega_0}{1 + \omega_0} \right) y_{n-1} = \frac{\omega_0}{1 + \omega_0} x_n + \frac{\omega_0}{1 + \omega_0} x_{n-1} \quad (6.6-10)$$

and the design is complete. For designing complicated filters, the technique can be mechanized on a digital computer. A plot of the frequency response of a 2 KHz filter with a 10 KHz sampling rate is shown in Figure 6.6-3. The plot has been normalized so that 1 Hz is equivalent to 10 KHz. On the figure is shown the response of a 2 KHz continuous filter given by equation (6.6-6) for comparison.

6.6.1 General Low-Pass to Band-Pass, Band-Stop, and High-Pass Transformations of an Analog Filter Via the Bilinear z Transform

An analogue low pass filter, $H(s)$, with its 3 dB point at $\omega = 1$ is called a "normalized" low pass filter. It may be any kind of filter such as a Tchebychev, Butterworth, elliptical, etc., as long as it is realizable as the ratio of two polynomials in s .

It is well known that certain transformations of the variable s produce, from a "normalized" low-pass filter, low-pass, high-pass, band-pass, and band-stop analog filters with similar roll-off and ripple characteristics.

The appropriate transformations to go from our normalized low pass designs are:

$$\begin{aligned} s &\Leftarrow \frac{s}{\omega_u} && : \text{low-pass to low-pass,} \\ s &\Leftarrow \frac{s^2 + \omega_u \omega_l}{s(\omega_u - \omega_l)} && : \text{low-pass to band-pass,} \\ s &\Leftarrow \frac{s(\omega_u - \omega_l)}{s^2 + \omega_u \omega_l} && : \text{low-pass to band-stop,} \\ s &\Leftarrow \frac{\omega_l}{s} && : \text{low-pass to high-pass,} \end{aligned} \quad (6.6.1-1)$$

where

$$\begin{aligned} s &\Leftarrow y = \text{replace } s \text{ by } y \text{ in } H(s), \\ \omega_u &= \text{upper cut-off frequency in radians,} \\ \omega_l &= \text{lower cut-off frequency in radians.} \end{aligned}$$

Under these transformations stable filters remain stable as the real part of the new s variable roots remain in the left hand plane.

All the previous remarks are well known to students of circuit synthesis as well as the experts.

To go to the z transform of a sampled data filter that has the same cut-off frequencies and identical ripple in pass and stop bands as the continuous filter, merely replace the variable s in the appropriate kind of filter by $(1 - z^{-1})/(1 + z^{-1})$ as mentioned before. Furthermore, the frequencies ω_u and ω_l above (equation 6.6.1-1) are computed from

$$\omega_u = \tan \frac{\Omega_u T}{2} \quad (6.6.1-2)$$

and

$$\omega_l = \tan \frac{\Omega_l T}{2}$$

where Ω_u and Ω_l are the desired cut-off upper and lower frequencies of the sampled data filter (See Section 6.6).

Rather than perform this two step procedure on a normalized low pass filter both operations can instead be done in a single step. Replace s in the equations by the bilinear z form $(1 - z^{-1})/(1 + z^{-1})$ which gives a direct transform to the appropriate digital filter (after simplification to the ratio of two polynomials in z^{-1}). Of course, ω_u and ω_l are still computed by the relation (6.6.1-2) above.

The direct z forms to convert a normalized low-pass filter to the other forms below:

$$s \Leftarrow \frac{1}{\omega_u} \frac{1 - z^{-1}}{1 + z^{-1}} \quad : \text{low-pass to low-pass,}$$

$$s \Leftarrow \frac{(1 + \omega_u \omega_l) - 2(1 - \omega_u \omega_l) z^{-1} + (\omega_u \omega_l) z^{-2}}{(\omega_u - \omega_l)(1 - z^{-2})} \quad : \text{low-pass to band-pass,} \quad (6.6.1-3)$$

$$s \Leftarrow \frac{(\omega_u - \omega_l)(1 - z^{-2})}{(1 + \omega_u \omega_l) - 2(1 - \omega_u \omega_l) z^{-1} + (\omega_u \omega_l) z^{-2}} \quad : \text{low-pass to band-stop,}$$

$$s \Leftarrow \omega_u \frac{1 + z^{-1}}{1 - z^{-1}} \quad : \text{low-pass to high-pass,}$$

where, as before, the symbol " $\Leftarrow$ " means, replace s in the normalized low-pass design by the symbols to the right of " $\Leftarrow$ ".

The utility of these direct conversions depend on the method of implementing the replacement. Paper and pencil methods probably are easier with the direct conversion procedure. Machine methods probably are easiest with a two step process as two subroutines can be written; each to handle a part of the process

The method is illustrated by implementing a band-pass third order Butterworth.

A third order Butterworth normalized to unit cut-off frequency in radians has a magnitude response given by

$$|H(j\omega)| = \frac{1}{\sqrt{1 + \omega^{2n}}} \quad (6.6.1-4)$$

with $n = 3$. In terms of its poles in the stable left hand plane, the Laplace transfer function. can be written as:

$$H(s) = \frac{1}{(s + a)(s + a^*)(s + b)} \quad (6.6.1-5)$$

where

$$a = \alpha + i\beta,$$

$$= \frac{1}{2} + i \frac{3}{2},$$

$$a^* = \frac{1}{2} - i \frac{3}{2},$$

and

$$b = +1.$$

To convert to a band-pass filter the transformation used is

$$s = \frac{(1 - \omega_u \omega_l)^{-2} (1 - \omega_u \omega_l) z^{-1} + (1 + \omega_u \omega_l) z^{-2}}{(\omega_u - \omega_l) (1 - z^{-2})} \quad (6.6.1-3)$$

Defining:

$$c = \frac{1 + \omega_u \omega_l}{\omega_u - \omega_l}, \quad \omega = \tan \frac{\Omega T}{2}, \quad (6.6.1-6)$$

$$d = \frac{1 - \omega_u \omega_l}{\omega_u - \omega_l};$$

gives

$$s \Leftarrow \frac{c - 2dz^{-1} + cz^{-2}}{1 - z^{-2}} \quad (6.6.1-7)$$

Replacing s gives the z for:

$$H_z(z) = H(s(z)) \quad (6.6.1-8)$$

$$H_z(z) = \frac{1}{\left(\frac{c - 2dz^{-1} + cz^{-2}}{1 - z^{-2}} + a \right) \left(\frac{c - 2dz^{-1} + cz^{-2}}{1 - z^{-2}} + a^* \right) \left(\frac{c - 2dz^{-1} + cz^{-2}}{1 - z^{-2}} + b \right)}$$

Simplifying to a polynomial ratio, gives the z transfer function for the band-pass filter with upper and lower cut-offs, Ω_u and Ω_l respectively. Upon simplification; it is found that the b_0 term obtained is not in the standard form

$$H_z(z) = \frac{\sum_{n=0}^6 a_n z^{-n}}{\sum_{n=0}^6 b_n z^{-n}}, \quad (6.6.2-9)$$

where b_0 equals one. Therefore, the numerator and denominator must be divided by b_0 to obtain the standard form. The above mentioned "simplifications" are quite laborious to perform by hand. A computer program can be written to perform these operations. Amplitude and phase plots of the digital filter obtained by this method are given in Figure 6.6.1-1. The sampling frequency was assumed normalized to 1 Hz and the upper and lower cut-offs at $f_u = .3$, and $f_l = .2$ Hz.

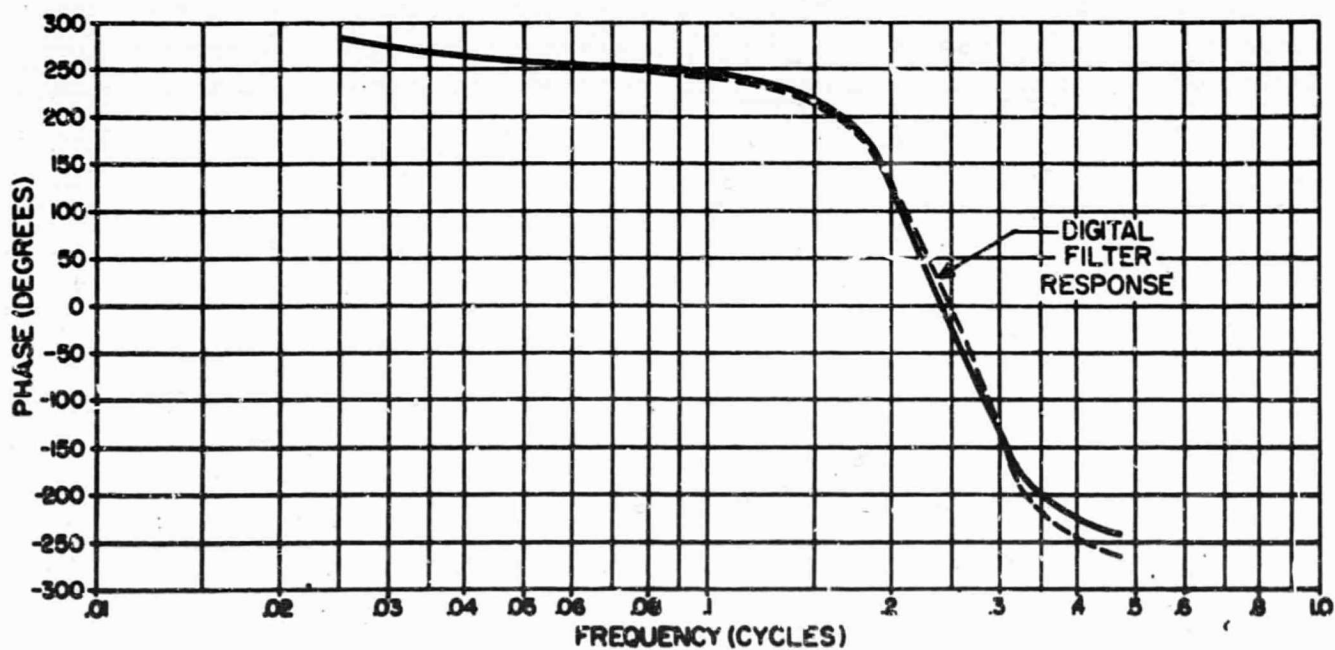
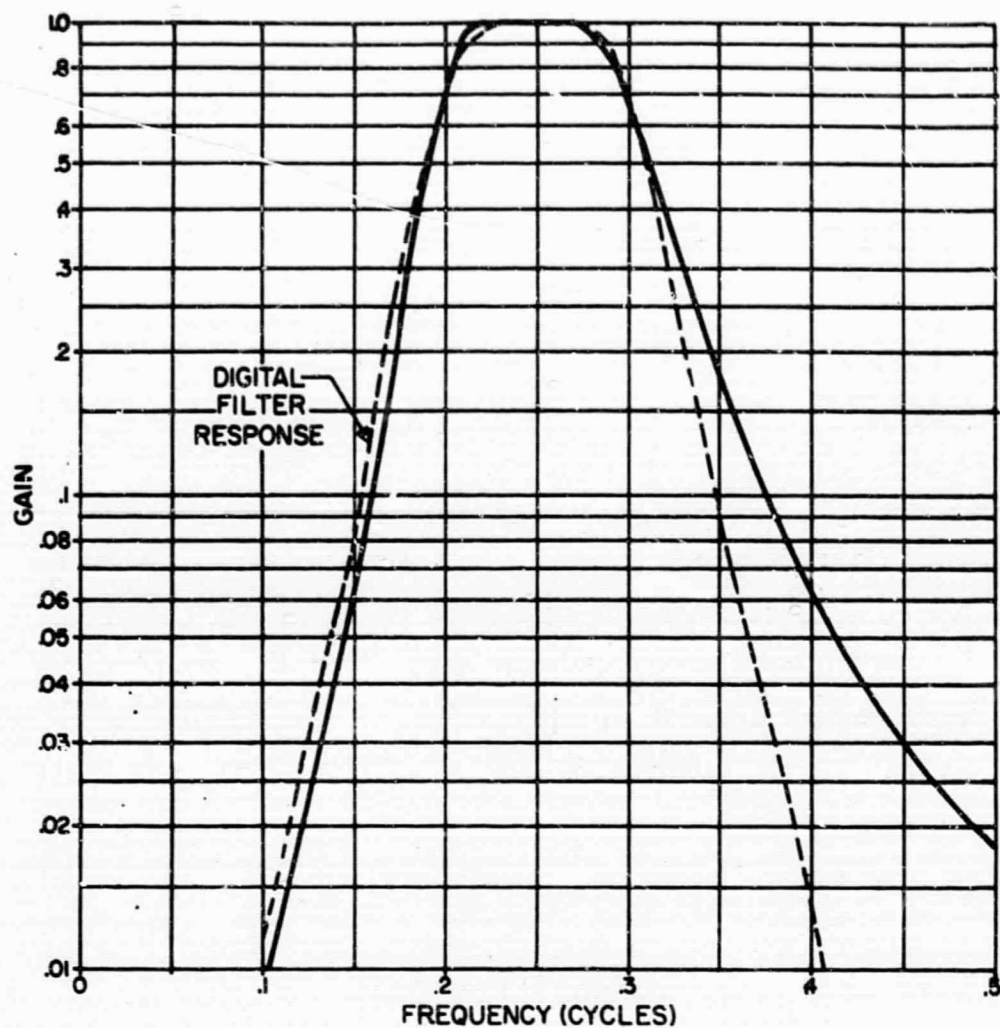


Figure 6.6.1-1. Bilinear z Transform Bandpass Filter Design via Analog Low-pass to Digital Bandpass Transformation

6.6.2 General Transformations of Normalized Low-Pass Digital Filter Designs to Low/High-Pass, Band-Pass and Band-Stop Filters in the Digital Filter Domain

The "bilinear z transform" has been used to convert well known continuous transfer function designs to digital filter designs.

Continuous, low-pass, normalized filters such as Butterworth, Tchebychef, and elliptical, could, after conversion by well known transformations of the "s" plane, (Section 6.6.1), be converted to low-pass, high-pass, band-pass, and band-stop filters with their significant properties intact. The bilinear z transform was then applied, after "pre-warping" significant frequencies in the "s" plane design, to convert to digital filters having the same significant properties as the original analog filter. Stability in the s plane resulted in stability in the z plane.

Furthermore, the bilinear z transform is a one-to-one mapping. One point in the s plane corresponds to one point in the z plane, and vice versa.

This one-to-one correspondence of the bilinear z transform can be exploited, in reverse, to develop transformation rules in the "z" plane so that digital low-pass designs, designs possibly made independently of any "s" plane considerations, can be converted to low-pass, high-pass, band-pass, and band-stop, filters with similar properties.

Utilization of the transformation, "in reverse" is a simple matter. The bilinear z transform is

$$s = \frac{z - 1}{z + 1} \quad (6.6.2-1)$$

or z in terms of s as

$$z = \frac{1 + s}{1 - s}$$

Therefore, if a stable, constant coefficient, digital filter $H_z(z)$ is given, a stable continuous filter can be developed,

$$H_z\left(\frac{1 + s}{1 - s}\right), \quad (6.6.2-2)$$

where the frequencies in the two domains are related (Section 6.6) by

$$\omega = \tan \frac{\Omega T}{2}, \quad (6.6.2-3)$$

Assuming a digital low pass filter, $H_z(z)$, with its "cut-off" frequency normalized at:

$$\frac{\Omega T}{2} = \frac{\pi}{4}, \quad (6.6.2-4)$$

or

$$\Omega_c = \frac{\pi}{2T},$$

$$= \frac{\omega_s}{4},$$

ω_s = radian sampling frequency;

then the corresponding continuous low pass filter,

$$H_z \left(\frac{1+s}{1-s} \right), \quad (6.6.2-2)$$

would have its cut-off at $\omega = 1$. Namely, a normalized, stable, low-pass analog filter design has been produced. Now there is liberty to use s plane transformations (Section 6.6.1) to the other band-pass and band-stop designs, and then use the bilinear z transform to reconvert to the equivalent z transform digital filters.

Diagrammatically, the route in Figure 6.6.2-1 is taken.

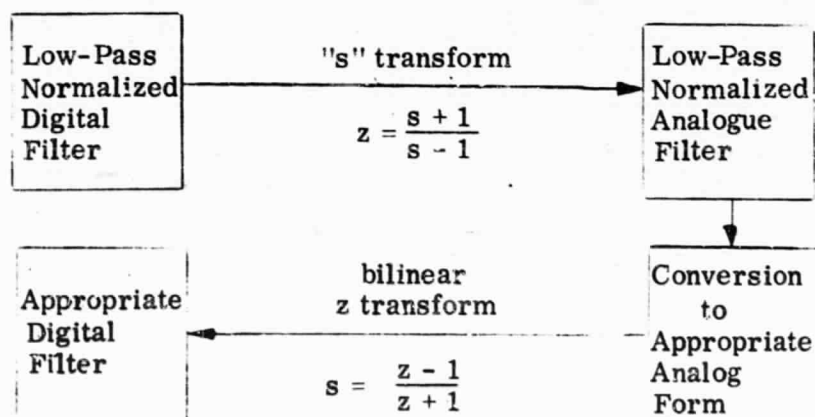


Figure 6.6.2-1. Low-Pass Digital Filter to Other Digital Filter Transformation Route

Actually, it is not necessary to go through this procedure, formally, each time the transform from one digital design to the next is desired. Transformation can be developed directly in the " z " variable from the known " s " plane transformations (equation 6.6.1-3).

The procedure for going from a normalized low-pass design to a high-pass at cut-off ω_c is demonstrated.

First, the normalized analog filter is derived from the digital one under the transformation.

$$z = \frac{1+s}{1-s}$$

namely;

$$H_z(z) = H_z \left(\frac{1+s}{1-s} \right) \quad (6.6.2-2)$$

Next, transform the analog filter to high pass via the transformation (equation 6.6.1-3)

$$s \Leftarrow \frac{\omega_c}{s}, \quad \omega_c = \text{low frequency cut-off}$$

where " $\Leftarrow$ " means replace the left hand side by the right side in transfer functions. Therefore, the high pass analog transfer function is

$$H_z \left(\frac{1 + \frac{\omega_c}{s}}{1 - \frac{\omega_c}{s}} \right) \quad (6.6.2-5)$$

Finally, reconvert the high pass analog design to a digital design, after adjusting ω_c so that the digital filter will have the proper cut-off Ω_c . That is, replace the s in the high-pass analog design

$$s \Leftarrow \frac{z-1}{z+1}$$

using

(6.6.2-6)

$$\omega_c = \tan \frac{\Omega_c T}{2}$$

and obtain the digital design

$$H_z \left(\frac{1 + \omega_\ell \left(\frac{z+1}{z-1} \right)}{1 - \omega_\ell \left(\frac{z+1}{z-1} \right)} \right) \quad (6.6.2-7)$$

The net result, is that a high-pass digital filter with a cut-off at $\omega_\ell = \tan(\Omega_c T/2)$ can be obtained from a normalized low pass ($\Omega_c = \omega_s/4$), by simply replacing z in the digital low pass:

$$z \Leftarrow \frac{1 + \omega_\ell \left(\frac{z+1}{z-1} \right)}{1 - \omega_\ell \left(\frac{z+1}{z-1} \right)}, \quad \omega_\ell = \tan \left(\frac{\Omega_c T}{2} \right); \quad (6.6.2-8)$$

or simplified

$$z \Leftarrow \frac{(\omega_\ell - 1) + (\omega_\ell + 1) z}{(\omega_\ell + 1) + (\omega_\ell - 1) z}$$

Hereafter, the s plane need not be discussed; the transformation in z is sufficient.

Using the same method to derive them, the four transformations in the z domain equivalent to the analog transformations (equation 6.6.1-3) are listed.

$$\begin{aligned} z &\Leftarrow \frac{(\omega_u - 1) + (\omega_u + 1) z}{(\omega_u + 1) + (\omega_u - 1) z} && : \text{Low-pass to Low-pass;} \\ z &\Leftarrow \frac{(\omega_\ell - 1) + (\omega_\ell + 1) z}{(\omega_\ell + 1) + (\omega_\ell - 1) z} && : \text{Low-pass to High-pass;} \\ z &\Leftarrow - \frac{(1-\alpha+\beta) + 2(\beta-1)z + (1+\alpha+\beta)z^2}{(1+\alpha+\beta) + 2(\beta-1)z + (1-\alpha+\beta)z^2} && : \text{Low-pass to Band-pass} \\ z &\Leftarrow \frac{(1-\alpha+\beta) + 2(\beta-1)z + (1+\alpha+\beta)z^2}{(1+\alpha+\beta) + 2(\beta-1)z + (1-\alpha+\beta)z^2} && : \text{Low-pass to Band-stop} \end{aligned} \quad (6.6.2-9)$$

with

$$\omega_u = \tan \left(\frac{\Omega_u T}{2} \right)$$

$$\omega_l = \tan \left(\frac{\Omega_l T}{2} \right)$$

$$\alpha = \omega_u - \omega_l$$

and

$$\beta = \omega_u \omega_l$$

Another convenient form of the same transformations is:

$$z^{-1} \Leftarrow \frac{(\omega_u - 1) + (\omega_u + 1) z^{-1}}{(\omega_u + 1) + (\omega_u - 1) z^{-1}} \quad : \text{Low-pass to Low-pass}$$

$$z^{-1} \Leftarrow - \frac{(\omega_l - 1) + (\omega_l + 1) z^{-1}}{(\omega_l + 1) + (\omega_l - 1) z^{-1}} \quad : \text{Low-pass to High-pass} \quad (6.6.2-10)$$

$$z^{-1} \Leftarrow - \frac{(1 - \alpha + \beta) + 2(\beta - 1) z^{-1} + (1 + \alpha + \beta) z^{-2}}{(1 + \alpha + \beta) + 2(\beta - 1) z^{-1} + (1 - \alpha + \beta) z^{-2}} \quad : \text{Low-pass to Band-pass}$$

$$z^{-1} \Leftarrow \frac{(1 - \alpha + \beta) + 2(\beta - 1) z^{-1} + (1 + \alpha + \beta) z^{-2}}{(1 + \alpha + \beta) + 2(\beta - 1) z^{-1} + (1 - \alpha + \beta) z^{-2}} \quad : \text{Low-pass to Band-stop}$$

with

$$\alpha = \omega_u - \omega_l,$$

$$\beta = \omega_u \omega_l,$$

and

$$\omega_u = \tan\left(\frac{\Omega_u T}{2}\right),$$

$$\omega_l = \tan\left(\frac{\Omega_l T}{2}\right),$$

as before.

Figure 6.6.2-2 illustrates the usage of these transformations on a "normalized" digital lowpass which was designed from an analog 3 pole Butterworth prototype by the methods of Section 6.6.1. Depicted are the direct transformations via equations (6.6.2-10) to a low pass filter at a cut-off of .3 Hz (a), a high pass filter at .2 Hz (b), and a band-stop filter with cut-offs at .2 and .3 Hz (c). The sampling frequency was assumed equal to unity. The band-pass design was not reproduced here as it is identical to that in Figure 6.6.1-1 which also illustrates the validity of the method.

These transformations may be performed on $H_z(z)$ in any realization, namely those of Section 3.7.

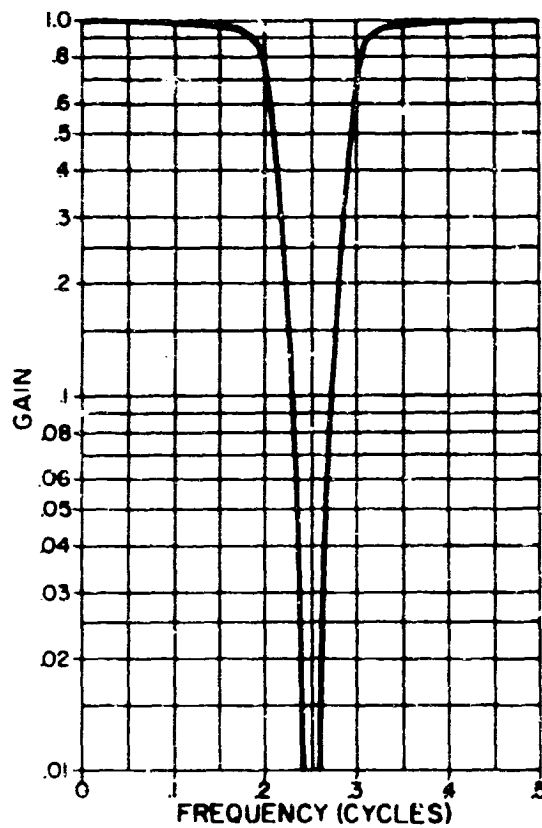
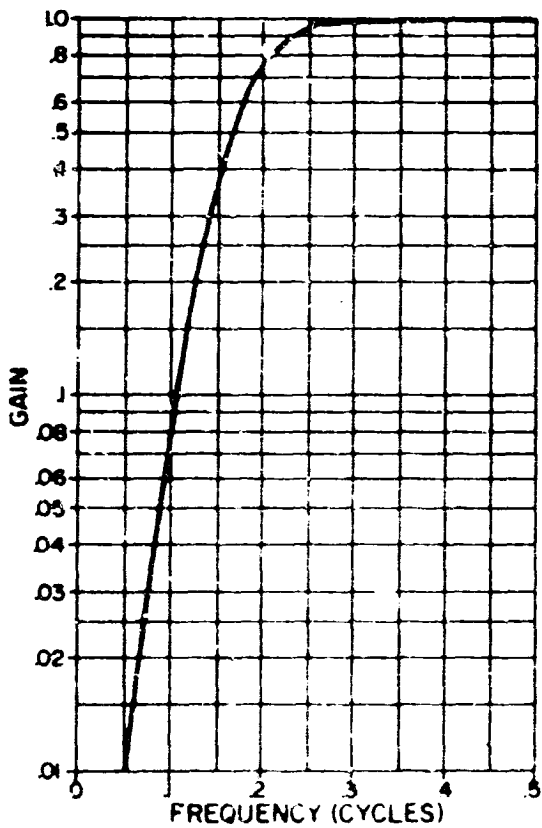
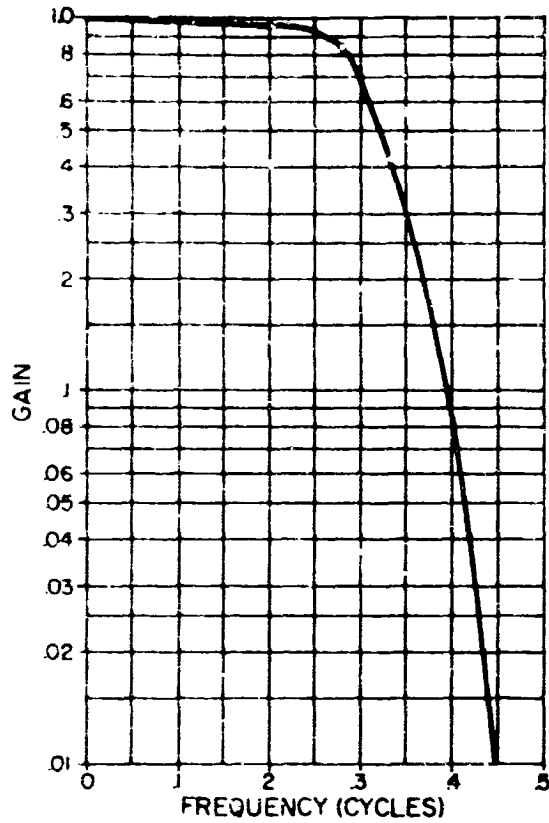


Figure 6.6.2-2. Direct Transformation of a 3 Pole "Normalized" Digital Low-Pass to Low-Pass, High Pass, and Band Stop Digital Filters

6.7 BILINEAR z TRANSFORM (SPECIAL FORM FOR NUMERICAL INTEGRATION)

Another method of creating a digital filter or numerical integration procedure for linear, constant coefficient, integro-differential equations is quite similar to the preceding method and is simple to implement though it requires a machine program to do the arithmetic for complicated transfer functions. It has the basic limitation of allowing operation to only about $1/16$ to $1/8$ of the sampling frequency, depending on the accuracy involved, and cannot be extended above these limits even with delay. Certain transfer functions, such as bandpass filters whose parameters: cut-off frequencies and band-widths, can be pre-adjusted (or "pre-warped") so that the "data processing" bi-linear transform of the previous section gives desirable results. Those are of special interest and will not be discussed here. See section 6.6.

The bilinear z transform used in this section is the replacement of s in $H(s)$ by a relation quite similar to that of the previous section:

$$s = \beta \frac{2}{T} \frac{z - 1}{z + 1} \quad (6.7-1)$$

where

T = data sampling interval

β = a constant which will be found most advantageous if set equal to 1.

The new sampled data transfer function $H_z(z)$ is given by

$$\begin{aligned} H_z(z) &= H(s) \\ &= H\left(\beta \frac{2}{T} \frac{z - 1}{z + 1}\right) \end{aligned} \quad (6.7-2)$$

As $H(s)$ is assumed to be given as the ratio of two polynomials in s ; the substitution gives a transfer function $H_z(z)$ that is a rational function of z (a ratio of two polynomials in z) and therefore describes a difference equation or recursive digital filter. The relationship between the frequency responses of the continuous filter $H(z)$ and the digital filter is of interest and follows developments of the preceding section.

It should be mentioned that this form of the bilinear z transform is one-to-one and produces stable digital filters. The transform is one-to-one as a given point in the s plane corresponds to a unique point in the z plane and conversely. Stability is guaranteed under the transform as the left half of the s plane maps into the interior of the unit circle of the z plane.

The relationship between the frequency responses of $H(s)$ and $H_z(z)$ is simply the connection between s on the imaginary axis, $s = i\omega$, and z on the unit circle, $z = e^{i\Omega T}$, ω is the frequency in the analog (s) domain and Ω is the frequency in the digital (z) domain. Given a value of Ω in the digital or sampled data domain, the corresponding analog frequency, ω can be calculated. At these corresponding values $H_z(e^{i\Omega T})$ and $H(i\omega)$ are identical.

Apparently it is only the relationship of the frequency scales that determines the behavior of the digital filter with respect to the analog filter. By setting

$$s = i\omega$$

and

$$z = e^{i\Omega T}$$

in the bilinear transform (eq. 6.7-1) it is found that

$$\omega = \beta \frac{2}{T} \tan \left(\frac{\Omega T}{2} \right) \quad (6.7-3)$$

Examination of this equation shows that the $i\omega$ axis is mapped into the unit circle in the z plane. Plus infinity in ω is mapped on to $\Omega_s T = \pi$, or $\Omega_s/2 = \pi/T$ (half the sampling frequency). Furthermore, at low frequencies ($\Omega T < \pi/4$)

$$\tan \left(\frac{\Omega T}{2} \right) \approx \frac{\Omega T}{2} \quad (6.7-4)$$

which gives $\omega \approx \beta \Omega$. Therefore β should be set equal to unity as stated previously, to have a decent transformation. Finally, it can be seen that this transformation is restricted to frequencies in the digital domain such that

$$\Omega \leq \frac{\pi}{4T}$$

or

$$(6.7-5)$$

$$\Omega \leq \Omega_s/8$$

for the linear approximation to remain valid. When more accurate approximations are desired the frequency range of Ω is further restricted.

The bilinear z transform is obviously quite easy to use and gives approximations good to about $1/8$ the data sampling frequency, Ω_s . If the transfer function being approximated varies rapidly the non-linearity of the mapping between ω and Ω can cause distortion of the approximation even below $\Omega_s/8$. It may be found that $\Omega_s/16$ or lower is the upper bound for the approximation region.

6.8 EXACT FITTING OF A TRANSFER FUNCTION BY A RECURSIVE FILTER AT $Q + 1$ POINTS

An arbitrary transfer function can be fitted exactly at $Q + 1$ points by a recursive filter transfer function. While the fit is exact at the points where the fit is specified, it is difficult to assess the error between sample points. Furthermore, as the number of sample points increases the order of the numerator and denominator of the recursive filter increases causing long computation times when using the filter. Finally, filter stability is not guaranteed though the filter may be altered to be stable at the expense of changing the phase characteristic; the magnitude will remain unchanged.

The exact fit problem in terms of a non-recursive transfer function means finding a function

$$H_A(\omega) = \frac{\sum_{m=0}^M a_m e^{-i\omega m T}}{\sum_{n=0}^N b_n e^{-i\omega n T}} \quad ; b_0 = 1, \quad (6.8-1)$$

over the interval $0 \leq \omega \leq \omega_s/2$, $\omega_s = 2\pi/T$, so that $H_A(\omega)$ equals $H(\omega)$ at $Q + 1$ points.

It may be noticed that $N + M + 1$ parameters have to be determined:

$$a_0, a_1, \dots, a_M$$

and

$$b_1, b_2, \dots, b_N; (b_0 = 1).$$

If there are samples of the transfer function $H(\omega)$ over the interval $0 \leq \omega \leq \omega_s/2$ and 0 and $\omega_s/2$ are included as sample points then there will be an even number of independent values of the real and imaginary parts (or amplitude and phase). If $\omega_s/2$ or zero, but not both were included then there will be an odd number of independent values. If both $\omega_s/2$ and zero were excluded there will again be an even number of independent values. This occurs because $H_A(\omega)$ must be real at $\omega = 0$ and $\omega_s/2$ due to its form. Transfer functions of continuous systems satisfy the reality condition at $\omega = 0$, but are arbitrary at other ω 's. The implication is that the imaginary part of $H(\omega)$ cannot fit at $\omega_s/2$ unless its phase is $\pm \pi$ (or ± 180 degrees). Other points will be fitted exactly.

Including the points $\omega = 0$ and $\omega = \omega_s/2$ sampling can be as in Figure 6.8-1.

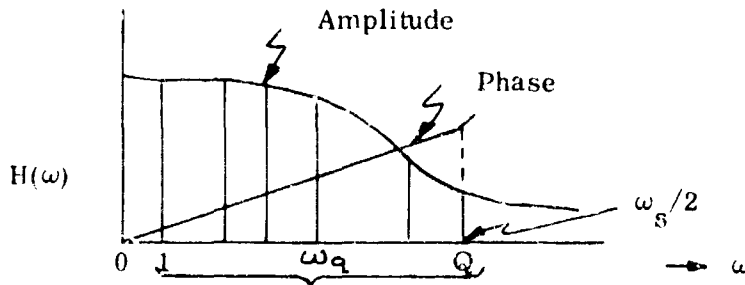


Figure 6.8-1. Transfer Function Sampling

It can be seen that there are real values at $\omega = 0$ and $\omega_s/2$ which are two independent numbers and $Q-1$ pairs of amplitude and phase values giving $2Q$ total independent numbers when sampled at $Q+1$ points. It is therefore expected that the number of coefficients a_k and b_k , $N+M+1$, must be equal to $2Q$.

The case where the sample point ω_Q does not lie at $\omega_s/2$ (or zero) gives $2Q+1$ values of phase and amplitude and, therefore, in this instance, $2Q+1 = N+M+1$.

To find the coefficients set the desired transfer function $H(\omega)$ equal to the approximation at the points ω_q , which are not necessarily equi-spaced. This gives

$$H(\omega_q) = \frac{\sum_{m=0}^M a_m e^{-i\omega_q mT}}{\sum_{n=0}^N b_n e^{-i\omega_q nT}}, \quad b_0 = 1 \quad (6.8-2)$$

or

$$\sum_{n=0}^N b_n H(\omega_q) e^{-i\omega_q nT} = \sum_{m=0}^M a_m e^{-i\omega_q mT} = H(\omega_q). \quad (6.8-3)$$

Now

$$H(\omega_q) e^{-i\omega_q nT} = [f(\omega_q) + ig(\omega_q)] e^{-i\omega_q nT} \quad (6.8-4)$$

if $H(\omega)$ is written in terms of its real and imaginary parts, $f(\omega)$ and $g(\omega)$, respectively. Carrying out the multiplication.

$$H(\omega_q) e^{-i\omega_q nT} = [f(\omega_q) \cos \omega_q nT + g(\omega_q) \sin \omega_q nT] \quad (6.8-5)$$

$$+ i [g(\omega_q) \cos \omega_q nT - f(\omega_q) \sin \omega_q nT]$$

Writing (6.8-5) in terms of its real and imaginary parts gives

$$x e_q \frac{n}{s} : \sum_{n=1}^N b_n (f(\omega_q) \cos \omega_q nT + g(\omega_q) \sin \omega_q nT) - \sum_{m=0}^M a_m \cos \omega_q mT = f(\omega_q) \quad (6.8-6)$$

$$y e_q \frac{n}{s} : \sum_{n=1}^N b_n (g(\omega_q) \cos \omega_q nT - f(\omega_q) \sin \omega_q nT) + \sum_{m=0}^M a_m \sin \omega_q mT = -g(\omega_q).$$

There are four sets of values for x and y , the number of equations, depending on the inclusion of the endpoint samples, zero and $\omega_s/2$. A table of the values for x and y , and their total versus sampling at the endpoints follows. In each case $Q + 1$ samples of $H(\omega)$ are assumed.

H (0) Present	H($\omega_s/2$) Present	x	y	v = x + y
Yes	Yes	Q + 1	Q - 1	2Q
Yes	No	Q + 1	Q	2Q + 1
No	Yes	Q + 1	Q	2Q + 1
No	No	Q + 1	Q + 1	2Q + 2

TABLE 6.8-1

Number of Simultaneous Equations for Various End Point Conditions

The equations (6.8-6) are linear in the variables a_m and b_n and may be solved by standard linear equation methods using library routines from any computer installation, or paper and pencil methods if the total number of equations is small (about 4 or less). Obviously, $v = N + M + 1$; the total number of equations must be equal to the number of unknown a_m 's and b_n 's (b_0 excluded).

When the number of samples of $H(\omega)$ is "large", the solution of the simultaneous equations (6.8-6) becomes difficult due to machine round-off errors. "Large", is about 7 or 8 samples which produces 14 or 16 simultaneous equations if endpoint samples are included. A reduction in the number of equations that must be solved is possible in the special, but important, case of equi-spaced sampling of $H(\omega)$. This reduction is to about half that for unequal sampling, namely, N equations.

6.8.1 Exact fit at Equi-Spaced Points

When the data is given at equi-spaced points the problem can be simplified, as far as the order of the system of simultaneous equations that must be solved. As usual, certain complications arise when the number of coefficients a_m and b_n to be determined, $N + M + 1$, is even or odd.

In the even case the transfer function samples are taken at

$$\omega_q = \frac{\omega_s}{2} \frac{q}{Q}, \quad (6.8.1-1)$$

$$= \frac{\pi q}{QT}, \quad q = 0, 1, \dots, Q;$$

where $Q + 1$ is the number of transfer function samples and $2Q = N + M + 1$. It may be noticed that the endpoint samples, $\omega = \omega_s/2$, are included.

For odd $N+M+1 = 2Q+1$, the samples are taken at

$$\begin{aligned}\omega_q &= \frac{\omega_s}{2} \cdot \frac{2q}{2Q+1} \\ &= \frac{2\pi q}{(2Q+1)T} \quad , \quad q = 0, 1, \dots, Q;\end{aligned}\tag{6.8.1-2}$$

with $Q+1$ samples as before.

The endpoint sample

$$\omega_Q = \frac{2Q}{2Q+1} \cdot \frac{\omega_s}{2}\tag{6.8.1-3}$$

does not allow inclusion of a sample at $\omega_s/2$.

Sampling at different places when $N+M+1$ is odd or even is necessary to allow use of the orthogonality relations, which give the subsequent simplification in solution of the simultaneous equations. The relations of interest are

$$\frac{1}{2Q} \sum_{q=-(Q-1)}^Q e^{i\omega_q(j-k)T} = \delta_N(j-k), \text{ the Kronecker delta}\tag{6.8.1-4a}$$

when $N+M+1 = 2Q$ (even) with

$$\omega_q = \frac{\omega_s}{2} \cdot \frac{q}{Q} \quad ,$$

and

$$\frac{1}{2Q+1} \sum_{q=-Q}^Q e^{i\omega_q(j-k)T} = \delta_N(j-k)\tag{6.8.1-4b}$$

when $N+M+1 = 2Q+1$ (odd) with

$$\omega_q = \frac{\omega_s}{2} \cdot \frac{2q}{2Q+1} \quad .$$

Obviously the discrete Fourier transform, DFT2, of Section 2.6.1 is being used.

In the rest of the work it will be assumed that $N + M + 1 = 2Q$ (even). The odd case is carried out in an identical manner. From equation 6.8-3 the relation

$$\sum_{n=0}^N b_n H(\omega_q) e^{-i\omega_q nT} = \sum_{m=0}^M a_m e^{i\omega_q mT}, \quad b_0 = 1. \quad (6.8.1-5)$$

can immediately be written. Assuming $\omega_s q/2Q$; $q = 0, 1, \dots, Q$, multiply both sides of this equation by $e^{i\omega_q kT/2Q}$ and sum over q as in (6.8.1-4). Then

$$\sum_{n=0}^N b_n \frac{1}{2Q} \sum_{q=-(Q-1)}^Q H(\omega_q) e^{i\omega_q (k-n)T} = \sum_{m=0}^M a_m \frac{1}{2Q} \sum_{q=-(Q-1)}^Q e^{i\omega_q (k-m)T} \quad (6.8.1-6)$$

Letting k take on the values 0 to M , and using relation (6.8.1-4) it is found that

$$a_k = \sum_{n=0}^N w_{k-n} b_n; \quad k = 0, 1, \dots, M; \quad (6.8.1-7)$$

where

$$w_j = \frac{1}{2Q} \sum_{q=-(Q-1)}^Q H(\omega_q) e^{i\omega_q jT} \quad (6.8.1-8)$$

has been defined. The a_k 's in terms of the b_n 's have been found. Note, before passing, that if $H(\omega)$ is written in terms of amplitude and phase:

$$H(\omega) = A(\omega) e^{i\phi(\omega)} \quad (6.8.1-9)$$

that

$$w_j = \frac{1}{2Q} A(0) \cos \phi(0) + \frac{1}{2Q} A(\omega_Q) \cos \phi(\omega_Q) + \frac{1}{Q} \sum_{q=1}^{Q-1} A(\omega_q) \cos((\omega_q)jT + \phi(\omega_q)), \quad (6.8.1-10)$$

as $H(-\omega) = H^*(\omega)$. The w_j 's are real.

To determine the b_n 's let k take on the values $M+1, M+2, \dots, M+N$, and arrive at

$$\sum_{n=1}^N w_{k-n} b_n = -w_k, \quad k = M+1, M+2, \dots, M+N; \quad (6.8.1-11)$$

where it has been remembered that $b_0 = 1$.

As the w_j 's are known, it is seen that (6.8.1-11) is a system of N simultaneous linear equations in N unknowns, the b_n 's, $n = 1, 2, \dots, N$. The number of simultaneous equations that must be solved has been reduced from $N+M+1$ to N .

Recapitulating, the unknown b_n 's are determined from (6.8.1-11) and then inserted in (6.8.1-7) to produce the a_m 's.

A useful piece of information when computing the w_j 's is that they are periodic in j with a period $2Q = N+M+1$, that is

$$w_{j \pm 2Q} = w_j. \quad (6.8.1-12)$$

This means that only $2Q = N+M+1$ of them are needed and the rest are obtained from (6.8.1-12). For example, if w_{-j} is desired then w_{2Q-j} can be used if it had been computed previously. This fact is of importance when computing the coefficients in (6.8.1-7) and (6.8.1-11).

Figure (6.8.1-1) shows the result of an exact fit using this method for $N = 8, M = 7$, and $Q = 8$, and for $N = 4, M = 3$, and $Q = 4$. It will be noticed that the fit for $N = 8, M = 7$ is astonishingly good.

6.8.2 Stability of Exact Fit Method

If it is found that the approximate transfer function obtained

$$H_A(\omega) = \frac{\sum_{m=0}^M a_m z^{-m}}{\sum_{n=0}^N b_n z^{-n}} \quad ; \quad b_0 = 1 \quad (6.8.2-1)$$

is unstable; namely by computing the roots of the denominator of $H_A(\omega)$, then corrective action may be taken. This corrective action will preserve the magnitude characteristic but will obviously alter the phase approximation.

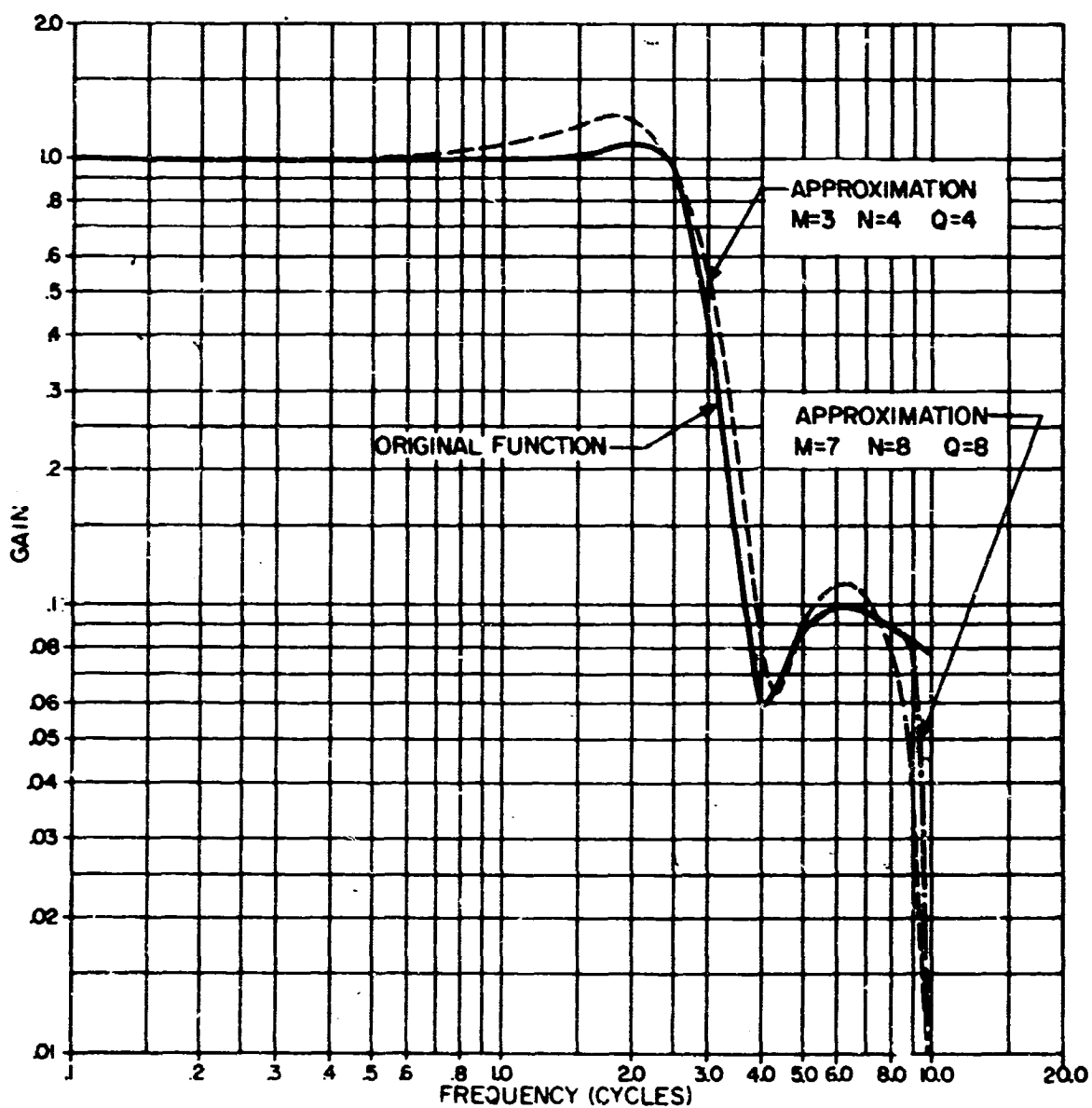


Figure 6.8.1-1. Exact Fit of Amplitude Characteristic of the Saturn 1B 203 SIB Stage Roll Attitude Rate Network

Suppose the denominator is factored into the form

$$\sum_{n=0}^N b_n z^{-n} = (1-\beta_1 z^{-1}) (1-\beta_2 z^{-1}) \dots (1-\beta_N z^{-1}) \quad (6.8.2-2)$$

where certain β_k 's have a magnitude greater than one. Then the filter is unstable (Section 3.4.2).

To correct this instability the factor corresponding to the root outside the unit circle may be modified. This modification consists of replacing the term $(1-\beta_j z^{-1})$ where β_j is a root whose magnitude is greater than one by

$$\beta_j^* (1 - \frac{1}{\beta_j^*} z^{-1}), \quad * = \text{complex conjugation.} \quad (6.8.2-3)$$

The magnitude of this term is the same as the replaced term when $z = e^{j\omega T}$. As the magnitude of the product of terms is the product of the magnitudes of the individual terms the magnitude of (6.8.2-1) has not been affected. Obviously, the new root, $1/\beta_j^*$, lies within the unit circle.

6.9 ZERO FREQUENCY CORRECTION

In many applications of the preceding techniques, and those to be mentioned in Section 7, it will be found that the value of the digital transfer function at zero frequency is not exactly the value desired due to aliasing when z transforming. In particular, differentiating techniques usually require a zero value at zero frequency.

A quick and useful remedy for slight errors is simply adding (subtracting) the necessary zero frequency correction from the entire digital transfer function $H_z(z)$. That is, if the desired value at zero frequency is $H_D(0)$, and the digital transfer function has the actual value $H_z^A(0)$ at zero frequency, then the corrected digital transfer function is

$$H'_z(z) = H_z(z) + H_D(0) - H_z^A(0)$$

which can be rearranged into its original form of realization (direct, partial fraction, etc.)

Without this correction many filtering applications will be found to have gross errors in the low frequency components even though the transfer function error at zero frequency is relatively small.

7. SYNTHESIZING OF RECURSIVE NUMERICAL INTEGRATION DIGITAL FILTERS WITH SMALL DELAY AND DATA RECONSTRUCTION

Preceding work has demonstrated the impossibility of producing linear numerical integration procedures that do not have delay unless the transfer function of the continuous linear system is bandlimited by half the sampling frequency. Of course, the high pass continuous system is also considered as bandlimited if the constant level is approximately reached by half the sampling frequency (Section 4. 3). The delay incurred using numerical integration procedures, which limits "black box" simulation, is caused basically by the requirement of using samples of a bandlimited process. It has been mentioned that infinite delay is required to reconstruct, or filter, a bandlimited function from the sampled values. Of course, approximate reconstruction with a finite delay with arbitrary accuracy can be obtained (Section 5). It should not be expected that any integration scheme would work in general without large delay over wide bandwidths using slow sampling.

Two basic numerical integration schemes will be discussed having various merits but achieving results by narrowing the approximation bandwidth:

1. Reconstruct the input using polynomial techniques. Recursive digital filters are obtained.
2. Low-pass filter preceding the original system function to achieve either of two equivalent goals: bandlimit the transfer function or reconstruct the input. Recursive digital filters are obtained.

These techniques have the distinct advantages of having relatively small delay and allowing recursive digital filter implementation. These advantages are accrued at the expense of only allowing approximation of the analog filter operation to about $1/8$ to $1/4$ of the sampling frequency for reasonable approximations. More accurate approximations lower the frequency to which the transfer function can be approximated for a given delay and allow usage of only the bandpass reconstruction method.

To discuss these methods a slight detour must be taken into the matter of data reconstruction

7.1 DATA RECONSTRUCTION TECHNIQUES

A subject of much practical interest is the reconversion of sampled data to a continuous analog waveform. Not only is this a useful enterprise in itself, but certain techniques found applicable are also quite fruitful in numerical integration and synthesis problems.

Methods to be described fall in a general category of linear, time independent methods. An example of a time dependent method is one that uses the convergence factors of Section 5. 7 on the Fourier transform (2.1.1-7), or reconstruction formula (2.1.1-9), for bandlimited functions to eliminate ripples caused by a finite sampling time. These time dependent methods are usually not useful for long data spans or for "simple" data processing. ("Adaptive" methods that use data characteristics are not included in this comment.)

All methods to be described have the property of limiting the output signal bandwidth to a greater or lesser extent. The more the bandwidth is limited to the "ideal" the greater the delay. However, certain methods are easier to implement and when used in conjunction with a subsequent digital filtering operation, result in less weights in the filter.

7.1.1 Time Invariant Reconstruction

A well known technique, "sample and hold reconstruction", plus a method not so familiar, use of a bandpass filter with a pulsed input, fall within a general category. Namely, a general linear form for reconstruction is used:

$$x_R(t) = \sum_{k=0}^{\infty} x_k \varphi(t-kT) \quad (7.1.1-1)$$

where the x_k are the data samples, $\varphi(t)$ is a "reconstruction" function and T the sampling interval. For bandlimited functions the "ideal" reconstruction function is

$$\varphi_I(t) = \frac{\sin \left[\frac{\omega_s}{2} (t-kT) \right]}{\left[\frac{\omega_s}{2} (t-kT) \right]} \quad (2.1.1-9)$$

as given in Section 2.1.1. However, it may be noted that $\varphi_I(t)$ is not "causal" in the sense that it has values for all time making $x_R(t)$ depend on all samples extending both in the past and future. An approximation may be made however by noting that $\varphi_I(t)$ tends to zero for large t . If this delay is allowable in the system under consideration then a truncated $\varphi_I(t)$ may be used, namely $\varphi_I(t-jT) u(t)$ where jT is large and $u(t)$ is unity for t greater than zero and zero for t less than zero.

A property of the reconstruction function that must be met in any physical system is that it vanish for negative time; namely,

$$\varphi(t) = 0, \quad t < 0; \quad (7.1.1-2)$$

a "causal" function.

It should be noted that the impulse response of a realizable analog filter must satisfy this requirement.

Upon taking the Laplace transform (or Fourier Transform) of the linear reconstruction (7.1.1-1), it is found that the transform of $x_R(t)$, $X_R(s)$ is given by:

$$X_R(s) = \Phi(s) \sum_{k=0}^{\infty} x_k e^{-skT} \quad (7.1.1-3)$$

where

$$\Phi(s) = \int_0^{\infty} \varphi(t) e^{-st} dt \quad (7.1.1-3)$$

= the Laplace transform of $\varphi(t)$,

and the well known transform relation

$$\int_0^{\infty} \varphi(t-\tau) e^{-st} dt = e^{-s\tau} \Phi(s) \quad (7.1.1-4)$$

has been used.

Now the summation

$$X_s(s) = \sum_{k=0}^{\infty} x_k e^{-skT} \quad (7.1.1-5)$$

can be recognized as the impulse sampled spectrum of $x(t)$ whose samples are $x_k = x(kT)$ of Section 2.1.4. It was shown that the impulse sampled spectrum was identical (within a factor of T) to that of $x(t)$ up to half the sampling frequency, if $x(t)$ is bandlimited by half the sampling frequency $\omega_s = 2\pi/T$. The impulse spectrum is periodic outside of the region.

The ratio

$$\Phi(s) = \frac{X_R(s)}{X_s(s)}, \quad (7.1.1-6)$$

then, is the reconstruction transfer function. $\Phi(s)$ should ideally be zero above half the sampling frequency and unity below. However, the time function corresponding to this $\Phi(s)$ is $\varphi_I(t)$ of equation (2.1.1-9) which is unrealizable without significant delay.

7.2 SAMPLE AND HOLD RECONSTRUCTION

A popular method of data reconstruction, polynomial curve fitting, is also commonly known as a "sample and hold". The degree of the polynomial fitted is assigned to the sample and hold, e.g., "first order sample and hold".

A convenient polynomial reconstruction formula that fits exactly at the sampled values is the backward difference Gregory-Newton interpolation formula.

$$\begin{aligned}
 x_R(t) = & x_k + u \Delta x_{k-1} + \frac{u(u+1)}{2} \Delta^2 x_{k-2} \\
 & + \frac{u(u+1)(u+2)}{3} \Delta^3 x_{k-3} + \dots + \frac{u(u+1)\dots(u+n-1)}{n} \Delta^n x_{k-n}
 \end{aligned}
 \tag{7.2-1}$$

where

$$u = \frac{t - kT}{T},$$

$$\Delta x_j = x_{j+1} - x_j,$$

n = order of "fitted" polynomial

and

T = sampling interval.

This formula* is used for $kT \leq t < (k+1)T$ for "extrapolation" holds and with t replaced by $t - T$ for "delayed" holds with the $n+1$ sampled values $x_k, x_{k-1}, \dots, x_{k-n}$.

7.2.1 Zero Order Sample and Hold

The extrapolation case for $n = 0$

$$x_R(t) = x_k, \quad kT \leq t < (k+1)T, \tag{7.2.1-1}$$

is often called a "box car" as it simply holds the value of the most recent sample until the next arrives. The "delayed" hold simply delays the reconstruction by a sampling period T . Figure 7.2.1-1 illustrates its operation

* C. Lanczos, Applied Analysis, Prentice-Hall, 1964

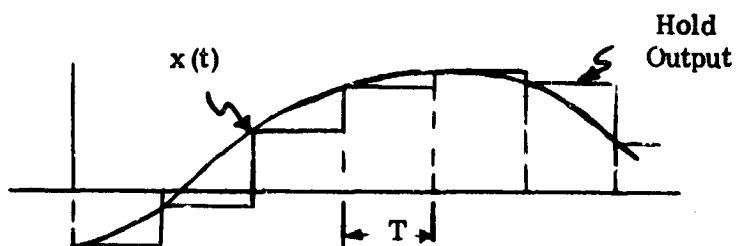


Figure 7.2.1-1 Zero Order Hold Output

To determine the transfer function (7.1.1-6) take the Laplace transform

$$\begin{aligned}
 X_R(s) &= \int_0^{\infty} x_R(t) e^{-st} dt \\
 &= \sum_{k=0}^{\infty} x_k \int_{kT}^{(k+1)T} e^{-st} dt \\
 &= \frac{1 - e^{-sT}}{s} \sum_{k=0}^{\infty} x_k e^{-skT}
 \end{aligned} \tag{7.2.1-2}$$

Now

$$X_s(s) = \sum_{k=0}^{\infty} x_k e^{-skT}$$

is the transform of an impulse sampled function $x(t)$.

Therefore the ratio:

$$\begin{aligned}
 R_o(s) &= \frac{X_R(s)}{X_s(s)} \\
 &= \frac{1 - e^{-sT}}{s} \\
 &= \Phi(s)
 \end{aligned}
 \tag{7.2.1-3}$$

is the Laplace transfer function of the reconstruction function with respect to an impulse sampled input.

The reconstruction function obviously is

$$\varphi(t) = \begin{cases} 1, & 0 \leq t < T, \\ 0, & \text{otherwise} \end{cases}
 \tag{7.2.1-4}$$

by taking the inverse Laplace transform of (7.2.1-3) and is depicted in Figure 7.2.1-2.

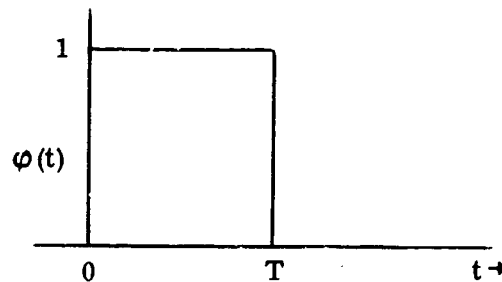


Figure 7.2.1-2 Zero Order Sample and Hold Reconstruction Function

Letting $s = i\omega$ the transfer function as a function of angular frequency is

$$R_o(i\omega) = T e^{\frac{-i\omega T}{2}} \frac{\sin \frac{\omega T}{2}}{\frac{\omega T}{2}}
 \tag{7.2.1-5}$$

The phase obviously is given by

$$\psi_o(\omega) = -\frac{\omega T}{2}; \quad (7.2.1-6)$$

a delay of $T/2$ is introduced as the phase shift is linear in ω .

The impulse sampled sequence is modified in amplitude by

$$|R_o(i\omega)| = \left| T \frac{\sin \frac{\omega T}{2}}{\frac{\omega T}{2}} \right| \quad (7.2.1-7)$$

Figure 7.2.1-3 is a plot of $|R_o(i\omega)|/T$ as a function of $\omega T/2 \pi = \omega/\omega_s$

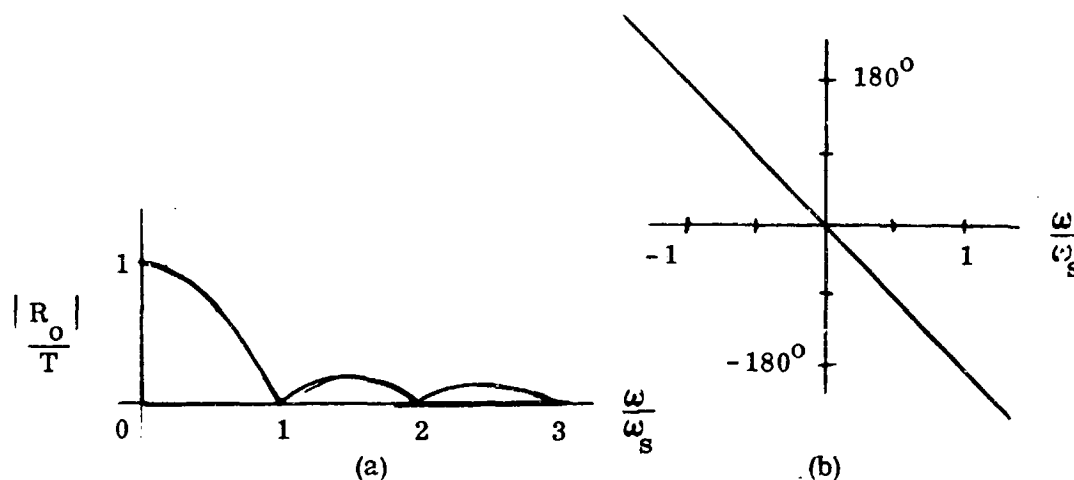


Figure 7.2.1-3 Amplitude (a) and Phase (b) of Zero Order Hold

The important features of $R_o(i\omega)$ are that significant sampling harmonic leakage can occur at $\omega = \omega_s$ unless the signal bandwidth is small or a subsequent system transfer function is small about ω_s and furthermore, significant error is caused by the lack of flatness of $|R_o(i\omega)|$ for ω close to zero.

In analog systems a zero order hold is quite useful as approximate performance is all that is usually desired and, further, a low pass filter can be used to remove sampling harmonic ripple or error. Sampling prior to this removal, as in a digital system, results in aliasing of the "ripple".

Typical values of the response of the sample and hold are tabulated:

$\frac{\omega}{\omega_s}$	Phase	$\frac{1}{f} R_o(i\omega) $
0	0°	.1
1/4	-45°	.901
1/2	-90°	.636
3/4	-135°	.309
1	-180°	0

Table 7.2.1-1 Zero Order Hold Values

It is quite evident from the table that the transfer function is not "flat" near zero frequency. 10% error is incurred at 1/4 the sampling frequency.

7.2.2 First Order Sample and Hold (Delayed)

The "delayed" first order sample and hold is implemented by the linear formula

$$x_R(t) = x_k + \frac{t - (k+1)T}{T} (x_k - x_{k-1}) \quad (7.2.2-1)$$

with

$$kT \leq t < (k+1)T$$

Taking the Laplace transform the transfer function is found to be

$$\begin{aligned}
 R_1^d(s) &= \frac{X_R(s)}{X_g(s)} \\
 &= \frac{(1 - e^{-sT})^2}{Ts^2} \\
 &= \bar{\Phi}(s)
 \end{aligned} \quad (7.2.2-2)$$

assuming $x_k = 0$ for $k < 0$.

This sample and hold method fits straight line segments to the between data points delayed by one sample time as in figure 7.2.2-1.

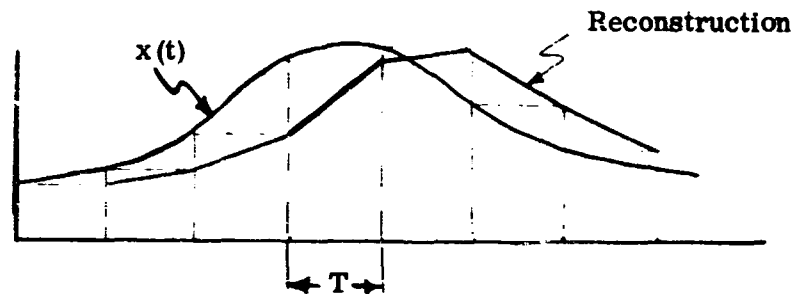


Figure 7.2.2-1 First Order Sampled and Hold Reconstruction (Delayed)

By taking the inverse transform of (7.2.2-2) the reconstruction function is found to be a simple triangle:

$$\phi(t) = \begin{cases} \frac{t}{T}, & 0 \leq t < T, \\ 1 - \frac{t-T}{T}, & T \leq t < 2T, \\ 0, & \text{otherwise} \end{cases} \quad (7.2.2-3)$$

It is depicted in Figure 7.2.2-2.

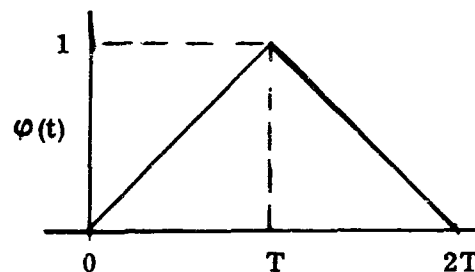


Figure 7.2.2-2 First Order Sample and Hold (Delayed) Reconstruction Function

With $s = i\omega$ the transfer function becomes

$$R_1^d(i\omega) = T e^{-i\omega T} \frac{\sin^2(\frac{\omega T}{2})}{(\frac{\omega T}{2})^2} \quad (7.2.2-4)$$

The delay is seen to be T seconds, twice that of the zero order hold, and the amplitude of $\Phi(i\omega)/T$ is identical to the square of that for the zero order hold. Besides the delay it has one deficiency over the zero order hold and one advantage. The deficiency is that the transfer function falls off faster with frequency in the useful range, zero to half the sampling frequency. Its advantage is the additional attenuation of higher sampling harmonics by the inverse square fall-off with frequency which assists to some extent in removing aliasing when sampling the output of a filter whose input is the reconstructed waveform.

Typical values of the transfer function are tabulated in Table 7.2.2-1.

$\frac{\omega}{\omega_s}$	Phase	$\frac{1}{T} R_1^d(i\omega)$
0	0	1
1/4	-90°	.812
1/2	-180°	.404
3/4	-270°	.09
1	-360°	0

Table 7.2.2-1 First Order Hold (Delayed) Values

7.2.3 First Order Sample and Hold (Extrapolation)

The polynomial fitting formula is (eq (7.2-1) $n = 1$)

$$x_R(t) = x_k + \frac{t - kT}{T} (x_k - x_{k-1}),$$

with

(7.2.3-1)

$$kT \leq t < (k+1)T$$

In pictorial form the effect of this form of extrapolation can be seen in Figure 7.2.3-1.

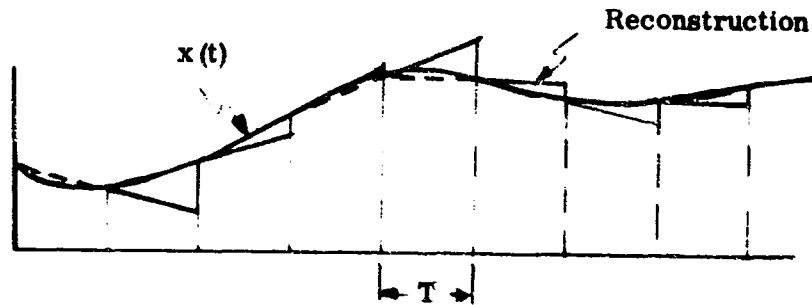


Figure 7.2.3-1 First Order Sample and Hold
(Extrapolated) Output

By taking the Laplace transform of (7.2.3-1) the transfer function is found to be

$$\begin{aligned}
 R_1^e(s) &= \left(\frac{1+Ts}{T} \right) \left(\frac{1-e^{-sT}}{s} \right)^2 \\
 &= \frac{X_R(s)}{X_s(s)}
 \end{aligned}
 \tag{7.2.3-2}$$

after considerable algebra. Letting $s = j\omega$ the frequency and phase responses are found to be

$$|R_1^e(j\omega)| = T [1 + (\omega T)^2]^{1/2} \frac{(\sin \frac{\omega T}{2})^2}{(\frac{\omega T}{2})^2}
 \tag{7.2.3-3}$$

$$\psi_1^e(\omega) = \tan^{-1}(\omega T) - \omega T$$

The amplitude and phase responses when plotted appear in figure 7.2.3-2.

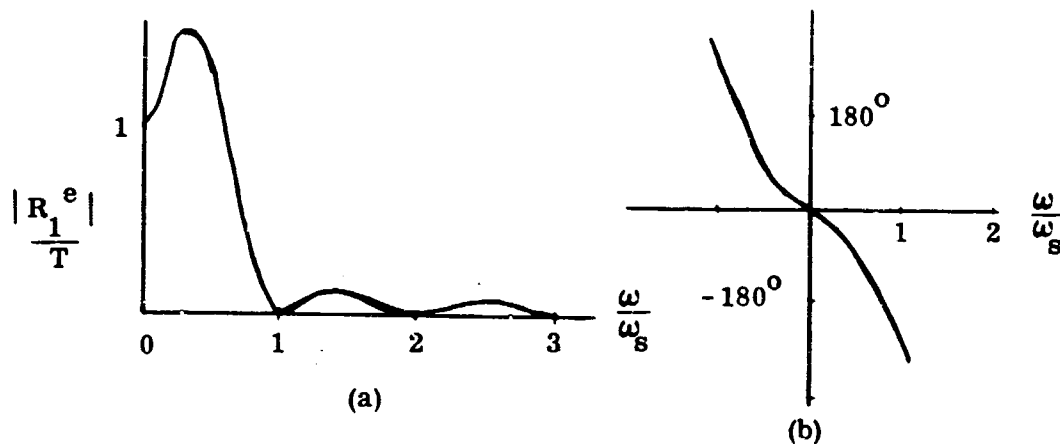


Figure 7.2.3-2 Amplitude (a) and Phase (b) Characteristics of a First Order Hold (Extrapolation)

Typical values of the first order hold transfer function phase and amplitude are tabulated below:

$\frac{\omega}{\omega_s}$	Phase	$\frac{1}{T} R_1(i\omega) $
0	0°	1.0
1/4	-32.5°	1.51
1/2	-107.7°	1.35
3/4	-192.0°	.424
1	-279.0°	0

Table 7.2.3-1 Amplitude and Phase Characteristics of a First Order Hold

Taking the inverse Laplace transform the reconstruction function is

$$\phi(t) = \begin{cases} 1 + \frac{t}{T}, & 0 \leq t < T, \\ -(\frac{t-T}{T}), & T \leq t < 2T, \\ 0, & \text{otherwise} \end{cases} \quad (7.2.3-4)$$

and is illustrated in Figure 7.2.3-3.

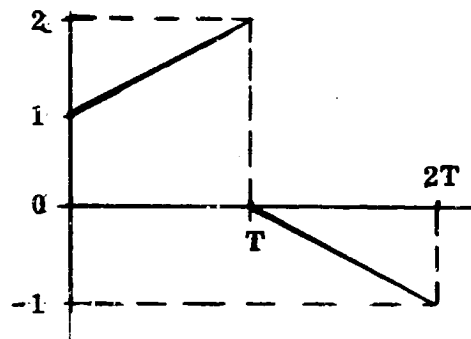


Figure 7.2.3-3 First Order Sample and Hold
(Extrapolation) Reconstruction Function

The first order hold (extrapolated) suffers from difficulties. Its phase shift is non-linear and the transfer function peaks markedly above unity gain at $\omega/\omega_s = .25$. Further the phase shift is more than 1.5 times that of the zero order hold at a given frequency. Finally, the normalized (by $1/T$) amplitude response only falls off as fast as $1/\omega$ for large ω does the zeroth order hold. However in many feedback systems these difficulties are unimportant as the delay is less than that for the delayed first order sample and hold.

7.3 NUMERICAL INTEGRATION BY INPUT RECONSTRUCTION WITH POLYNOMIALS

This method of numerical integration has been used extensively and successfully by electronics engineers for many years.* In many applications a time function is sampled, converted to digital form, transmitted to some point, and then reconstructed for analog system processing. Analysis of this sampling, reconstruction, and subsequent processing by a linear system leads to a difference equation if input and output at the input sampling instants is of interest. This difference equation can be mechanized on a digital computer and, therefore, a numerical integration scheme is obtained. General reconstruction and the transfer function of polynomial fitting for low orders was discussed thoroughly in Section 7.2.

* Julius T. Tou, Digital and Sampled - Data Control Systems, McGraw-Hill 1959

A major drawback of these systems is the introduction of an error in the reconstructed signal which also appears in the output. This error is due to the sampling harmonic leakage, especially at the sampling frequency. Polynomial reconstruction does not allow as much rejection of sampling harmonics as the lowpass filter method (Section 7.4) but requires no extra filter coefficients in the digital filter.

Sometimes it is stated in the literature that polynomial reconstruction allows exact solutions at the sampling instants. This statement is misleading. While it is true that a numerical integration procedure can be developed to describe the input-output relation between the reconstructed signal input and the linear system output with no delay, the desired relationship is a numerical integration scheme between the continuous input and the linear system output. Sampled signal reconstruction is, however, a fundamental part of the numerical integration. Polynomial reconstruction does not allow exact numerical integration of the total continuous input/linear system problem and, in fact, produces significant output error for low order polynomial reconstruction at higher frequencies though time delay is small. As the order of reconstruction, the degree of the polynomial, increases the time delay must increase.

Intuitively it is almost obvious that delay must increase as reconstruction order increases. More data points are needed to define a polynomial as the order increases: $n + 1$ data points are needed for an n^{th} order polynomial.

7.3.1 Conversion of Hold Systems to Digital Recursive Filters

The Laplace transfer function of the data hold was defined as the ratio of the continuous reconstructed output transform to the impulse sampled input transform and was given as (See Section 7.2)

$$R_o(s) = \frac{1 - e^{-sT}}{s} \quad (7.2.1-3)$$

for the zero order hold,

$$R_1^e(s) = \frac{s + \frac{1}{T}}{s^2} (1 - e^{-sT})^2 \quad (7.2.3-2)$$

for the first order (extrapolated) hold, and

$$R_1^d(s) = \frac{(1 - e^{-sT})^2}{TS^2} \quad (7.2.2-2)$$

for the first order (delayed) hold.

Combined with a low-pass transfer function $H(s)$, the overall transfer function is:

$$R(s) H(s)$$

where $R(s)$ corresponds to the hold involved.

To compute the time response the Laplace transform, $Y(s)$, the output of $H(s)$ is used:

$$Y(s) = H(s) R(s) X_s(s) \quad (7.3.1-1)$$

By the convolution theorem, (Section 2.7.1), $y(t)$, the inverse of $Y(s)$, is given by:

$$y(t) = \int_0^{\infty} \bar{w}(t - \tau) x(\tau) \sum_{j=0}^{\infty} \delta(\tau - jT) d\tau \quad (7.3.1-2)$$

where

$$x(\tau) \sum_{j=0}^{\infty} \delta(\tau - jT)$$

is the inverse transform of $X_s(s)$ (See Section 2.1.4)

and

$\bar{w}(\tau)$ is the "impulse response" of $H(s) R(s)$.

Carrying out the integration:

$$y(t) = \sum_{j=0}^{\infty} \bar{w}(t - jT) x(jT) \quad (7.3.1-3)$$

At the sampling instants $t = nT$

$$y(nT) = \sum_{j=0}^{\infty} \bar{w}((n-j)T) x(jT) \quad (7.3.1-4)$$

Taking the z transform and adjusting indices it is found that

$$Y(z) = H_z(z) X(z),$$

where

$$Y(z) = \sum_{k=0}^{\infty} y_k z^{-k},$$

$$X(z) = \sum_{k=0}^{\infty} x_k z^{-k},$$
(7.3.1-5)

and

$$H_z(z) = \sum_{k=0}^{\infty} \bar{w}(kT) z^{-k}$$

If $H_z(z)$ can be represented as the ratio of two polynomials in z^{-1} , then (7.3.1-5) is equivalent to a difference equation. Fortunately, if $H(s)$ is derived from a linear constant coefficient differential equation then $H_z(z)$ is the ratio of two polynomials in z^{-1} . The method of converting $H_z(z)$ to a ratio of two polynomials in z^{-1} is common to this method of input reconstruction and that of the low-pass filter reconstruction method. It was described separately in Section 6.5.

It should be noted that the particular forms of the zero and first order holds allow simplification of the z transform procedure upon the impulse response of the transfer functions: As e^{-sT} in the Laplace transform domain corresponds to a unit, T , of time delay it can be seen that the z transform of the impulse responses of the reconstruction systems are:

1) Zero Order Hold

$$H_z(z) = (1 - z^{-1}) \mathcal{Z}\left(\frac{H(s)}{s}\right)$$
(7.3.1-6)

where $\mathcal{Z}\left(\frac{H(s)}{s}\right)$ means the z transform of the impulse response of the transfer function $H(s)/s$.

2) First Order Hold (extrapolated)

$$H_z(z) = (1 - z^{-1})^2 \mathcal{Z} \left(\frac{(s + \frac{1}{T})}{s^2} H(s) \right) \quad (7.3.1-7)$$

where $\mathcal{Z} \left(\frac{(s + \frac{1}{T})}{s^2} H(s) \right)$ is interpreted identically as for the zero order hold.

3) First Order Hold (delayed)

$$H_z(z) = (1 - z^{-1})^2 \mathcal{Z} \left(\frac{H(s)}{s^2} \right) \quad (7.3.1-8)$$

It should be noted that care must be taken with these formulae. When, for example, the z transform is taken of the impulse response of $H(s)/s$,

$$\mathcal{Z} \left(\frac{H(s)}{s} \right),$$

a term $(1 - z^{-1})$ is introduced in the denominator of the sampled data transfer function $H_z(z)$ due to the factor $1/s$, (See Section 6.2). For efficient digital filters with a small number of coefficients and for stability (a pole on the unit z circle) this term $(1 - z^{-1})$ in the denominator should be cancelled by the $(1 - z^{-1})$ introduced in the numerator by (7.3.1-6).

For the first order hold systems a similar kind of term $(1 - z^{-1})^2$ appears due to the $1/s^2$ term. This should be taken care of in the same manner as for a zero order hold.

7.4 BANDPASS FILTER RECONSTRUCTION FUNCTIONS

From previous discussions it should be somewhat apparent that bandpass analog filters possess desired properties. They are, for filters of the Butterworth, Tchebychev, elliptic, etc. type; a "relatively" flat amplitude characteristic, controllable attenuation for frequencies outside the "pass-band", and an approximately linear phase shift in the "pass-band". Therefore, these kinds of filters should be considered as reconstruction networks as they can be used to remove sampling harmonics.

An apparent difficulty, in analog reconstruction, might seem to be the need for impulse modulated inputs. However, instead of impulses, triangles or pulses of a width narrow with respect to the sampling period may be used.

The reconstruction function is then the impulse response of the reconstruction filter transfer function (though the filter transfer function is of more practical importance):

$$\varphi(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} R(i\omega) e^{i\omega t} d\omega \quad (7.4-1)$$

As the design of recursive filters from an analog prototype can be considered from the point of view of an impulse sampled input a pertinent low-pass reconstruction filter will be considered.

7.4.1 Design of a Reconstruction Low-Pass Filter

A particular, well known, low-pass filter which is realizable and has desirable characteristics is the Butterworth filter.* It is also known as a "maximally flat" filter because of its extreme flatness at $\omega \leq 0$ which is related to the number of derivatives which are zero at $\omega = 0$.

The Butterworth filter was chosen for use here for a number of reasons:

- a) Approximately linear phase through the "pass-band", giving a constant delay to a signal localized to the "pass-band".
- b) Approximately constant amplitude through the "pass-band" produces little amplitude modification of the system transfer function of interest.
- c) Synthesis for numerical integration purposes is virtually trivial.
- d) Only a filter with a few poles (2 to 6) is usually needed.

The low-pass Butterworth filter is specified by the magnitude function

$$|R(i\omega)| = \frac{1}{\sqrt{1 + \left(\frac{\omega}{\omega_0}\right)^{2n}}} \quad (7.4.1-1)$$

where n is the number of filter "poles" in the " s " plane, and ω_0 is the cut-off frequency (the filter response is down 3 decibels or a factor of $1/\sqrt{2}$ from that at $\omega = 0$). As the magnitude is completely specified so must the phase. Fortunately, it is found that the phase is approximately linear through most of the "pass-band", zero to ω_0 , of the Butterworth filter producing delay without distortion.

The stable transfer function corresponding to the magnitude function above is:

$$R(s) = \frac{\omega_0^n}{(s + s_1)(s + s_2) \dots (s + s_n)} \quad (7.4.1-3)$$

* M. E. Van Valkenburg, Introduction to Modern Network Synthesis, Wiley

where the negative roots s_j are given by

$$s_j = \omega_0 e^{\frac{j(n+2k-1)}{n} \frac{\pi}{2}}, \quad K = 1, 2, \dots, n \quad (7.4.1-3)$$

Each of the poles is a first order pole and exists, as it must, along with its complex conjugate pole for n even. When n is odd the poles are also first order and exist in conjugate pairs with the exception of a single real pole at $s = -\omega_0$.

The phase shift per pole of the filter is -45° , or $-\pi/4$ radians to the cut-off frequency. However the approximately linear portion does not extend to ω_0 but the extent approaches ω_0 as n increases. The linear portion of the phase is approximately given by

$$\psi_n(\omega) = \frac{\omega}{\omega_0} \frac{2}{\pi} n \quad (7.4.1-4)$$

where

ω_0 = cut-off frequency in radians per second,

n = number of poles or the "order" of the filter.

The Butterworth filter has two parameters which may be adjusted for data reconstruction, the cut-off frequency, ω_0 and the "order", n . It will be seen that these parameters need not be adjusted independently, but in a manner that minimizes the time delay for a given attenuation of the sampling harmonics (or of a transfer function that is being bandlimited).

To determine the filter parameters consider figure 7.4.1-1.

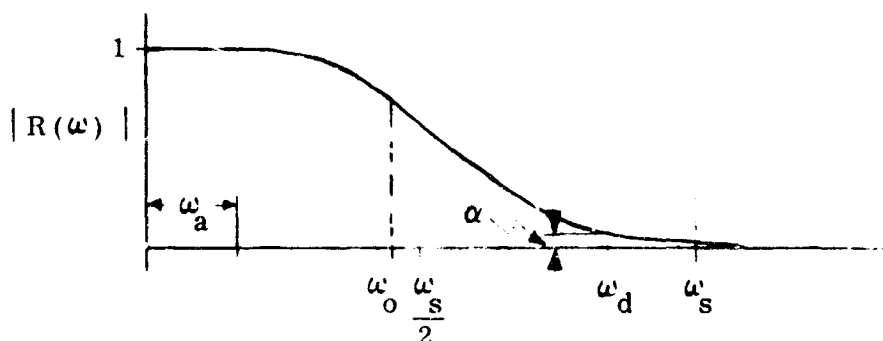


Figure 7.4.1-1 Reconstruction Filter

where:

ω_a = (a) bandlimited of sampled analog input or (b) desired region of approximation of analog transfer function (radians/sec)

ω_s = sampling frequency (radians/sec)

ω_0 = Butterworth filter cut-off frequency (radians/sec)

α = (a) least attenuation of first impulse sampling harmonic with respect to unity ($\alpha = .01$ is a reduction of the harmonic by a factor of 100) or
(b) the least reduction of the first image harmonic of the desired transfer function in the region $0 \leq |\omega| \leq \omega_a$ (Section 2.1.2).

$$\omega_d = \omega_s - \omega_a$$

Setting the Butterworth filter attenuation at the edge, ω_d , of the band of the first sampling harmonic to a desired value,

$$\frac{1}{\sqrt{1 + \left(\frac{\omega_d}{\omega_0}\right)^{2n}}} = \alpha$$

or

$$\frac{\omega_0}{\omega_d} = \frac{1}{(\beta^2 - 1)^{1/(2n)}}, \quad \beta^2 = \frac{1}{\alpha^2} \quad (7.4.1-5)$$

Replacing ω_d by $\omega_s - \omega_a$ and letting ω_a be represented as a fraction of half the sampling frequency,

$$\omega_a = \mu \frac{\omega_s}{2}, \quad 0 \leq \mu \leq 1, \quad (7.4.1-6)$$

then (7.4.1-5) can be written as

$$\begin{aligned} \frac{\omega_0}{\omega_s} &= \frac{(1 - \frac{\mu}{2})}{(\beta^2 - 1)^{1/(2n)}} \\ &= \gamma/2 \end{aligned} \quad (7.4.1-7)$$

γ is the cut-off frequency normalized to half of the sampling frequency. Now the cut-off frequency ω_0 must be greater than ω_a if distortion of the input signal (or the transfer function being approximated) is to be minimized. Therefore,

$$\gamma = \frac{\omega_0}{\omega_s/2}$$

(7.4.1-8)

$$\geq \mu: 0 \leq \gamma \leq 2 - \mu$$

On the other hand the time delay of the Butterworth filter is approximately given by

$$\tau_d = \frac{\psi_n(\omega)}{\omega}$$

(7.4.1-9)

$$= \frac{2n}{\pi \omega_0}$$

in the filter pass band from equation (7.4.1-4). Also τ_d can be written as

$$\frac{\tau_d}{T} = \frac{n(\beta^2 - 1)^{1/(2n)}}{\pi^2 \left(1 - \frac{\mu}{2}\right)}$$

(7.4.1-10)

from equation (7.4.1-7) and as $\omega_s = 2\pi/T$ where T is the time between input data samples. Examination of this equation shows that it expresses the time delay, normalized to the sampling interval, in terms of the Butterworth filter order, n , the desired normalized approximation frequency, μ , and the desired approximation accuracy, $\alpha = 1/\beta$. For a fixed μ and β it can be shown that τ_d/T has a minimum as a function of filter order n . n can be adjusted to minimize the time delay.

This adjustment of filter order is subject to the constraint (7.4.1-8) above. Carrying out these computations it is found that meeting the constraint, $\gamma \geq \mu$, often requires a filter order that is greater than that which minimizes (7.4.1-10) independently (a minimum pole filter).

Recapitulating, the procedure for determining the Butterworth filter parameters ω_0 and n is the following

- a) Specify the attenuation of the first sampling harmonic (the approximation accuracy) $\alpha = 1/\beta$ as a fraction of unity.

- b) Specify the fraction, μ , of half the sampling frequency to which the approximation is desired.
- c) Compute the normalized Butterworth cut-off frequency such that

$$\gamma = \frac{2 \left(1 - \frac{\mu}{2}\right)}{(\beta^2 - 1)^{1/(2n)}}$$

is greater than μ by increasing n from 1 to some value n_1 . n_1 is the first value of n that satisfies the inequality

$$\gamma \geq \mu.$$

- d) Insert $n = n_1$ into the formula

$$\frac{\tau_d}{T} = \frac{(\beta^2 - 1)^{1/(2n)} n}{\pi^2 \left(1 - \frac{\mu}{2}\right)}$$

and see if this value of (τ_d/T) is less than that obtained by using $n_1 + 1$. If not continue increasing n until the minimum value of τ_d/T is found. This is the minimum time delay. If the delay increased by using $n_1 + 1$ then the value n_1 is the minimum that satisfies all the criteria.

- e) Using the value of n obtained from (d) compute the final value of ω_0 :

$$\begin{aligned} \omega_0 &= \frac{\omega_s}{2} \cdot \gamma \\ &= \omega_s \frac{1 - \frac{\mu}{2}}{(\beta^2 - 1)^{1/(2n)}} \end{aligned}$$

The Butterworth low-pass is completely defined from steps (d) and (e) above plus 7.4.1-2 for the transfer function in terms of n and ω_0 .

Table 7.4.1-1 below has been computed using the procedure above for various useful approximation accuracies, and approximation regions, μ , of the analog transfer function.

As the minimum number of poles to satisfy the accuracy requirement (α) is also of interest, for numerical accuracy and for the least number of coefficients in the recursive digital filter ultimately obtained, it has been included as a separate item in the table.

For convenience the various parameters used in table 7.4.1-1 are summarized.

α = approximation accuracy as:

- (a) minimum attenuation as a fraction of unity of the first impulse sampling harmonic or
- (b) minimum reduction factor of the first image-term of the desired analogue transfer function,

$\mu = \omega_a/(\omega_s/2)$, the normalized transfer function approximation region,

ω_a = un-normalized transfer function approximation region ($0 \leq \omega \leq \omega_a$; $\omega_a < \omega_s/2$) in radians per second,

$\omega_s = 2\pi/T$, data sampling frequency in radians per second.

T = data sampling interval in seconds,

$\gamma = \omega_0/(\omega_s/2)$, the normalized Butterworth filter cut-off frequency,

ω_0 = un-normalized cut-off frequency in radians per second

τ_d = Butterworth filter time delay,

n = Butterworth filter order.

TABLE 7.4.1-1 Minimum Delay and Minimum Pole Reconstruction Butterworth Filter Design Parameters

μ	n	γ	τ_d/T
1/16	2	.614	.814
1/8	2	.594	.841
1/4	2	.555	.901
1/2	3	.697	1.075
3/4	5	.789	1.583

$\alpha = .1$
Minimum Delay
(a)

μ	n	γ	τ_d/T
1/16	1	.195	1.284
1/8	1	.188	1.327
1/4	2	.555	.901
1/2	3	.697	1.075
3/4	5	.789	1.583

$\alpha = .1$
Minimum Number of Poles
(b)

μ	n	γ	τ_d/T
1/16	5	.771	1.621
1/8	5	.746	1.675
1/4	5	.697	1.794
1/2	5	.597	2.093
3/4	10	.789	3.170

$\alpha = .01$
Minimum Delay
(c)

μ	n	γ	τ_d/T
1/16	2	.194	2.581
1/8	2	.188	2.667
1/4	3	.377	1.989
1/2	5	.597	2.093
3/4	10	.789	3.170

$\alpha = .01$
Minimum Number of Poles
(d)

μ	n	γ	τ_d/T
1/16	7	.722	2.423
1/8	7	.699	2.504
1/4	7	.652	2.683
1/2	7	.559	3.130
3/4	14	.763	4.586

$\alpha = .001$
Minimum Delay
(e)

μ	n	γ	τ_d/T
1/16	3	.194	3.871
1/8	3	.188	4.000
1/4	4	.311	3.213
1/2	7	.559	3.130
3/4	14	.763	4.586

$\alpha = .001$
Minimum Number of Poles
(f)

7.4.2 Using an Analog Low-Pass Filter for Recursive Digital Filter Synthesis

This method of approximating the analog transfer function is quite simple in principle. In Section 4.6, it was seen that an analog transfer function bandlimited at half the sampling frequency had the same frequency response as the digital integration formula. Obviously, it is possible to approximately bandlimit any transfer function by preceding it with a low-pass filter which is essentially zero above one-half the sampling frequency; a price of additional delay must be paid.

Once the transfer function is bandlimited, then the common z transform method may be used to obtain a recursive digital filter. In practice, it is found that the number of poles of the low-pass filter needed to reduce the original analogue transfer function to a bandlimited one is quite large if the low-pass filter is "flat" or constant up to approximately half the sampling frequency.

These extra poles tend to produce three undesirable effects.

1. Additional terms (and poles) in the resultant digital filter
2. Numerical accuracy necessary in the z transform conversion to a digital filter becomes greater
3. More time delay, as mentioned before.

However, these difficulties can be circumvented to a large extent in a practical situation if certain basic facts about sampled data (or sampled impulse responses) are recalled, (Section 2.1).

This method can be made to work up to about 1/4 of the sampling frequency for reasonable accuracies for most transfer functions. If the input function $x(t)$ is considered as being impulse sampled, as in Sec(2.1.4), then the spectrum of the modulated impulse train is the periodic extension in ω of the spectrum of $x(t)$ with a period of the sampling frequency ω_s , assuming $x(t)$ is bandlimited below half the sampling frequency. Figure

7.4.2-1 illustrates this situation. Impulse sampling produces a spectrum

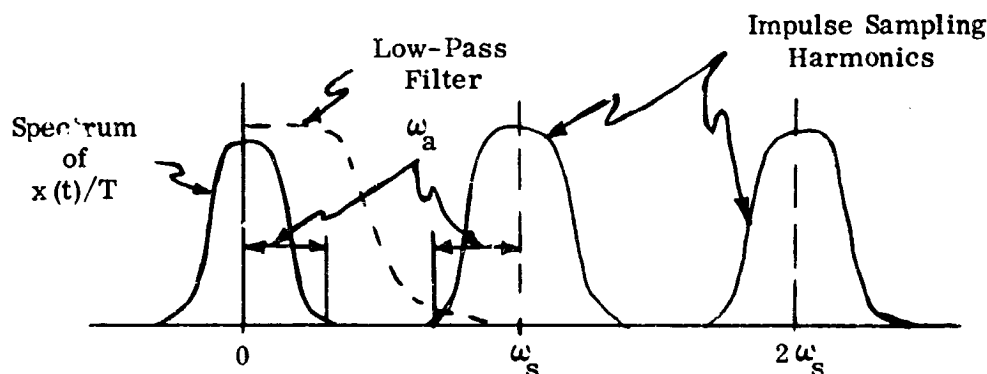


Figure 7.4.2-1 Impulse Sampling Spectrum

with harmonics at multiples of the sampling frequency and amplitude scaled by a factor of $1/T$.

Although the number sequence, $x(kT)$, in a computer is not obtained by impulse sampling, the resultant numerical integration formula that will be obtained is not dependent on impulse sampling actually being performed. Impulse sampling is performed in concept only.

To reconstruct $x(t)$ from the impulse samples, it is "only" necessary to low-pass filter the impulse samples with a filter that removes the higher harmonics.

From Figure 7.4.2-1 it can be seen that if the spectrum of the continuous input extends to ω_a which is less than half the sampling frequency then the first impulse sampling harmonic becomes appreciable at $\omega_d = \omega_s - \omega_a$. A low-pass filter acting upon the impulse samples, which has sufficient attenuation above ω_d , will serve to reconstruct $x(t)$ from the samples. Ideally, the low-pass filter would have perfect transmission (except for a scale factor of T) for $|\omega| \leq \omega_a$ and infinite attenuation when $|\omega| \geq \omega_d$ or

$$R(\omega) = \begin{cases} 1, & |\omega| \leq \omega_a \\ 0, & |\omega| \geq \omega_d = \omega_s - \omega_a \\ \text{Arbitrary,} & \omega_a < |\omega| < \omega_d \end{cases} \quad (7.4.2-1)$$

If the input signal, $x(t)$, has a spectrum such that its upper limit ω_a extends to half the sampling frequency then $R(\omega)$ would be unity in magnitude up to $\omega_s/2$ and zero above.

Unfortunately a filter of this type is impossible to realize without infinite delay. However, as the signal bandwidth, ω_a , is reduced below $\omega_s/2$ then it becomes easier to realize a filter that approximates the characteristics of (7.4.2-1).

Another way of looking at the problem of obtaining the appropriate numerical integration formula is that it is not necessary that the transfer function $H(s)$ (or $H_1(s)$), be bandlimited at $\omega_s/2$ if an approximation only to $\omega_a < \omega_s/2$ is desired. The low pass filter mentioned previously can be combined (multiplied) with the desired system transfer function $H(s)$, such that the product transfer function

$$R(\omega) H(\omega)$$

is approximately zero for $\omega > \omega_d = \omega_s - \omega_a$. In this case, the first image tail of the digital integration transfer function (eq 6.1-10) will be insignificant at $0 \leq |\omega| \leq \omega_a$. Of course, the digital transfer function will not be the same at frequencies higher than ω_a so that the data must also be constrained to contain frequencies less than ω_a . Obviously, a time delay will be introduced by the filter $R(\omega)$ used to attenuate the image tail of $H(\omega)$.

In previous work, discussion of the use of a particular filter for the use of data reconstruction, or equivalently, the nullifying of the transfer function image tail in terms of necessary filter poles, delay, and attenuation specification was made. A "Butterworth" filter was used. Other filters such as Tchebycheff and elliptical designs may be used but little is gained at the expense of significant complexity.

In Section 6.1 the digital input-output relation was derived through appropriate z transformation for a network or transfer function whose input was a sampled function.

It is only necessary with this method to design the reconstruction filter, $R(s)$ of Section 7.4.1 to meet specifications and take the appropriate z transform of the product,

$$R(s) H(s)$$

to form the digital integration procedure $\bar{H}_z(z)$ of Section 6.1.

A warning is in order. The z transform of the product filter must be taken, not the z transform of the individual filters.

7.4.3 Example of Digital Filter Design Using a Butterworth Reconstruction Filter

Assume that a numerical integration formula or simulation of an analogue transfer function, $H(s)$, is desired from 0 to 10 Hz with a sampling frequency of 40 Hz. A delay of one sampling interval is considered tolerable (1/40 of a second). Further, an attenuation of 100 ($\alpha = .01$) is also desired for the transfer function image tail or the first impulse sampling harmonic.

From Table 7.4.1-1c ($\alpha = .01$) it can be seen that a delay of 1.621 sampling intervals only allows an approximation to $1/32 \times 40$ Hz. To obtain an approximation to 10 Hz it is necessary to use $\mu = (10 \text{ Hz}/20 \text{ Hz}) = 1/2$ which implies

a) a delay,

$$\begin{aligned} \tau_d &= 1.794 \text{ sampling intervals} \\ &= 1.794 \times 1/40 \text{ seconds} \end{aligned}$$

b) a Butterworth filter order,

$$n = 5$$

c) a Butterworth filter cut-off,

$$\frac{\omega_0}{\omega_s/2} = \gamma$$
$$= .697$$

$$\omega_0 = 2 \pi \times \frac{40}{2} \times .697 ,$$
$$= 2 \pi \times 13.94 \text{ radians/sec},$$

also

$$f_0 = 13.94 \text{ Hertz}$$

The Butterworth transfer function is, then,

$$R(s) = \frac{\omega_0^5}{(s + s_1)(s + s_1^*)(s + s_2)(s + s_2^*)(s + \omega_0)} , \quad (7.4.3-1)$$

with

$$\omega_0 = 2 \pi \times 13.94$$

$$s_1 = \omega_0 \left(\cos \left(\frac{6}{5} \frac{\pi}{2} \right) + i \sin \left(\frac{6}{5} \frac{\pi}{2} \right) \right) ,$$

$$s_1^* = \omega_0 \left(\cos \left(\frac{6}{5} \frac{\pi}{2} \right) - i \sin \left(\frac{6}{5} \frac{\pi}{2} \right) \right) ,$$

$$s_2 = \omega_0 \left(\cos \left(\frac{8}{5} \frac{\pi}{2} \right) + i \sin \left(\frac{8}{5} \frac{\pi}{2} \right) \right) ,$$

$$s_2^* = \omega_0 \left(\cos \left(\frac{8}{5} \frac{\pi}{2} \right) - i \sin \left(\frac{8}{5} \frac{\pi}{2} \right) \right) .$$

The desired transfer function for numerical integration is $R(s) H(s)$.

To obtain the numerical integration formula (or digital filter) the appropriate z transform Section 6.1 of $R(s)H(s)$ must be taken.

Of course, the desired delay of 1 sampling unit (1/40 sec) cannot be achieved if the attenuation of 100 ($\alpha = .91$) must be attained in the transfer function image tail. If this requirement can be relaxed to a factor of 10 ($\alpha = .1$) then Table 7.4.1-1 indicates that a delay of 1.075, which closely approximates 1, can be obtained. Other "trade-offs" among parameters can be easily found from examination of Table 7.4.1.

7.5 ILLUSTRATIONS OF DESIGN TECHNIQUES

Many techniques have been described in the preceding pages. It is quite interesting to see the effects of the various techniques in actual designs. Ten applications of various techniques are illustrated with comparison to the original analog transfer function.

Figures 7.5-1 and 7.5-2 demonstrate that an essentially bandlimited transfer function may be successfully z transformed by the method of Section 6.1 with zero frequency level correction for the slight aliasing that does occur (Section 6.9).

Practically the entire repertoire of techniques of Sections 6 and 7 were used on a servo-control transfer function

$$H(s) = \frac{(1 + b_1 s + b_2 s^2 + b_3 s^3)}{.0012174 (s + a_1) (s + a_2) (s + a_3) (s + a_4)} \quad (7.5-1)$$

where

$$b_1 = 10.0644$$

$$b_2 = .648635$$

$$b_3 = .058991$$

and

$$a_1 = .04718$$

$$a_2 = 11.13939$$

$$a_3 = 30.05429$$

$$a_4 = 52.00895$$

This transfer function is of the low pass type with a numerator and denominator of degrees 3 and 4, respectively. It therefore requires the impulse modification of Section 6.1. The data sampling rate (10 Hertz) used is such that the transfer function is quite significant

above half the sampling frequency.

Figure 7.5-3 exemplifies the heavy aliasing that can be obtained when direct z transformation is used blindly. The procedure is virtually useless.

The situation is dramatically improved using the delayed first order sample and hold, Figure 7.5-4. Ensuing digital phase plots have been adjusted to remove the linear phase gradient caused by the reconstruction techniques to allow phase comparison with the analog original. Section 7.2.2 discusses the first order sample and hold reconstruction method. As might be expected errors increase after 2.5 cycles though not as large as might be expected as the combination of aliasing and high frequency rejection by the hold actually help in this particular case.

The zero order sample and hold (Section 7.2.1) does not offer the improvement of the first order delayed hold as the aliasing due to high frequencies is not removed as well by the slow high frequency roll-off of this reconstruction method.

Figure 7.5-6 illustrates the distortion caused by the high peaking in the extrapolation sample and hold reconstruction transfer function above $1/4$ of the sampling frequency.

Butterworth filter reconstruction techniques with various aliasing levels (Section 7.4) are illustrated in Figures 7.5-7 and 7.5-8. Both figures illustrate a reconstruction technique designed with an ω_a of $1/4$ of the sampling frequency. As may be seen this design parameter should usually be slightly higher than the limiting design frequency actually desired due to the slight roll-off of the filter within the pass-band. Namely, if reconstruction to $1/4$ the sampling frequency is desired an ω_a of $5/8$ would not be inappropriate. The effect of varying the aliasing level from .1 to .01 is plainly evident in the figures. The greater degree of attenuation above $1/4$ of the sampling frequency (2.5 Hertz) for an aliasing level of $\alpha = .01$ is caused by the faster roll-off of the reconstruction filter in this case.

In all the Figures 7.5-3 through 7.5-8, it should be clear that operating above $1/4$ the sampling frequency is difficult with a recursive filter when the original transfer function is not bandlimited. The Butterworth reconstruction method may be extended to slightly higher frequencies at the expense of more poles and filter weights.

Another interesting example is illustrated in Figures 7.5-9 and 7.5-10. The previous control function (7.5-1) was multiplied by s to create a digital filter whose output would be the time derivative of the output of $H(s)$; namely

$$F(s) = sH(s)$$

As $s = i\omega$ increases rapidly with frequency $F(s)$ is more liable to aliasing than $H(s)$.

A combined technique was used in Figure 7.5-9: first a Butterworth filter with an $\alpha = .01$ was used to reduce the high frequency content by a significant amount and then a first order delayed sample and hold was used on the composite. The combination of techniques sometimes allows the final product digital filter to have less poles and therefore less weights. If the Butterworth filter had not also been used the first order sample and hold would not have reduced the aliasing sufficiently. An adequate differentiating control filter is obtained to about 1.5 Hertz.

Figure 7.5-10 illustrates the disastrous effect of using only a simple zero order sample and hold on $F(s)$.

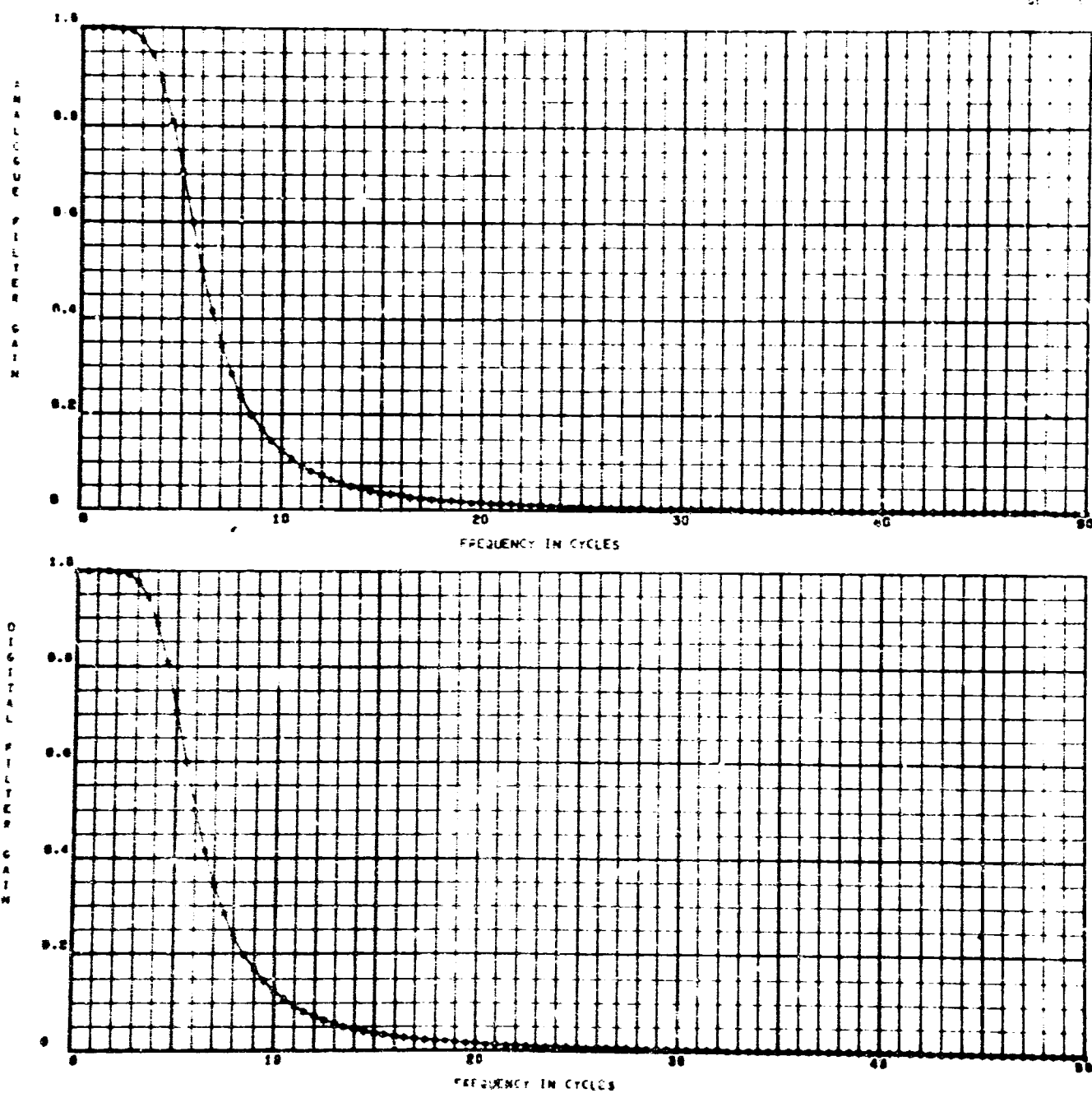


Figure 7.5-1. $N = 3$ Butterworth Low Pass Design
via z Transformation with D.C. Level Set. No Reconstruction

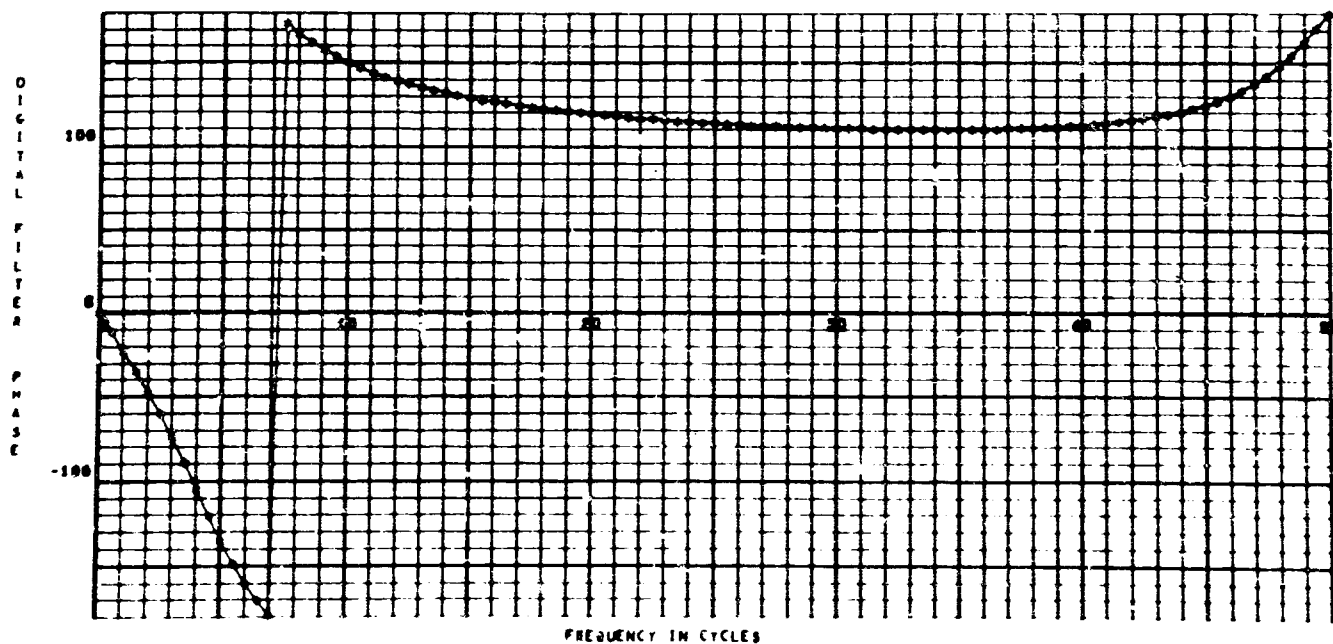
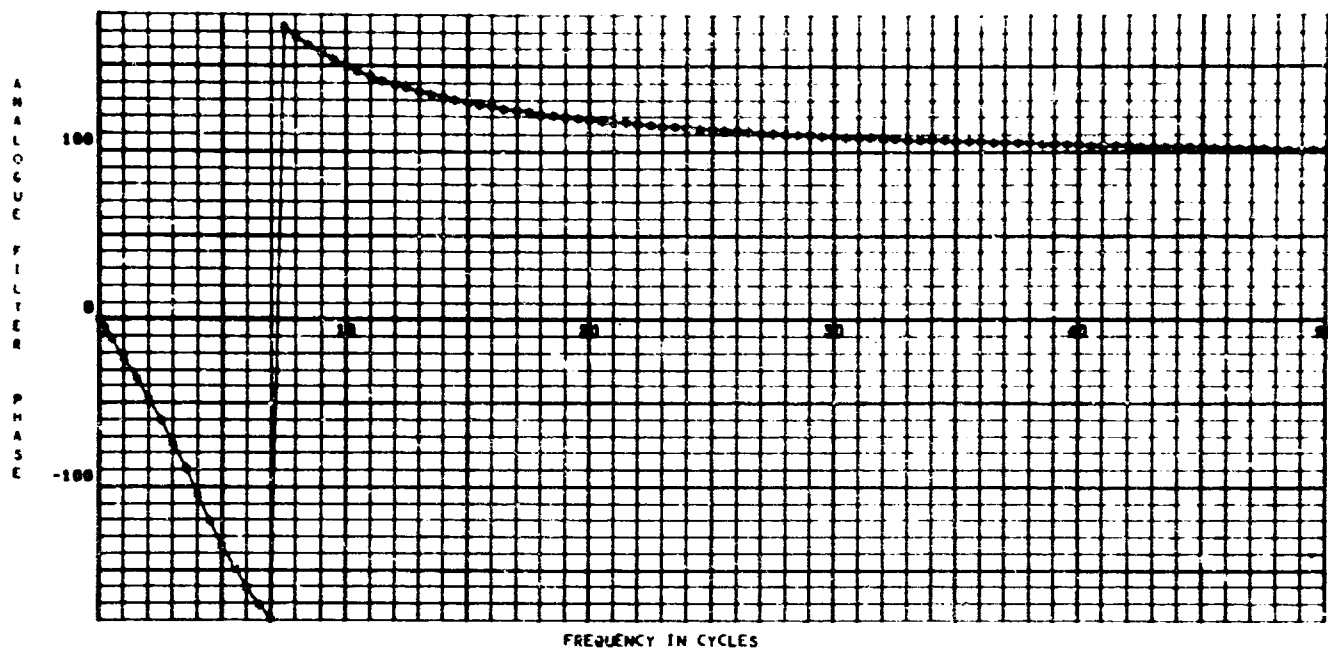


Figure 7.5-1. Continued

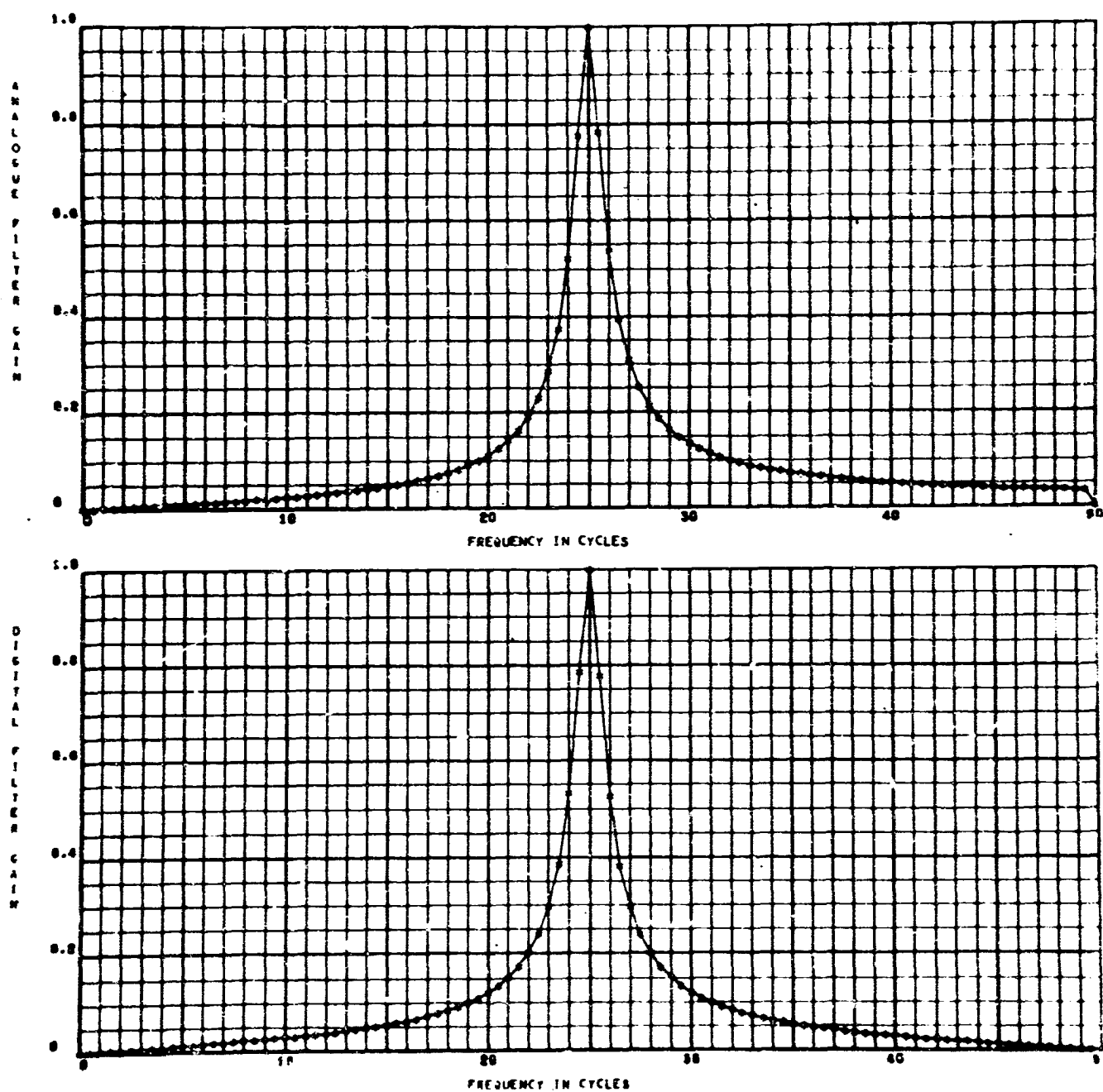


Figure 7-3-2. Bandpass Filter Design
 A/D Transformation with D.C. Level Set, No Reconstruction

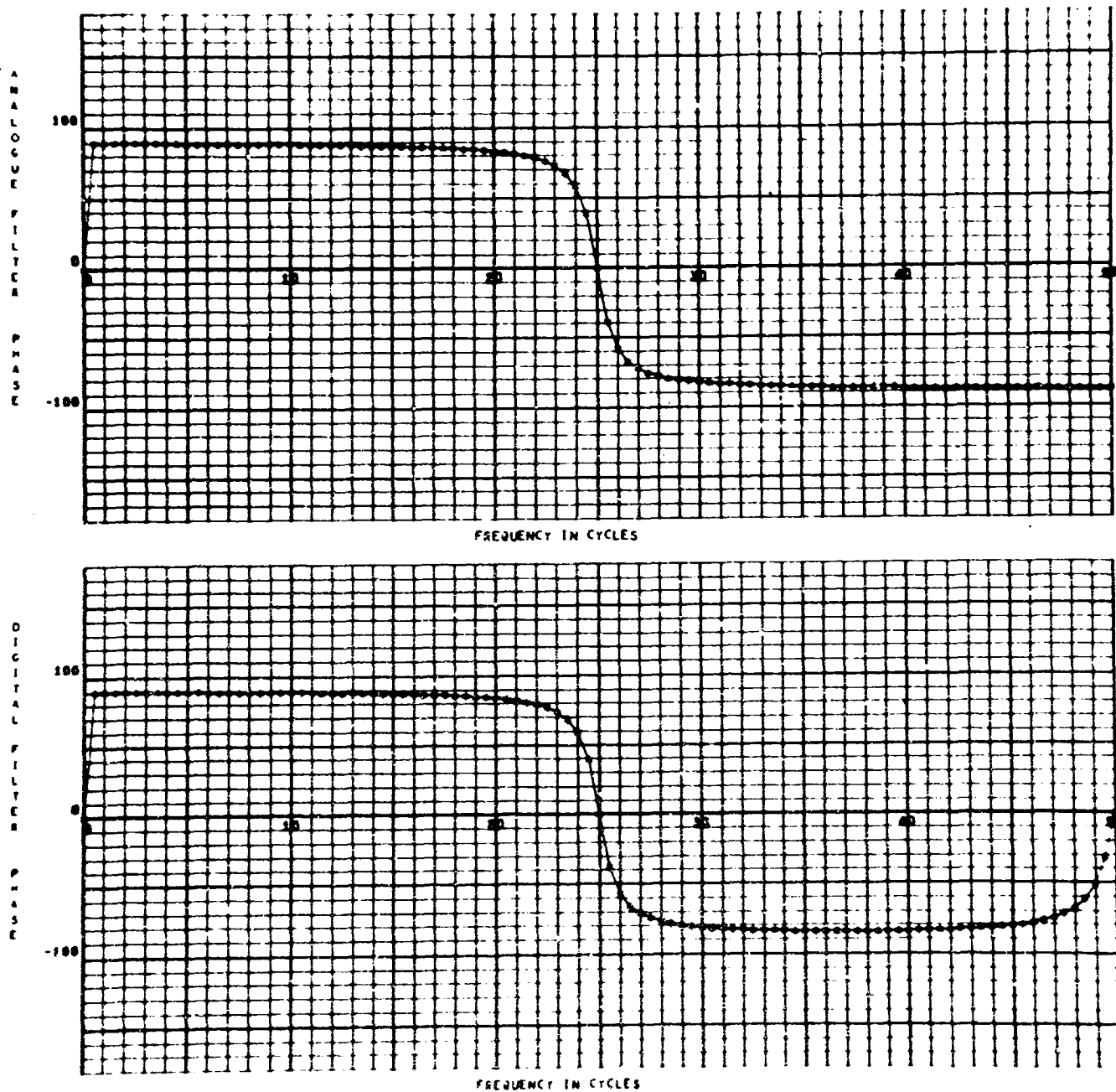


Figure 7.5-2. Continued

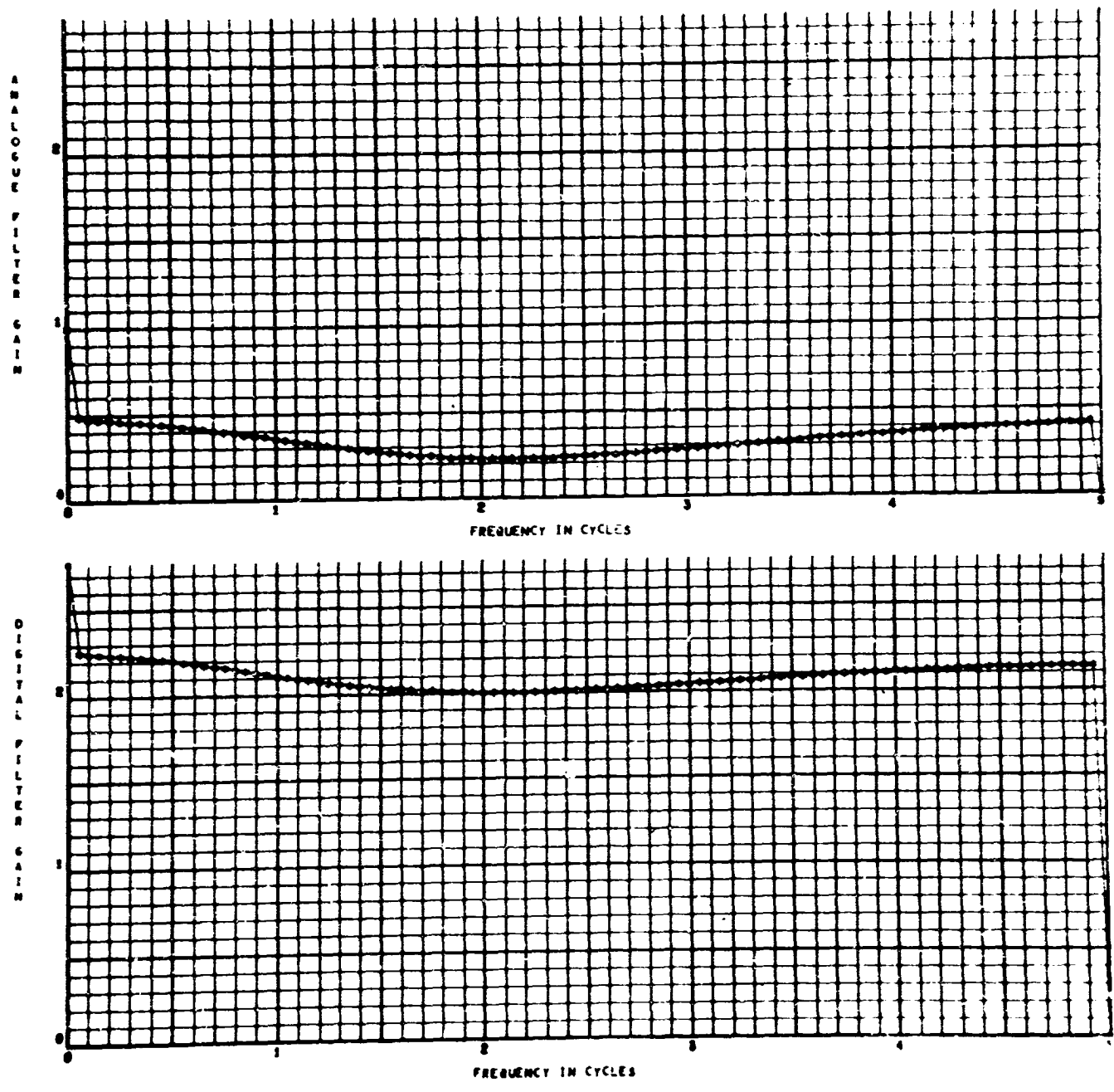


Figure 7.5-5 Control Function via z Transformation,
No Reconstruction

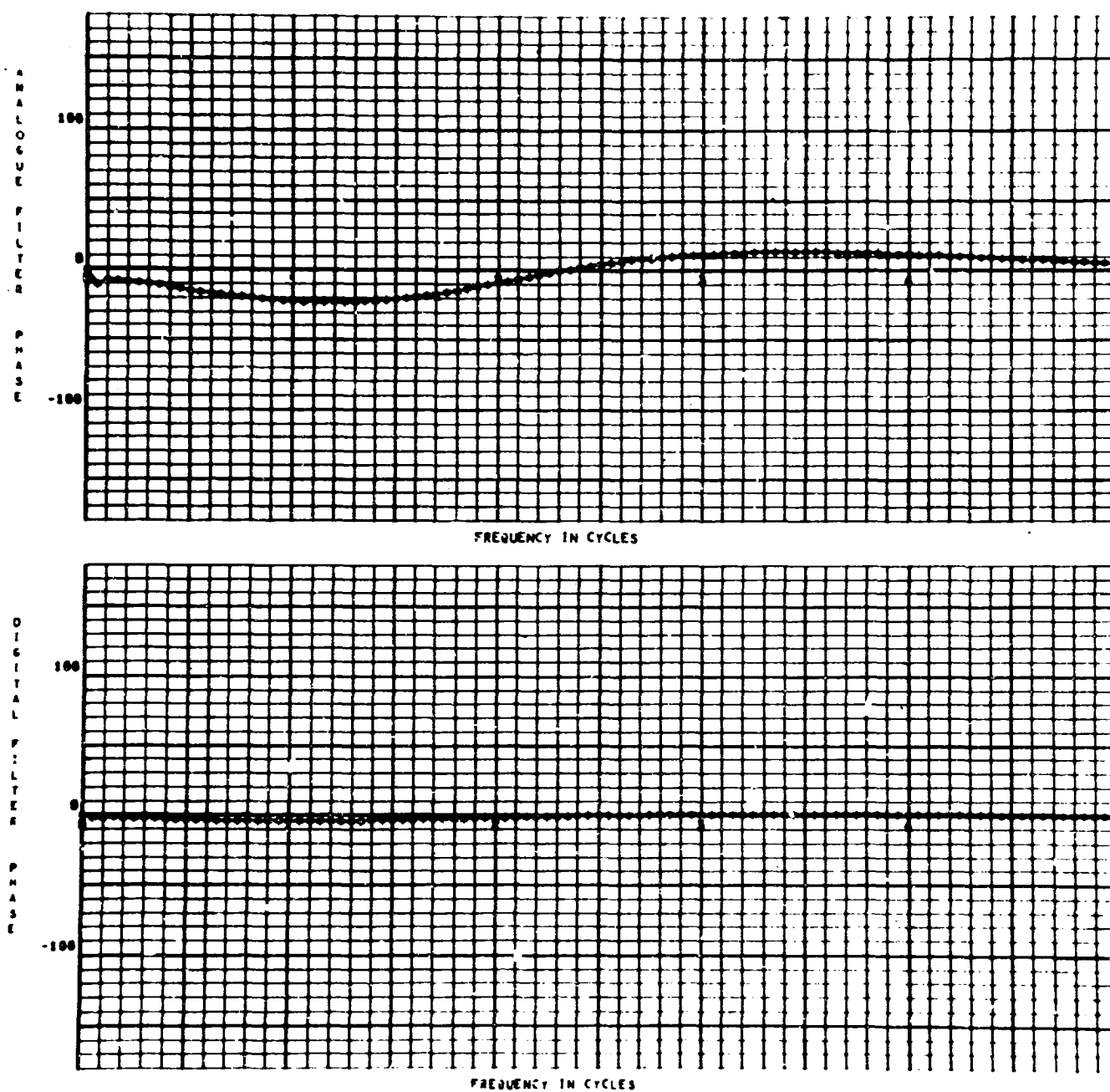


Figure 7.5-3. Continued

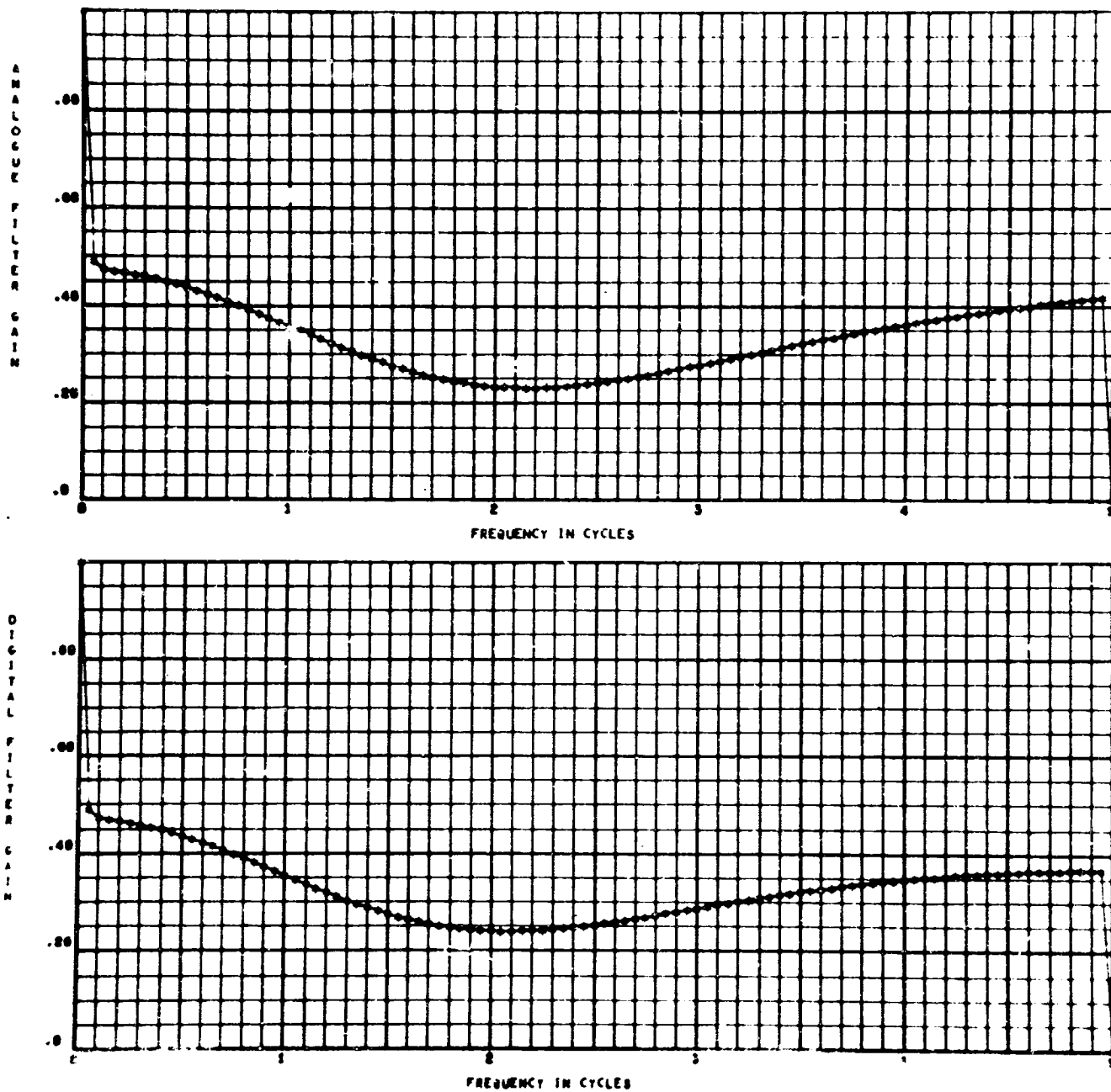


Figure 7.5.4 Control Function,
z Transformation with Delayed 1st Order S&H

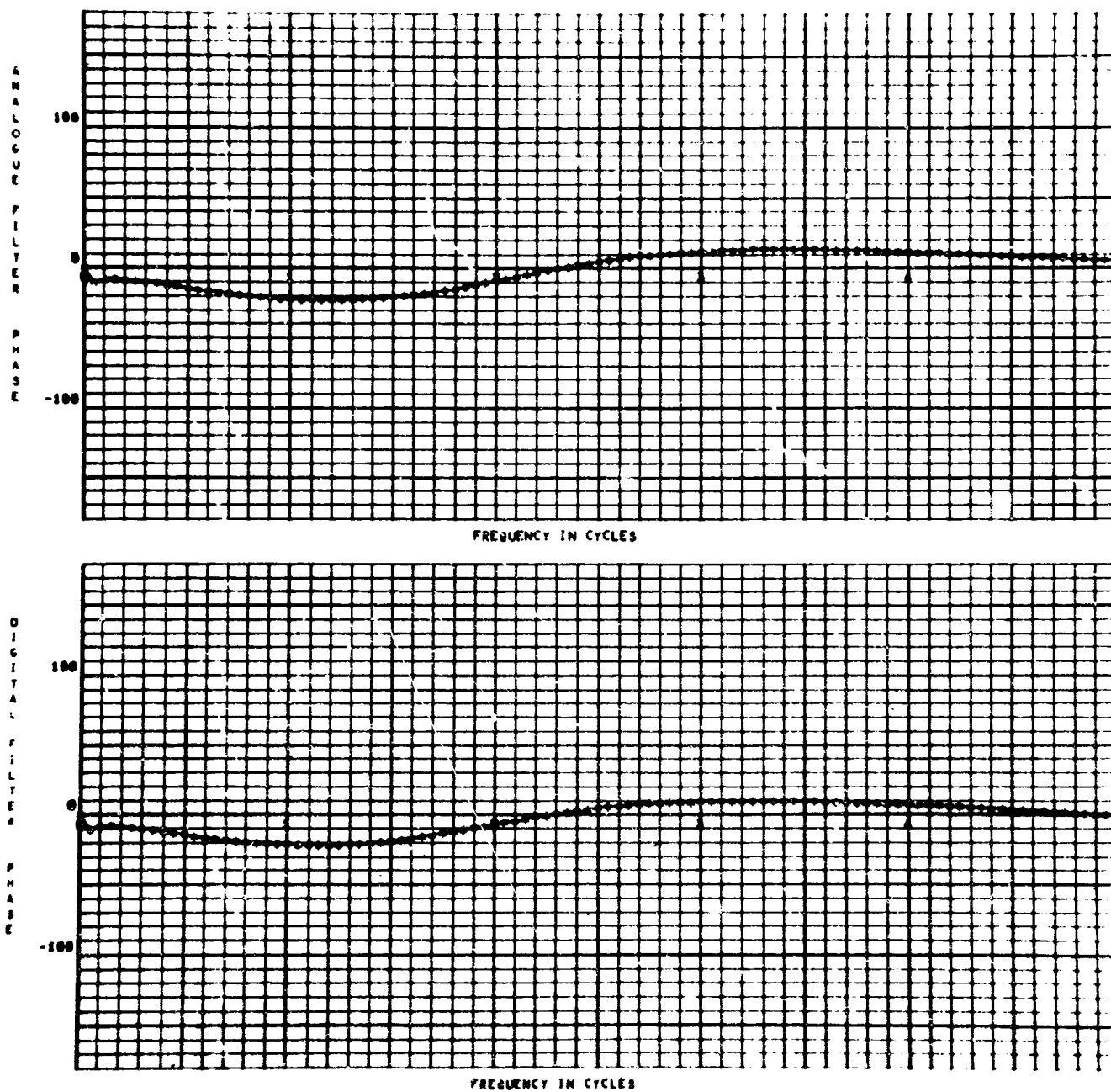


Figure 2.5.4 Continued

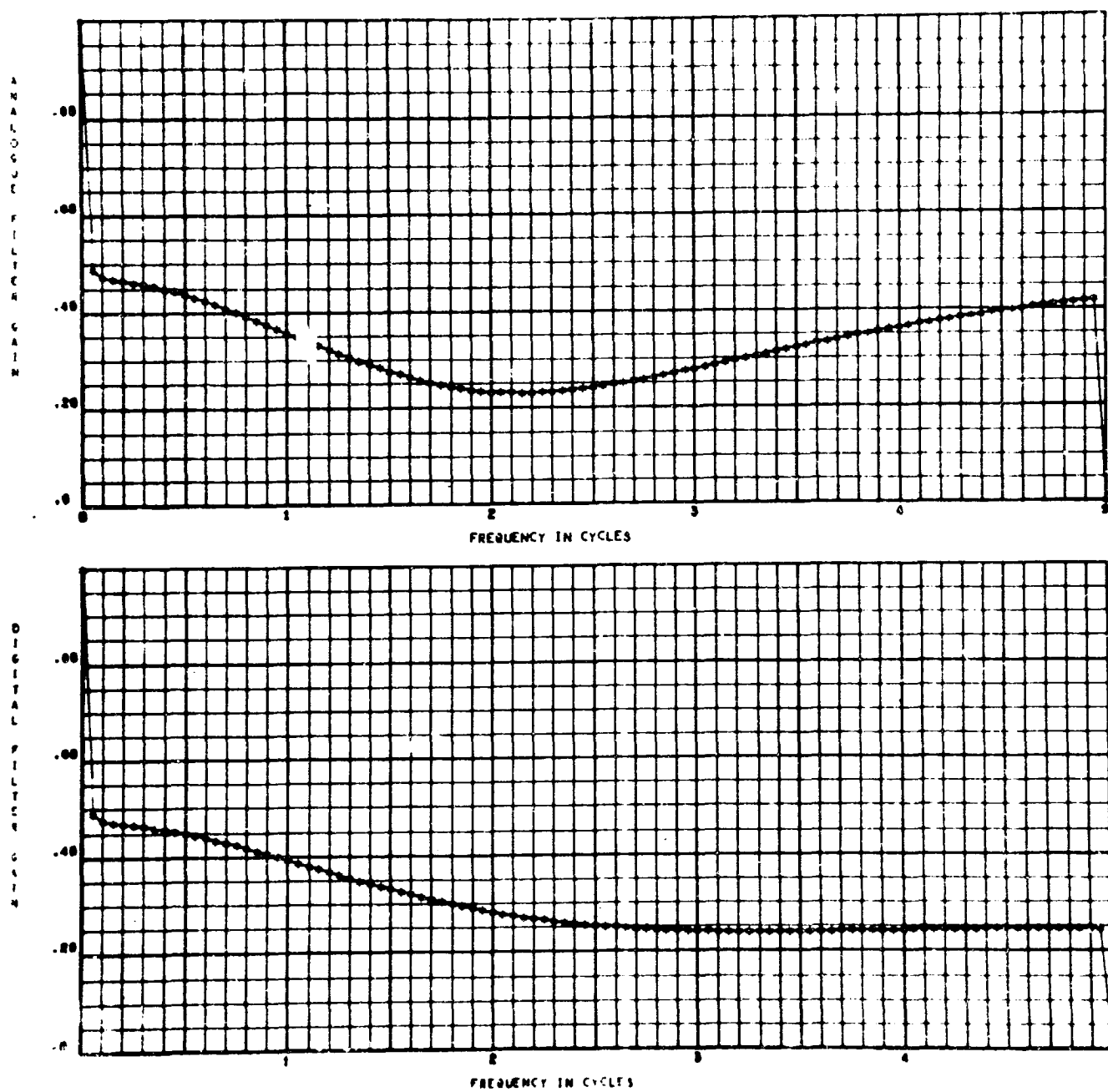


Figure 7.5.5. Control Function,
 z Transformation, Zero Order S&H

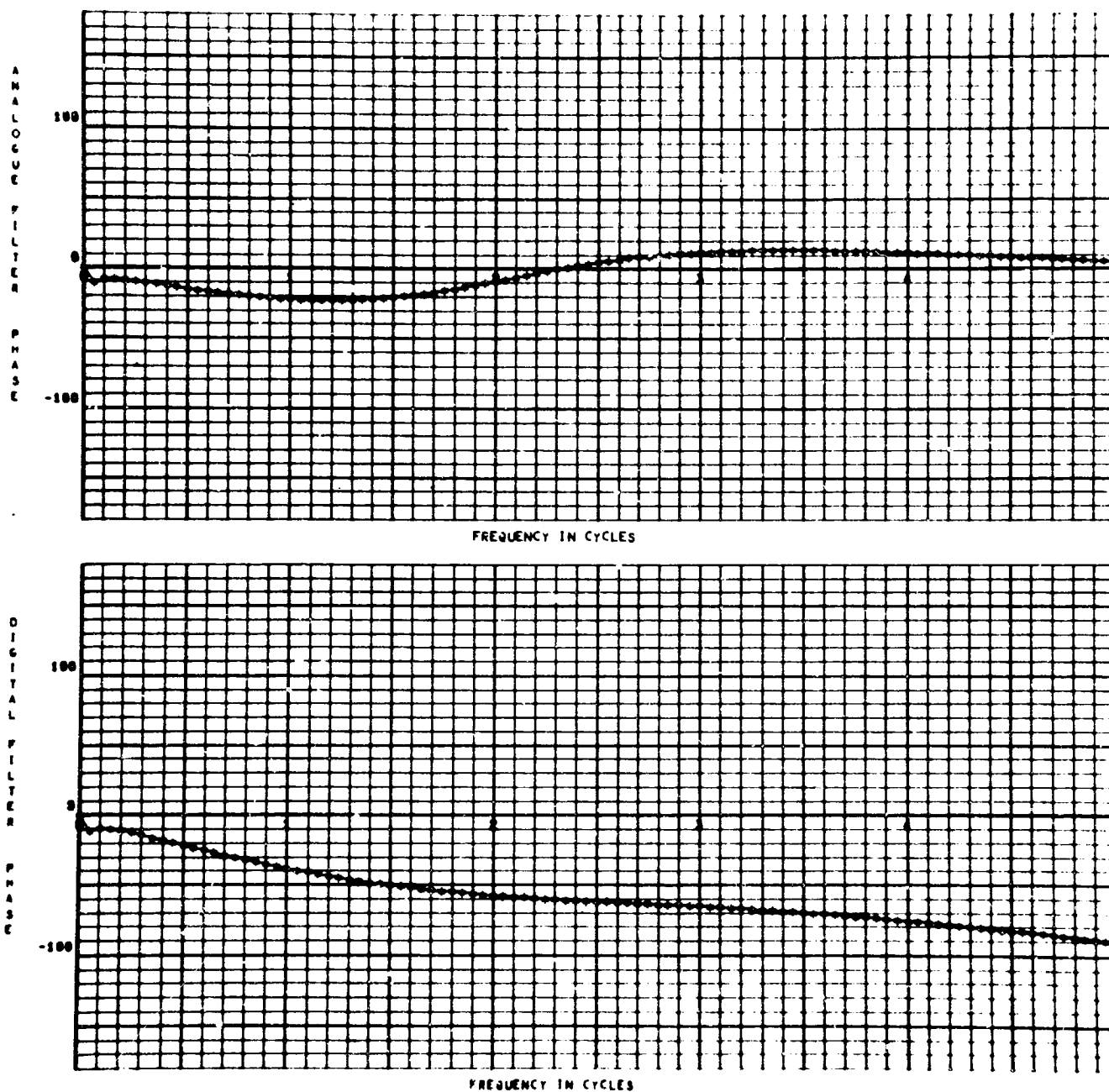


Figure 2.5-5. Continued

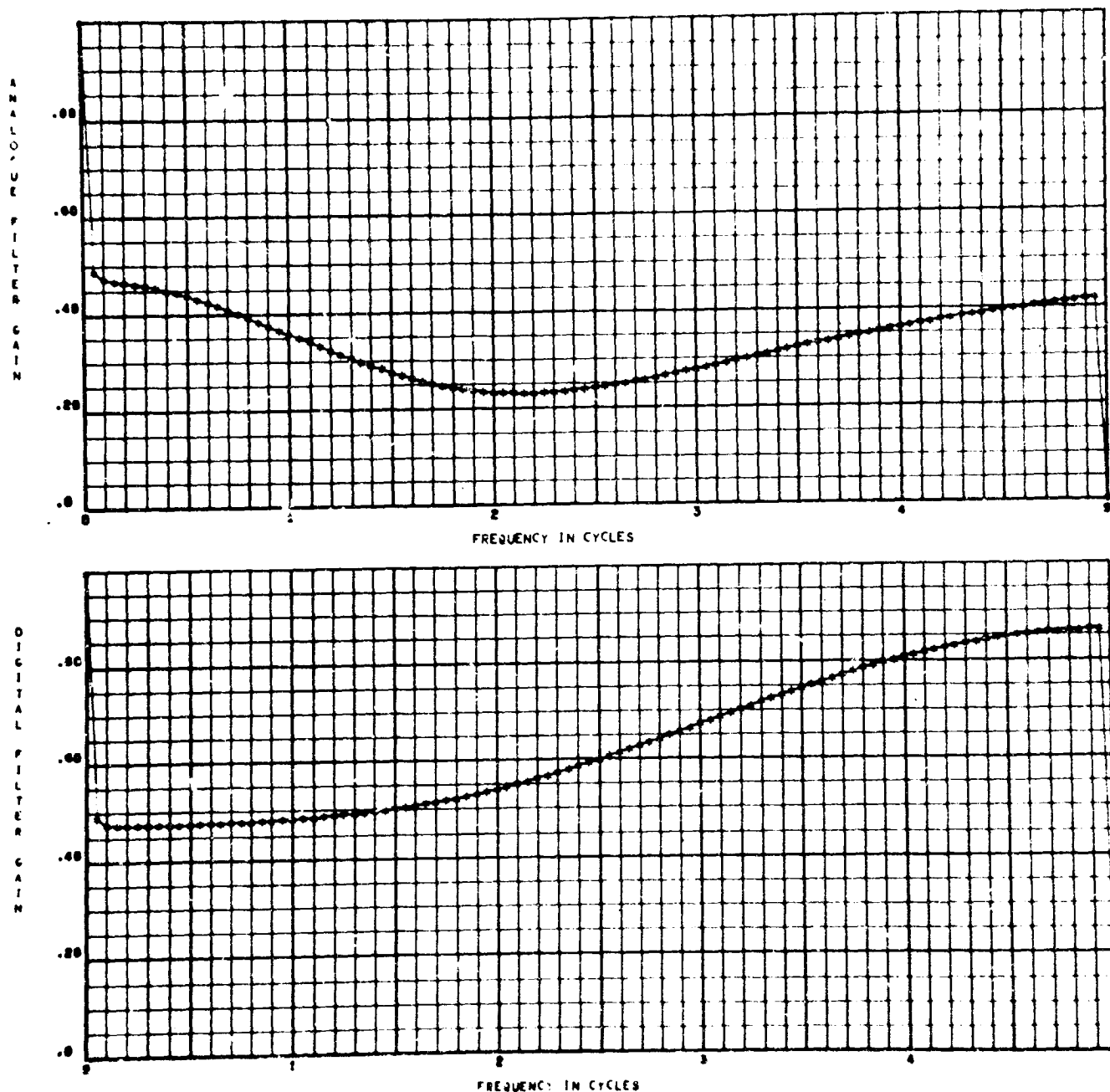


Figure 7.5.6. Control Function.
z Transformation, 1st Order Extrapolation S&H

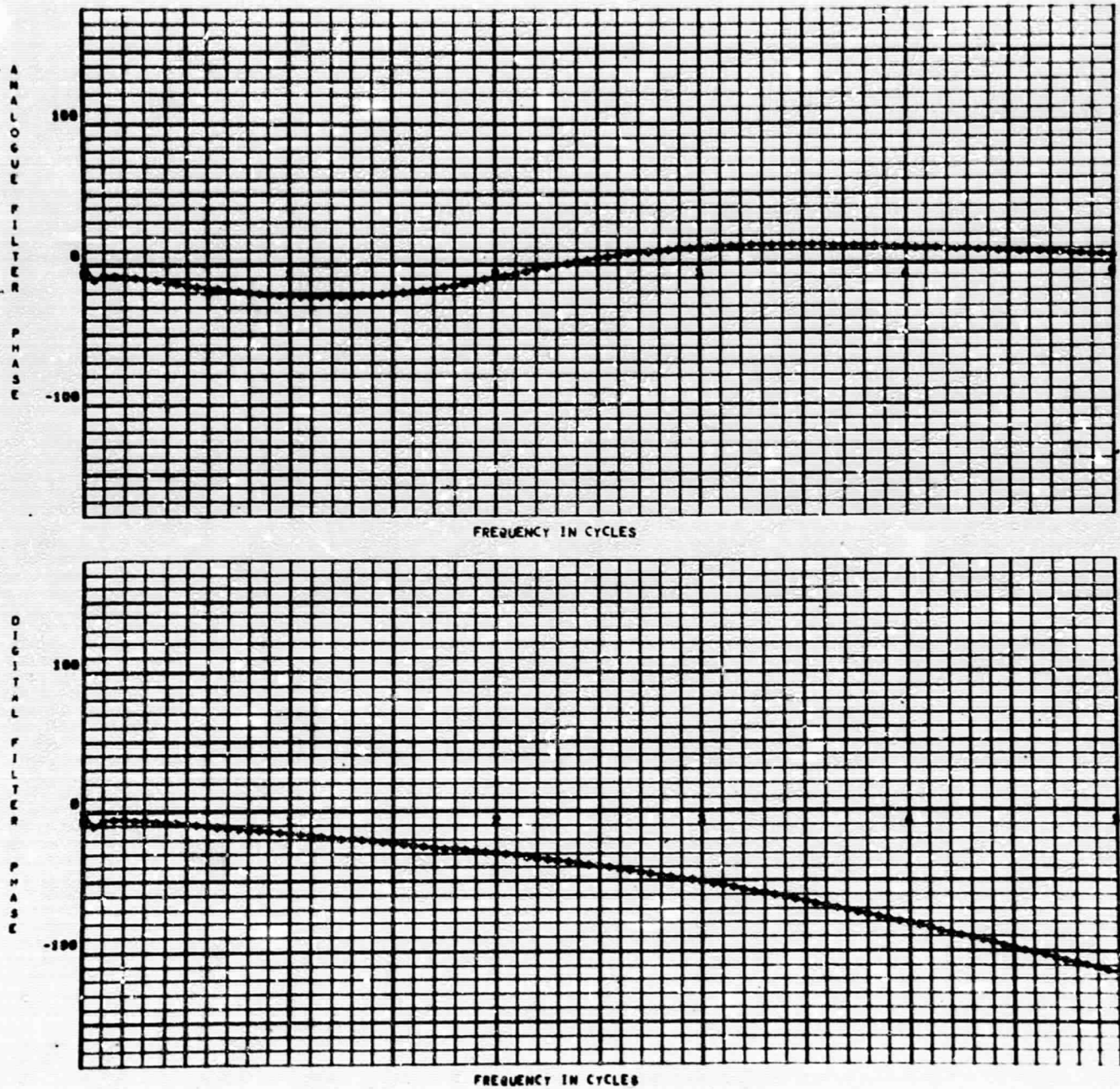


Figure 7.5-6. Continued

140016
G29 G29

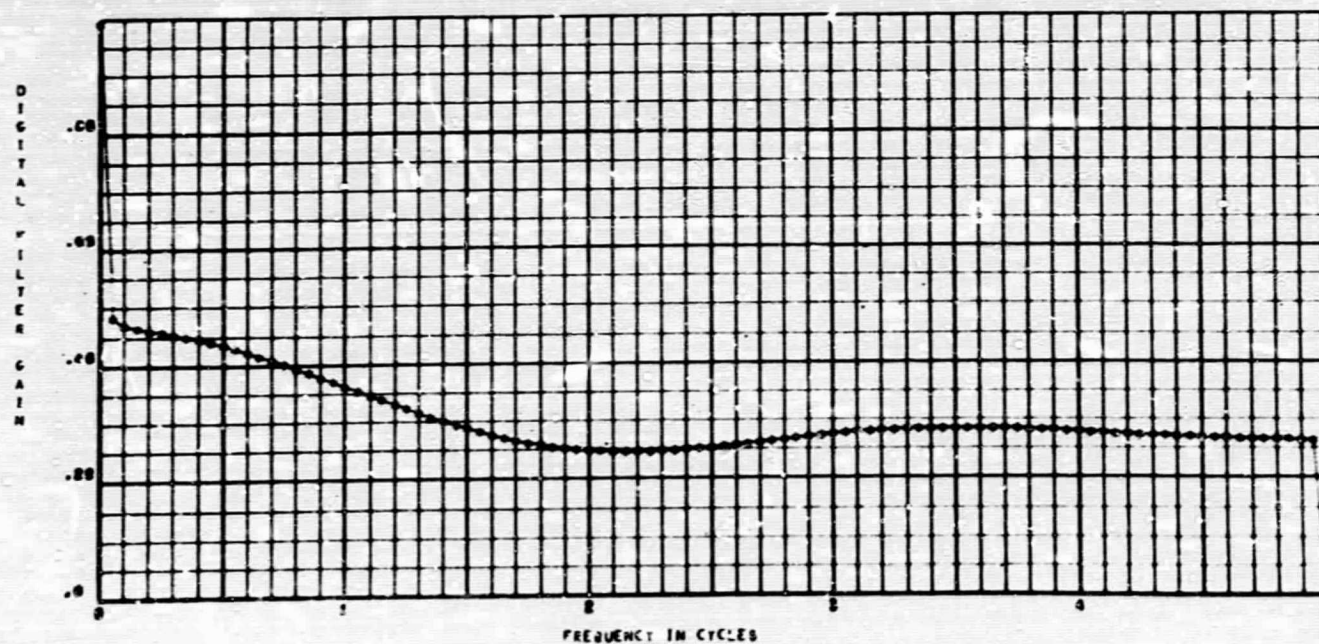
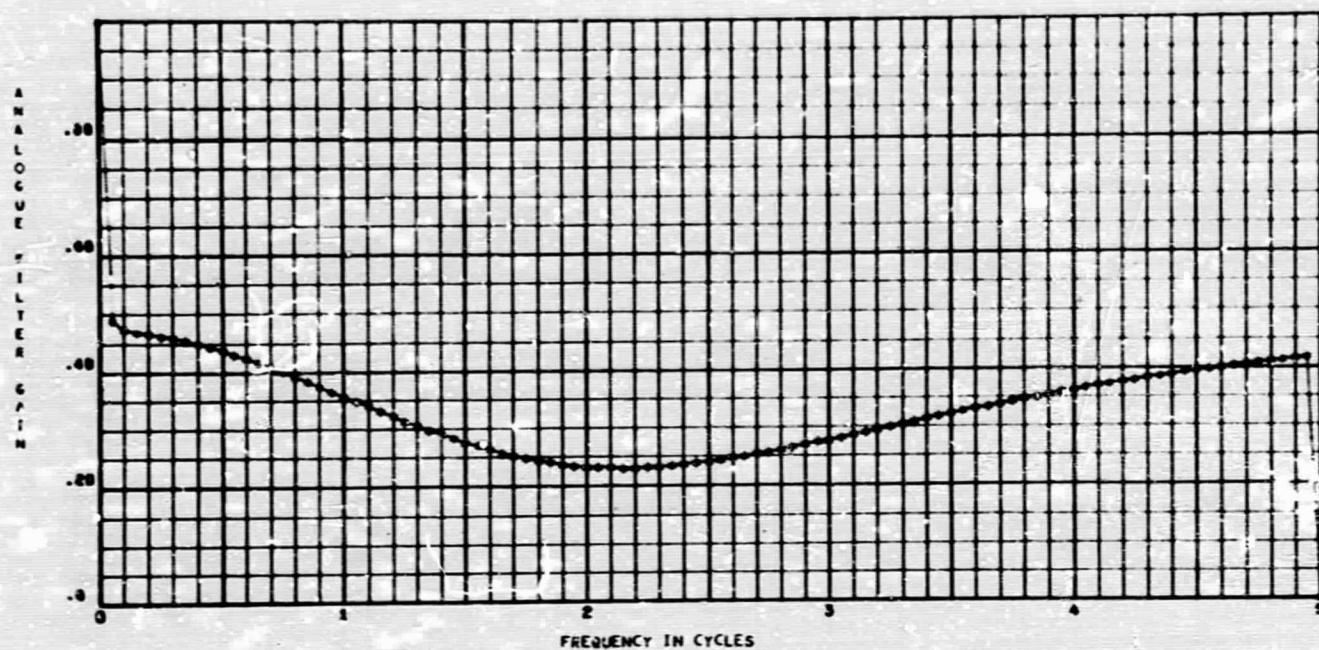


Figure 7.5-7. Control Function, z Transformation
Butterworth Recons., $\alpha = .1$, $\omega_a = 1/4 \omega_s$

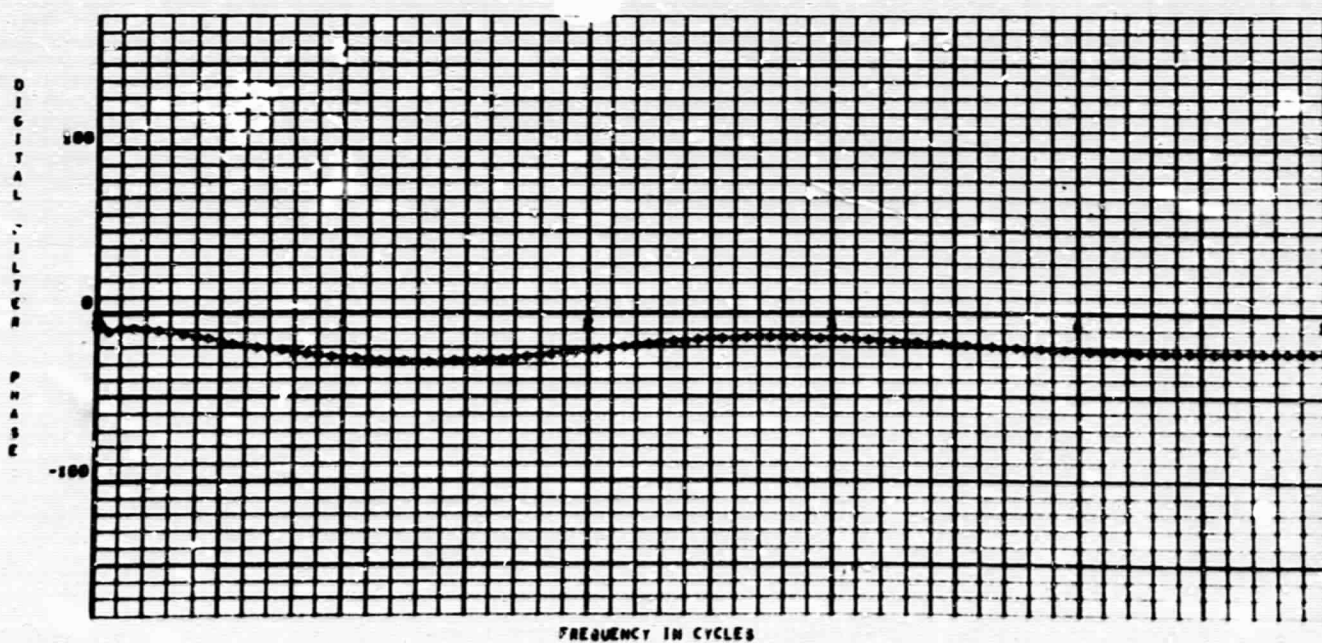
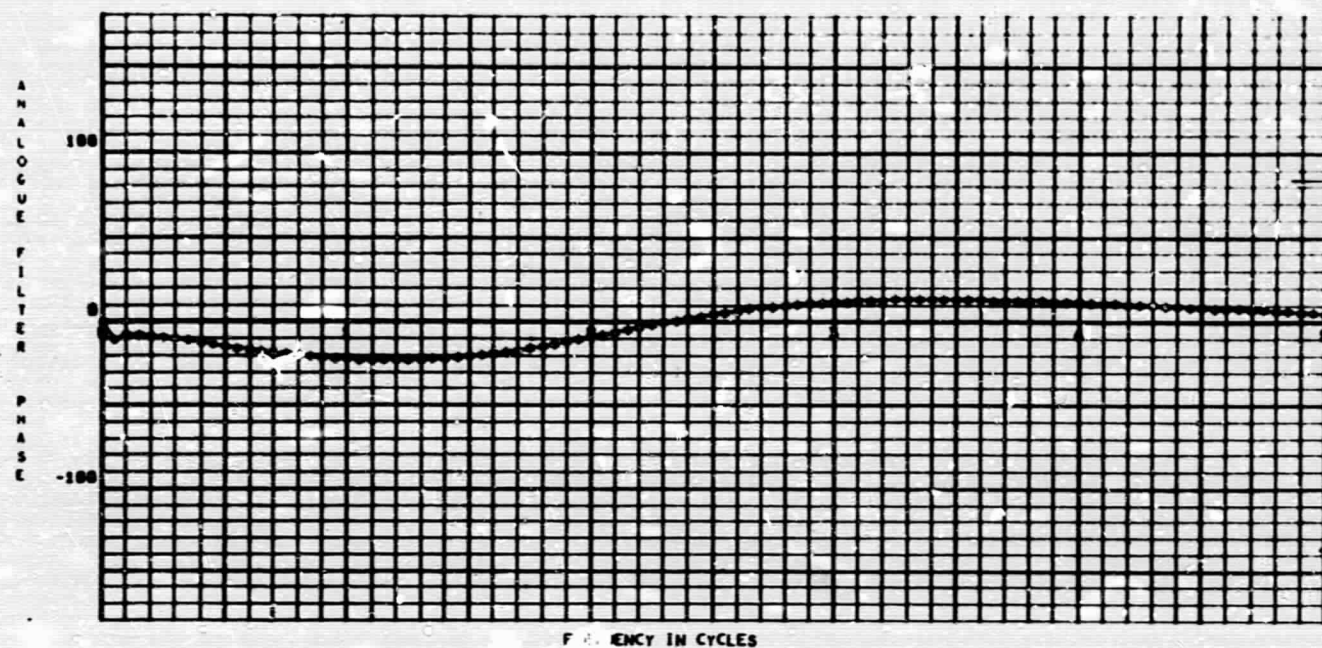


Figure 7.5-7. Continued

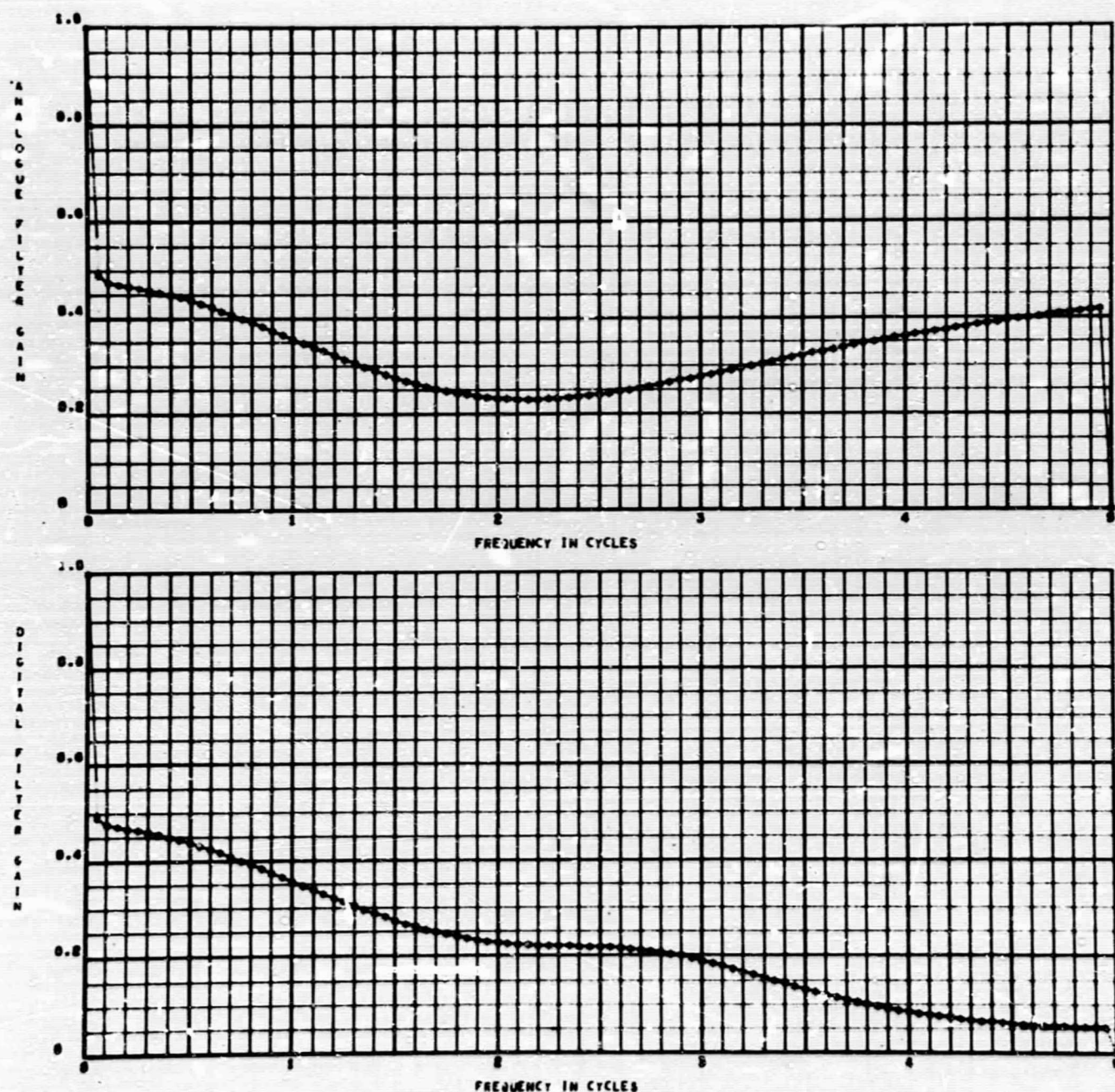


Figure 7.5-8. Control Function, z Transformation
 Butterworth Recurs.; $\alpha = .01$, $\omega_a = 1/4 \omega_s$

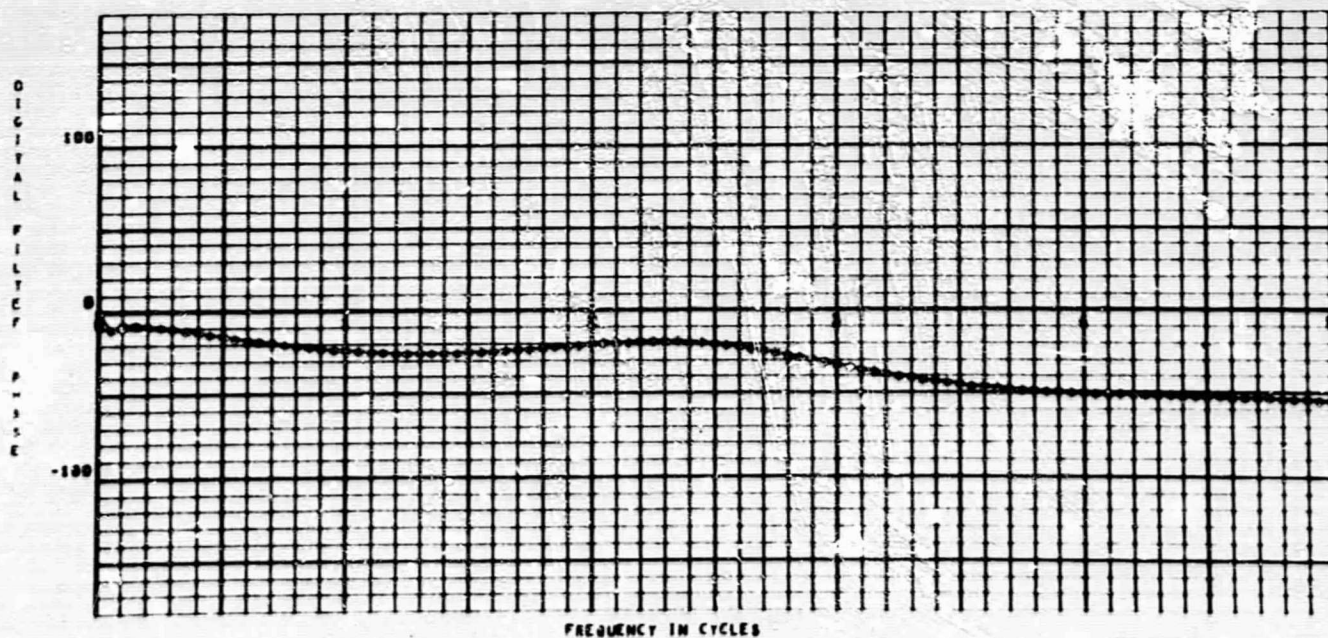
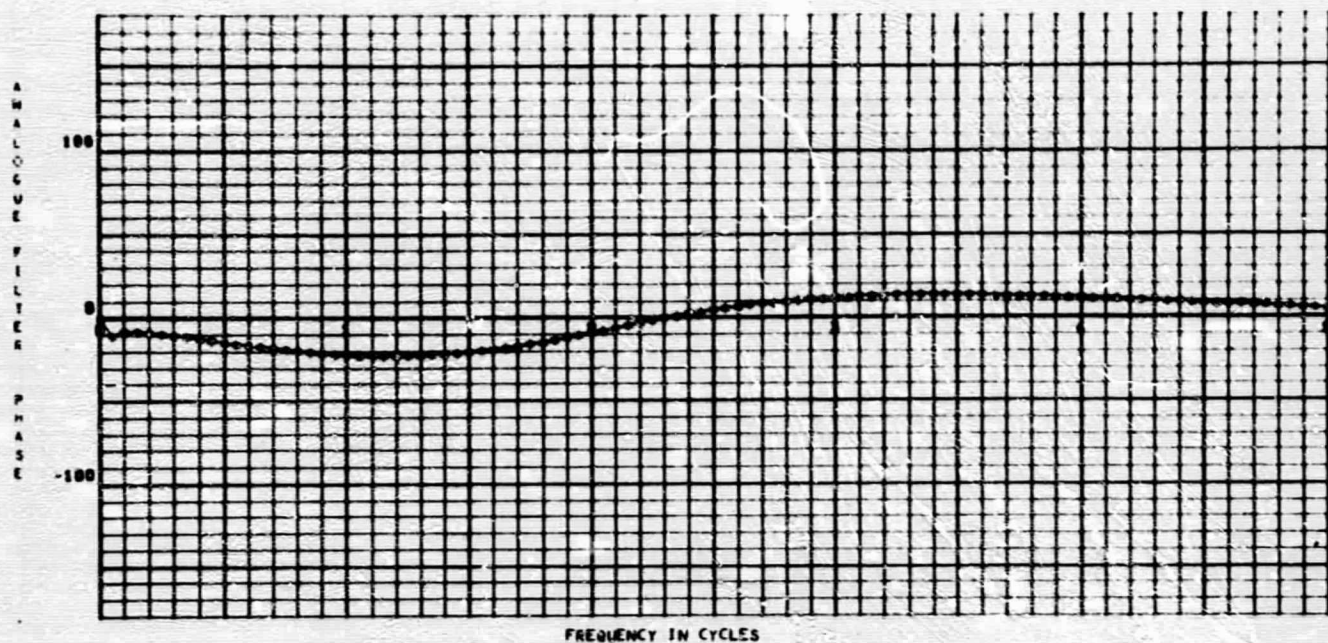


Figure 7.5-8. Continued

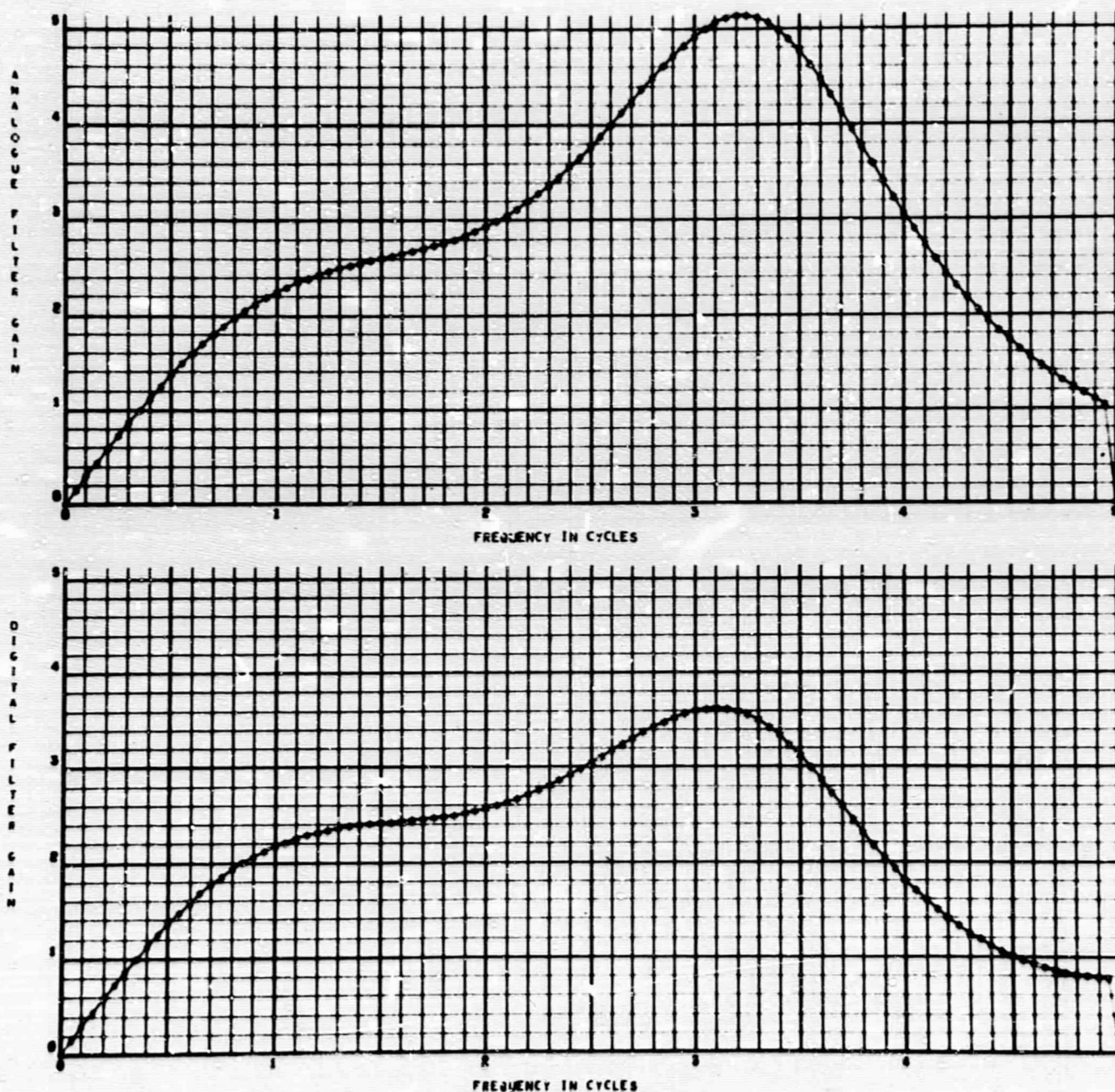


Figure 7.5-9. Derivative Control Function
 z Transformation, Compound Reconstruction Filters:
 Butterworth, $\alpha = .01$, $\omega_a = 1/3 \omega_s$; plus 1st Order Delayed S&H

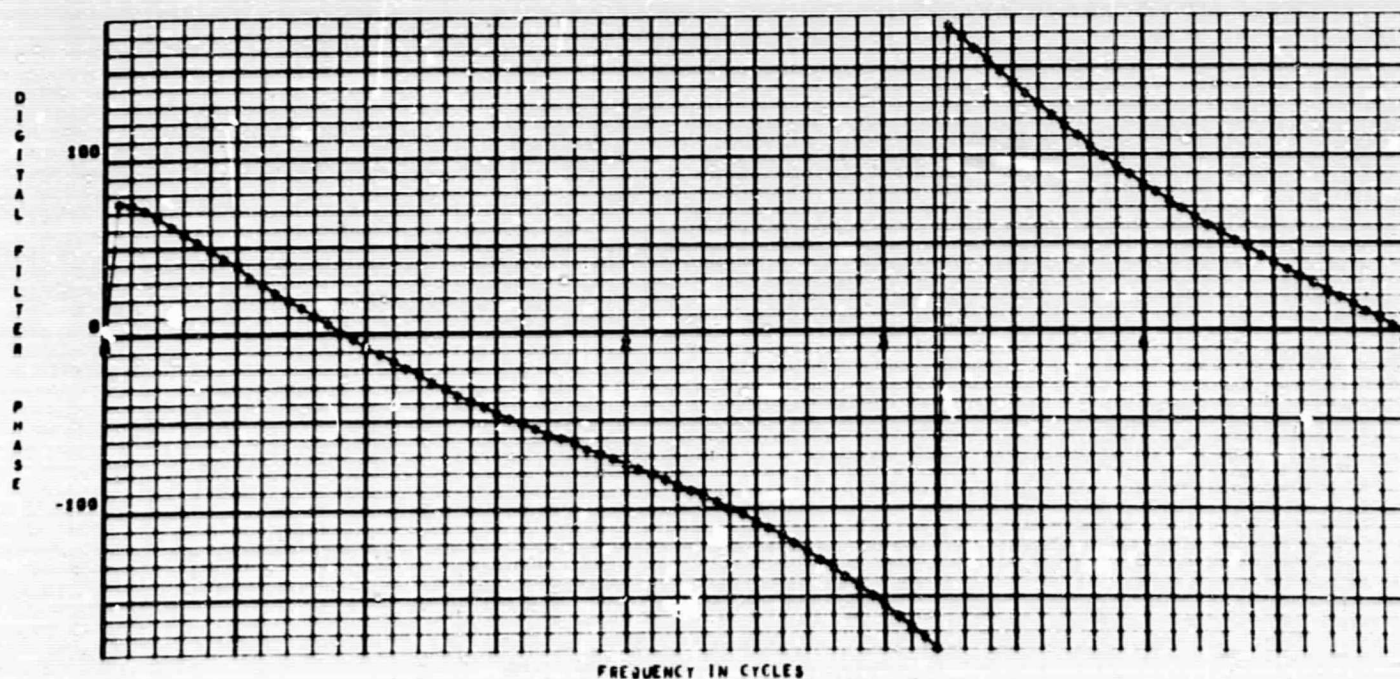
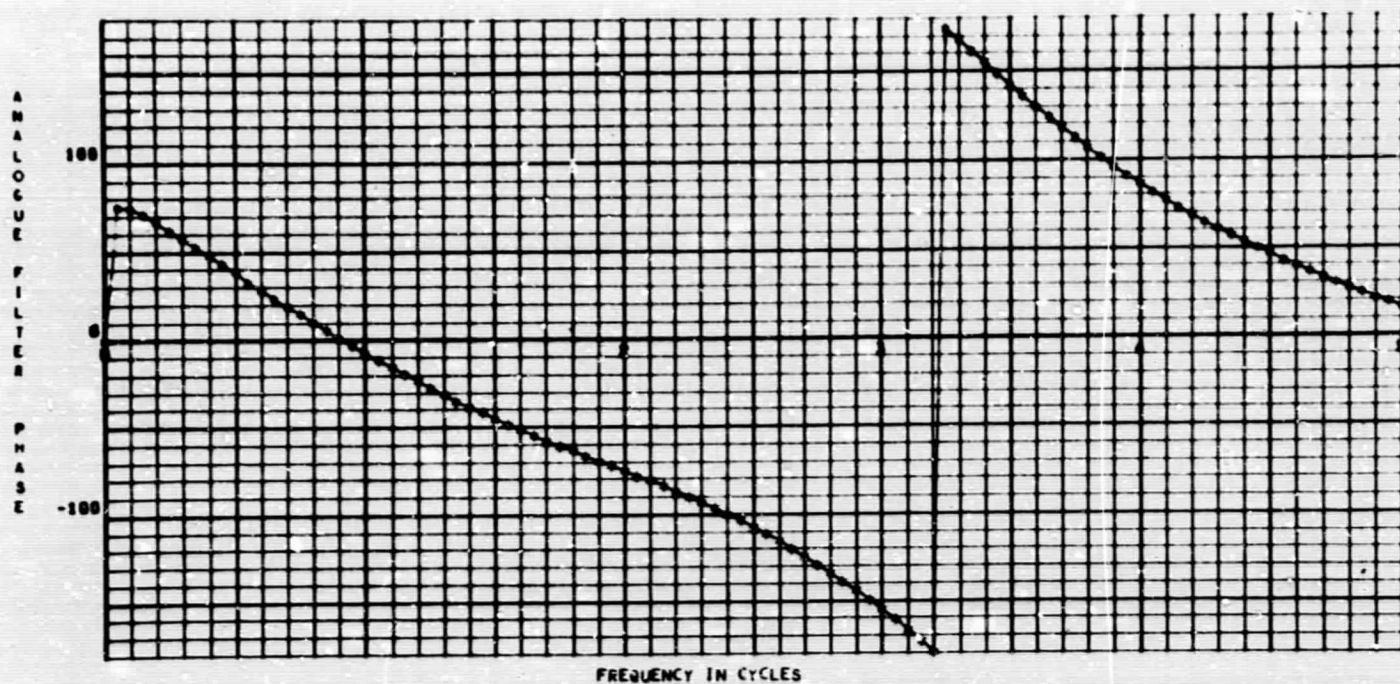


Figure 7.5-9. Continued

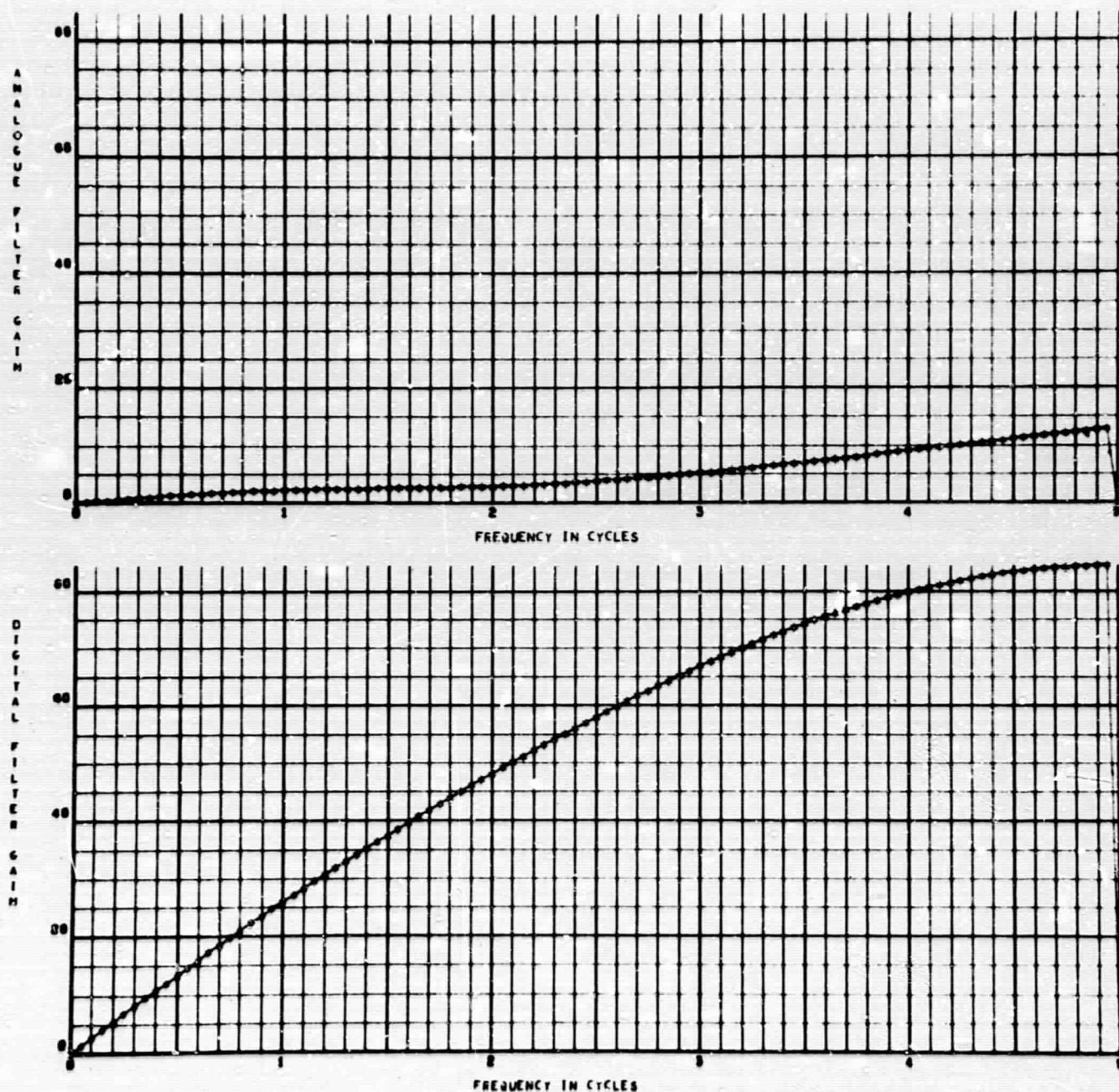


Figure 7.5-10. Derivative Control Function,
z Transformation, Zero Order S&H

100010
G4C 040

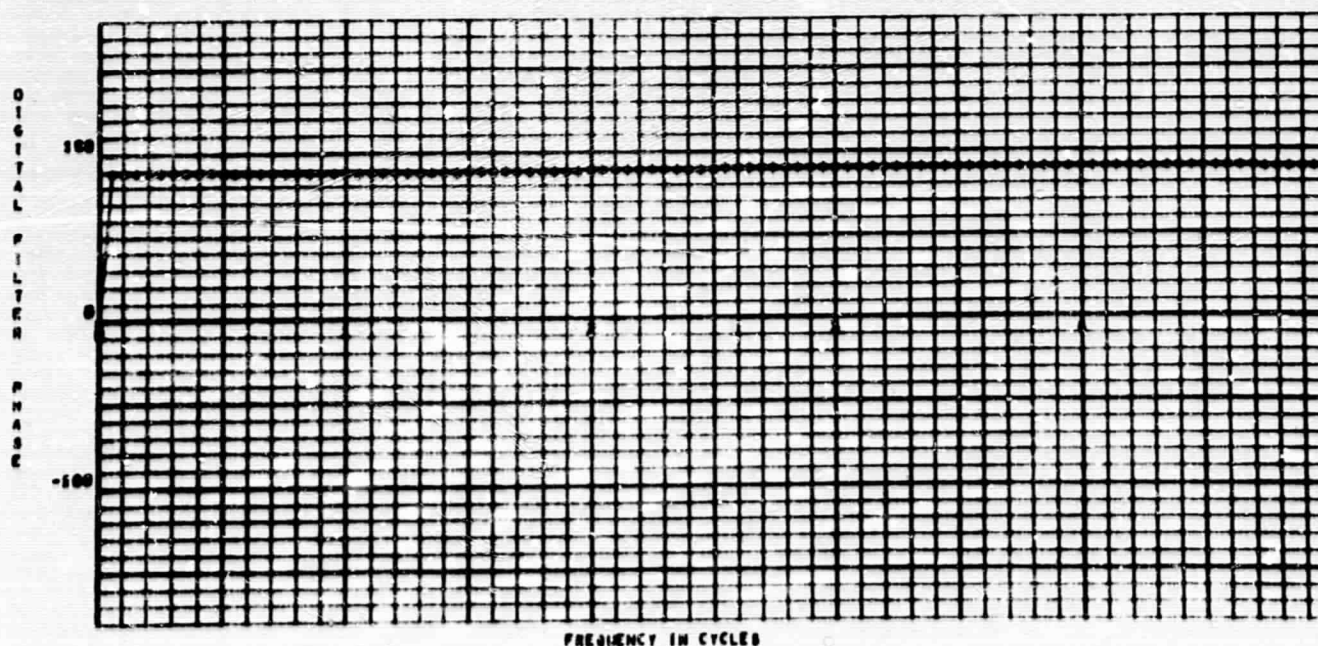
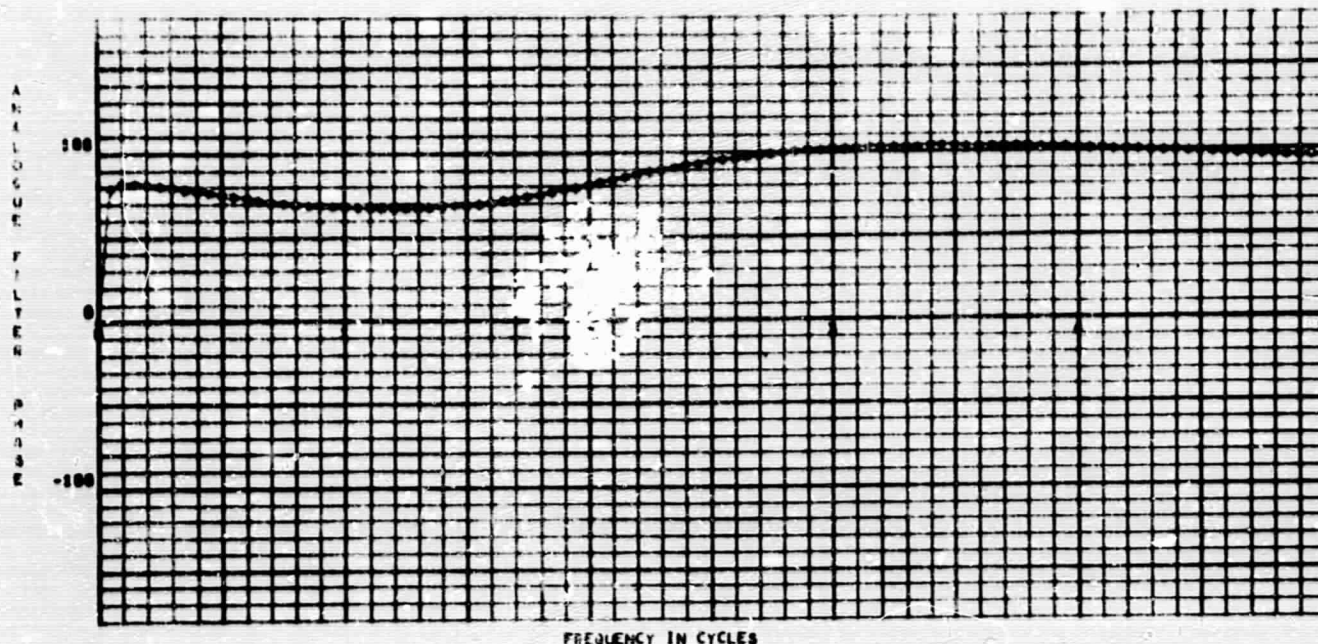


Figure 7.5-10. Continued

8. "COMPLEX" FILTERING

Discussion to this point has centered on filters with real valued impulse responses that operate upon real valued data and produce real valued outputs. Subsequent developments will demonstrate that certain, particular, important types of waveforms and filtering operations are conveniently represented in complex, or two component, form.

Digital equivalents of well-known analog formula will show that narrow-band spectrums and narrow-band filters, those that vanish outside of some frequency band which is above zero frequency, are capable of representation by two component, complex low frequency signals or impulse responses. A narrow band, often high frequency waveform can be filtered after decomposition into components by "low" frequency filters in a manner equivalent to filtering at the original frequency band.

Formula developed for narrow-band digital waveforms are identical to those for analog waveforms within a factor of T as the inverse z transform and Fourier transform are identical under the assumptions involved within a factor of T . However, it should be remembered that digital waveforms are not the same as analog waveforms unless certain special conditions are met (Section 2). One particular application discussed in some detail, with the ideas involved, is the digital filtering, or numerical integration of a bandlimited, narrow-band, high frequency waveform.

8.1 CERTAIN COMPLEX FILTER RELATIONS

Previously, a digital filter transfer function and impulse response were defined along with the interconnecting relationships (Section 3). However, emphasis was placed primarily on real impulse response filters. Certain other aspects of digital filters, when complex impulse responses are allowed, will be clarified presently.

It should be noted that the property $H(-\omega) = H^*(\omega)$ for real impulse response filters does not hold for complex impulse response filters. This means that the entire range $-\omega_s/2 \leq \omega \leq \omega_s/2$ is needed to describe a complex filter. However if the filter is decomposed into real and imaginary part impulse responses then each of the constituents possesses the property.

In this section the z transform will be applied to the complex filter form. The filter will be seen to be describable by a z transfer function exactly as in the real number case; stability conditions will also be derived. A more general filter form will be constructed from the simpler form. It is sufficient to discuss only the more simple form at first.

A digital filter is defined by the relations (Section 3.

$$u_n = \sum_{k=-\infty}^{\infty} w_{n-k} v_k$$

or

(8.1-1)

$$u_n = \sum_{k=0}^{\infty} w_k v_{n-k}$$

where

u_n = the output sequence,

= $f_n + ig_n$; f_n, g_n real;

v_n = the input sequence

= $x_n + iy_n$; x_n, y_n real;

and

$w_k = w_k^R + i w_k^I$; w_k^R, w_k^I real and $w_k = 0, k < 0$.

Taking the single sided transform exactly as in Section 3.2

$$U(z) = H_z(z) V(z) \quad (8.1-2)$$

where the various transforms are obvious:

$$\begin{aligned}
 U(z) &= \sum_{n=0}^{\infty} u_n z^{-n}, \text{ the output transform} \\
 &= \sum_{n=0}^{\infty} f_n z^{-n} + i \sum_{n=0}^{\infty} g_n z^{-n};
 \end{aligned}
 \tag{8.1-3}$$

$$\begin{aligned}
 V(z) &= \sum_{k=0}^{\infty} k z^{-k}, \text{ the input transform.} \\
 &= \sum_{k=0}^{\infty} x_k z^{-k} + i \sum_{k=0}^{\infty} y_k z^{-k};
 \end{aligned}
 \tag{8.1-4}$$

and

$$\begin{aligned}
 H_z(z) &= \sum_{j=0}^{\infty} w_j z^{-j}, \text{ the transfer function} \\
 &= \sum_{j=0}^{\infty} w_j^R z^{-j} + i \sum_{j=0}^{\infty} w_j^I z^{-j}
 \end{aligned}
 \tag{8.1-5}$$

As before the ratio

$$\frac{U(z)}{V(z)} = H_z(z)$$

is only dependent on the weights, or filter impulse response, w_j .

By defining the two "real" transfer functions

$$H_z^R(z) = \sum_{j=0}^{\infty} w_j^R z^{-j} \quad (8.1-6a)$$

and

$$H_z^I(z) = \sum_{j=0}^{\infty} w_j^I z^{-j} \quad (8.1-6b)$$

the total transfer function may be written

$$H_z(z) = H_z^R(z) + i H_z^I(z) \quad (8.1-6c)$$

Similarly, by separating the input and output transforms into "real" and "imaginary" parts

$$U(z) = F(z) + iG(z) \quad (8.1-7a)$$

and

$$V(z) = X(z) + iY(z) \quad (8.1-7b)$$

can be written where F , G , X and Y are the transforms of f_n , g_n , x_k , and y_k , respectively.

Upon using these forms in equation (8.1-2) for $H_z(z)$, $U(z)$, and $V(z)$, and equating coefficients of unity and $i = \sqrt{-1}$

$$F(z) = H_z^R(z) X(z) - H_z^I(z) Y(z)$$

and

$$G(z) = H_z^I(z) X(z) + H_z^R(z) Y(z) \quad (8.1-8)$$

Obviously, these equations express in transform form the physical filtering that is actually performed. They are the transforms of equation (8.1-1) which expressed the impulse response form of filtering which is implemented.

Examination of the proof of Section 3.4.2 indicates that the stability criteria is identical; namely

$$\sum_{j=0}^{\infty} |w_j| < \infty \quad (3.4.2-6)$$

if bounded outputs are to be obtained for bounded inputs. As w_j is composed of two "real" filters

$$|w_j|^2 = (w_j^R)^2 + (w_j^I)^2 \quad (8.1-9)$$

which implies that

$$\sum_{j=0}^{\infty} |w_j^R| < \infty \quad (8.1-10)$$

and

$$\sum_{j=0}^{\infty} |w_j^I| < \infty$$

That is, both of the constituent filters that are actually implemented must satisfy the stability criteria individually which seems intuitively obvious.

In turn, the stability criteria implies not only a constraint on the impulse response but, as for totally real impulse responses, also that the singularities or poles of $H_z(z)$ must lie inside the unit circle in the z plane.

As is the case for "real" digital filters the output sequence is given by the inverse z transform (Section 3.2).

$$\begin{aligned} u_n &= \frac{1}{2\pi i} \oint U(z) z^n \frac{dz}{z}; \\ &= \frac{1}{2\pi i} \oint H_z(z) V(z) z^n \frac{dz}{z} \end{aligned} \quad (8.1-11)$$

Assuming that all poles are within the unit circle and that the integrand is analytic elsewhere, evaluation of the contour about the unit circle gives

$$u_n = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H_z(e^{i\omega T}) V(e^{i\omega T}) e^{i\omega n T} d\omega \quad (8.1-12)$$

where $\omega_s = 2\pi/T$, T = the sampling interval. If the poles do not lie within the unit circle then (8.1-11) must be used.

By abbreviating notation, $H_z(\omega)$ for $H_z(e^{i\omega T})$ as has been done many times previously, and $V_z(\omega)$ for $V(e^{i\omega T})$ then the inverse transform can be written

$$u_n = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H_z(\omega) V_z(\omega) e^{i\omega n T} d\omega \quad (8.1-13)$$

Interpretation of this form of the inverse z transform can often be made, in many cases of analytical and practical interest, in a manner identical to the Fourier transform (Section 2.1) with t considered only at multiples of a sampling time T .

Complex filters possess a "frequency response" exactly as "real" filters.

Assume a complex sinusoid input

$$v_n = e^{i\omega n T}, \quad T = \text{sampling interval} \quad (8.1-14)$$

and an output of the same form modified by a constant (complex), $G(\omega)$:

$$u_n = G(\omega) e^{i\omega n T} \quad (8.1-15)$$

Then, from relation (8.1-1) it is found that

$$G(\omega) = \sum_{k=0}^{\infty} w_k e^{-i\omega k T} \quad (8.1-16)$$

exactly as for real filters.

Comparing this result with (8.1-5) it can be seen that

$$G(\omega) = \lim_{z \rightarrow e^{i\omega T}} (e^{i\omega T}),$$

$$= H_z(z) \Big|_{z = e^{i\omega T}}.$$
(8.1-17)

A simple generalization of a "complex" filter can also be made to "two-dimensional" vector filters. The fact that the input consists of two components, x_k and y_k , allows a more general linear operation expressed by four independent filtering operations, namely

$$f_n = \sum_{k=-\infty}^{\infty} \left[w_{n-k}^{(11)} x_k + w_{n-k}^{(12)} y_k \right]$$

and

$$g_n = \sum_{k=-\infty}^{\infty} \left[w_{n-k}^{(21)} x_k + w_{n-k}^{(22)} y_k \right]$$
(8.1-18)

where all $w_j^{(mn)} = 0$ for $j < 0$.

This generalization allows, among other things, more freedom in the design of optimal filters for noise rejection which will not be discussed here.

Alternatively, the more general filter can be implemented with two of the simpler filters (equation 8.1-1) in terms of complex arithmetic which is often more useful. Given two complex impulse responses w_j^1 and w_j^2 such that

$$w_j^1 = w_j^{1R} + i w_j^{1I}$$

and

$$w_j^2 = w_j^{2R} + i w_j^{2I}$$
(8.1-19)

then the general filtering operation may be expressed by

$$v_n = f_n + ig_n \quad (8.1-20)$$

$$= \sum_{n=-\infty}^x w_{n-k}^1 v_k + \sum_{k=-\infty}^{\infty} w_{n-k}^2 v_k^*$$

where

$$v_k = x_k + iy_k$$

The w_j 's of the complex representation are related to the general representation by

$$w_j^{11} = w_j^{1R} + w_j^{2R},$$

$$w_j^{12} = -w_j^{1I} + w_j^{2I},$$

$$w_j^{21} = w_j^{1I} + w_j^{2I},$$

$$w_j^{22} = w_j^{1R} - w_j^{2R}.$$

(8.1-21)

As the transformation of the representation is linear, and the determinant non-zero, the two complex impulse responses are also independent filters.

8.2 FUNCTIONAL FORM OF "HIGH FREQUENCY," NARROW-BAND SEQUENCES AND IMPULSE RESPONSES

If a spectrum or filter of a real sequence or impulse response has significant value only in a certain region, say

$$\omega_0 - \omega_1 \leq |\omega| \leq \omega_0 + \omega_1; \quad \omega_1 < \omega_0 < \omega_s/2;$$

then the sequence, or impulse response, can be written in a certain form. This form is that of a "modulated carrier," namely

$$s(kT) = x(kT) \cos(\omega_0 kT) - y(kT) \sin(\omega_0 kT) \quad (8.2-1)$$

where:

$s(kT)$ = signal or impulse response

$x(kT)$ = "low" frequency cosine modulation

$y(kT)$ = "low" frequency sine modulation.

The spectrums of $x(kT)$ and $y(kT)$ extend from $-\omega_1$ to ω_1 and, therefore, deserve the connotation of "low" frequency modulations.

The proof of these remarks is relatively simple if the appropriate definitions are used. Transforms will be abridged, for obvious simplicity, so that, e.g., $F(\omega)$ will replace $F_z(e^{i\omega T})$.

Let $S(\omega)$, the signal spectrum or filter transfer function, be non-zero (which can only be approximately true in physical systems) only in the region $\omega_0 - \omega_1 < |\omega| < \omega_0 + \omega_1$, then the time signal or impulse response can be obtained by the inverse z transform:

$$s(kT) = \frac{1}{2\pi i} \oint S(z) z^n \frac{dz}{z} \quad (8.2-2)$$

where $S(z)$ is the sequence spectrum z transform (Section 8.1). In terms of frequency let $z = e^{i\omega T}$, as usual, and have

$$s(kT) = \frac{T}{2\pi} \int_{\frac{-\omega_s}{2}}^{\frac{\omega_s}{2}} S(\omega) e^{i\omega nT} d\omega, \quad (8.2-3)$$

with $\omega_s = 2\pi/T$, the sampling frequency, and $S(e^{i\omega T})$ has been abridged to $S(\omega)$ for simplicity. As $S(\omega)$ has been assumed zero outside of a narrow-band region

$$\omega_s s(kT) = \int_{\omega_0 - \omega_1}^{\omega_0 + \omega_1} S(\omega) e^{i\omega kT} d\omega + \int_{-\omega_0 - \omega_1}^{-\omega_0 + \omega_1} S(\omega) e^{i\omega kT} d\omega \quad (8.2-4)$$

where the negative frequencies have been included ($S(-\omega) = S^*(\omega)$). Narrow-band spectrums or filters are also assumed to be zero at zero frequency.

Changing variables in the integrands produces

$$\omega_s s(kT) = e^{i\omega_0 kT} \int_{-\omega_1}^{\omega_1} S(\omega + \omega_0) e^{i\omega kT} d\omega + e^{-i\omega_0 kT} \int_{-\omega_1}^{\omega_1} S(\omega - \omega_0) e^{i\omega kT} d\omega. \quad (8.2-5)$$

Define the two linearly independent quantities

$$X(\omega) = S(\omega + \omega_0) + S(\omega - \omega_0), \quad |\omega| < \omega_1 \quad (8.2-6)$$

and

$$Y(\omega) = -i \left[S(\omega + \omega_0) - S(\omega - \omega_0) \right], \quad |\omega| < \omega_1.$$

Both $X(\omega)$ and $Y(\omega)$ are transforms of real sequences as

$$X(-\omega) = X^*(\omega).$$

and

$$(8.2-7)$$

$$Y(-\omega) = Y^*(\omega).$$

Furthermore, both $X(\omega)$ and $Y(\omega)$ have spectrums that extend from zero to $\pm\omega_1$, namely low frequency spectrums.

In terms of $X(\omega)$ and $Y(\omega)$, the original spectrum can be written as

$$\begin{aligned} S(\omega + \omega_0) &= 1/2 \left[X(\omega) + i Y(\omega) \right], \\ S(\omega - \omega_0) &= 1/2 \left[X(\omega) - i Y(\omega) \right]. \end{aligned} \quad (8.2-8)$$

Therefore, using the inverse z transforms of $X(\omega)$ and $Y(\omega)$, $x(kT)$ and $y(kT)$, respectively, equation (8.2-5) can be written as,

$$\begin{aligned} s(kT) &= 1/2 \left[x(kT) + iy(kT) \right] e^{i\omega_0 kT} + 1/2 \left[x(kT) - iy(kT) \right] e^{-i\omega_0 kT} \\ &= \operatorname{Re} \left\{ \left[x(kT) + iy(kT) \right] e^{i\omega_0 kT} \right\} \\ &= x(kT) \cos \omega_0 kT - y(kT) \sin \omega_0 kT \end{aligned} \quad (8.2-9)$$

8.2.1 "Complex Modulation" of a Narrow Band, High Frequency Sequence.

A narrow band, high frequency sequence or waveform can be represented as a complex number, hereafter called the "complex modulation," if the center frequency or carrier is known. This fact makes the algebra and understanding of linear filtering considerably simpler, and it has the physical interpretation of removing the carrier frequency from consideration as long as it is known (along with the carrier phase).

The narrowband signal can be written as (Section 8.2).

$$\begin{aligned} s(kT) &= x(kT) \cos \omega_0 kT - y(kT) \sin \omega_0 kT \\ &= \operatorname{Re} \left\{ \left[x(kT) + iy(kT) \right] e^{i\omega_0 kT} \right\} \end{aligned} \quad (8.2-9)$$

It can be seen that if ω_0 is known, then $s(kT)$ is completely defined, via equation (8.2-9) by the quantity

$$v(kT) = x(kT) + i y(kT) \quad (8.2.1-1)$$

which will be called the "complex modulation."

Often the pair of quantities $x(kT)$ and $y(kT)$ are called "observables" in coherent radar and communications systems. "Coherent" systems are simply ones in which the carrier frequency (and phase) are known.

Similarly, it can be seen from equation (8.2-6) that the transform of a narrow band signal, $s(kT)$, can be expressed in terms of the z transform of the complex low frequency modulation:

$$\mathcal{Z}(v(kT)) = X(\omega) + i Y(\omega) \quad (8.2.1-2)$$

as long as the carrier frequency is known and $X(\omega)$ and $Y(\omega)$ are known independently.

Subsequent endeavors will show that filtering operations performed at a precisely known carrier frequency and phase can be expressed in terms of the complex modulation input, such as $v(kT)$ and an output $u(kT) = f(kT) + i g(kT)$ at the same carrier frequency.

8.3 DIGITAL FILTERING OF A NARROW-BAND, HIGH FREQUENCY SIGNAL USING THE COMPLEX MODULATION

If the frequency of the band center, ω_0 , is known then linear processing of the narrow-band signal may be performed equivalently on the low frequency complex modulation components $x(kT)$ and $y(kT)$ as is the situation for analog signals.

From equation (8.2-8), $S(\omega)$ the z transform of the narrow-band signal $s(kT)$ can be written as

$$S(\omega) = \frac{1}{2} \left[X(\omega - \omega_0) + i Y(\omega - \omega_0) \right] + \frac{1}{2} \left[X(\omega + \omega_0) - i Y(\omega + \omega_0) \right] \quad (8.3-1)$$

As $S(\omega)$ vanishes outside the region $\omega_0 - \omega_1 < \omega < \omega_0 + \omega_1$ for positive ω , and the region $-\omega_0 - \omega_1 < \omega < \omega_0 + \omega_1$ for negative ω , the effect of a linear filter $H_z(\omega)$ is to produce another narrow-band high frequency signal that vanishes in the same regions. This obviously occurs because

$$R(\omega) = H_z(\omega) S(\omega) \quad (8.3-2)$$

is the z transform of the output signal.

It is of considerable interest to find $R(\omega)$ in the form

$$R(\omega) = \frac{1}{2} \left[F(\omega - \omega_0) + i G(\omega - \omega_0) \right] + \frac{1}{2} \left[F(\omega + \omega_0) - i G(\omega + \omega_0) \right]; \quad (8.3-3)$$

where $F(\omega)$ and $G(\omega)$ possess low frequency spectrums in terms of $H(\omega)$, $X(\omega)$, and $Y(\omega)$.

To this effect the computation of $r(kT)$, the output of the filter in the time domain, is a fruitful avenue. Taking the inverse z transform of $R(\omega)$ and using the form

$$\begin{aligned} r(kT) &= \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} R(\omega) e^{i\omega kT} d\omega \\ &= \frac{1}{2} \frac{1}{\omega_s} \int_{\omega_0 - \omega_1}^{\omega_0 + \omega_1} \left[H_z(\omega) X(\omega - \omega_0) + i H_z(\omega) Y(\omega - \omega_0) \right] e^{i\omega kT} d\omega \\ &\quad + \frac{1}{2} \frac{1}{\omega_s} \int_{-\omega_0 - \omega_1}^{-\omega_0 + \omega_1} \left[H_z(\omega) X(\omega + \omega_0) - i H_z(\omega) Y(\omega + \omega_0) \right] e^{i\omega kT} d\omega \end{aligned}$$

Upon changing variables in the integrals and using the relations (8.2-7), plus $H_z^*(-\omega) = H_z(\omega)$ for the filter, the relation

$$r(kT) = -\frac{1}{2} \frac{1}{\omega_s} \int_{-\omega_1}^{\omega_1} \left[H_z(\omega + \omega_0) (X(\omega) + i Y(\omega)) \right] e^{i\omega kT} d\omega e^{i\omega_0 kT} \\ + \frac{1}{2} \frac{1}{\omega_s} \int_{-\omega_1}^{\omega_1} \left[H_z^*(\omega + \omega_0) (X^*(\omega) - i Y^*(\omega)) \right] e^{-i\omega kT} d\omega e^{-i\omega_0 kT} \quad (8.3-5)$$

is obtained.

Recognizing the second integral as the conjugate of the first, then the expression may be written as

$$r(kT) = \text{Re} \left\{ \frac{1}{\omega_s} \int_{-\omega_1}^{\omega_1} \left[H_z(\omega + \omega_0) (X(\omega) + i Y(\omega)) \right] e^{i\omega kT} d\omega e^{i\omega_0 kT} \right\} \quad (8.3-6)$$

If the integral is defined as

$$u(kT) = f(kT) + i g(kT) \\ = \frac{1}{\omega_s} \int_{-\omega_1}^{\omega_1} \left[H_z(\omega + \omega_0) (X(\omega) + i Y(\omega)) \right] e^{i\omega kT} d\omega \quad (8.3-7)$$

then

$$r(kT) = f(kT) \cos(\omega_0 kT) - g(kT) \sin(\omega_0 kT)$$

which is the time representation expected (see Section 8.2). To find $f(kT)$ and $g(kT)$ and their transforms, as desired, the simple relations

$$f(kT) = \frac{1}{2} \left[u(kT) + u^*(kT) \right]$$

and

$$g(kT) = \frac{-i}{2} \left[u(kT) - u^*(kT) \right] \quad (8.3-8)$$

can be used effectively. Now $u^*(kT)$ is given by

$$u^*(kT) = \frac{1}{\omega_s} \int_{-\omega_1}^{\omega_1} \left[H_z(\omega - \omega_0) (X(\omega) - i Y(\omega)) \right] e^{i\omega kT} d\omega \quad (8.3-9)$$

upon changing variables $\omega \leftarrow -\omega$ in equation (8.3-7)

$$f(kT) = \frac{1}{\omega_s} \int_{-\omega_1}^{\omega_1} \left[\frac{H_z(\omega + \omega_0) + H_z(\omega - \omega_0)}{2} X(\omega) + i \frac{H_z(\omega + \omega_0) - H_z(\omega - \omega_0)}{2} Y(\omega) \right] e^{i\omega kT} d\omega. \quad (8.3-10a)$$

and

$$g(kT) = \frac{1}{\omega_s} \int_{-\omega_1}^{\omega_1} \left[-i \frac{H_z(\omega + \omega_0) - H_z(\omega - \omega_0)}{2} X(\omega) + \frac{H_z(\omega + \omega_0) + H_z(\omega - \omega_0)}{2} Y(\omega) \right] e^{i\omega kT} d\omega. \quad (8.3-10b)$$

Obviously, the transforms for which a search has been made are now available: the transforms of $f(kT)$ and $g(kT)$ are

$$F(\omega) = \frac{H_z(\omega + \omega_0) + H_z(\omega - \omega_0)}{2} X(\omega) + i \frac{H_z(\omega + \omega_0) - H_z(\omega - \omega_0)}{2} Y(\omega),$$

and

$$G(\omega) = -i \frac{H_z(\omega + \omega_0) - H_z(\omega - \omega_0)}{2} X(\omega) + \frac{H_z(\omega + \omega_0) + H_z(\omega - \omega_0)}{2} Y(\omega). \quad (8.3-11)$$

respectively. A check for the requisite property: $F^*(-\omega) = F(\omega)$, $G^*(-\omega) = G(\omega)$, demonstrates that $f(kT)$ and $g(kT)$ are indeed, real and that $F(\omega)$ and $G(\omega)$ are, indeed, the required transforms.

These expressions are of considerable interest. They indicate that the digital filtering of a narrow-band high frequency signal can be performed on the low frequency complex modulation components $x(kT)$ and $y(kT)$, if they are available, rather than at the carrier frequency. Of course, if the total output at the high frequency is desired; then either filtering at the high frequency ω_0 is necessary, or low frequency filtering can be performed with remodulation back to the frequency ω_0 . Another, somewhat obvious, utility of the preceding discussion is in digital filtering of the sampled complex modulation derived from a high frequency analog signal.

The two filters that produce the complex modulation output components from the quadrature input components, $x(kT)$ and $y(kT)$, can be defined as

$$H_z^{(1)}(\omega) = \frac{H_z(\omega + \omega_0) + H_z(\omega - \omega_0)}{2}, |\omega| < \omega_1$$

and

(8.3-12)

$$H_z^{(2)}(\omega) = -i \frac{H_z(\omega + \omega_0) - H_z(\omega - \omega_0)}{2}, |\omega| < \omega_1$$

As $H_z(\omega)$ is a "causal" or realizable filter, the impulse response doesn't occur before the impulse, it can be shown that both $H_z^{(1)}(\omega)$ and $H_z^{(2)}(\omega)$ are causal filters. Furthermore, the impulse responses of both $H_z^{(1)}(\omega)$ and $H_z^{(2)}(\omega)$ are real, as $H_z(-\omega) = H_z^*(\omega)$ implies that they both possess this same property.

In terms of $H_z^{(1)}(\omega)$ and $H_z^{(2)}(\omega)$

$$F(\omega) = H_z^{(1)}(\omega) X(\omega) - H_z^{(2)}(\omega) Y(\omega), \quad (8.3-13)$$

and

$$G(\omega) = H_z^{(2)}(\omega) X(\omega) + H_z^{(1)}(\omega) Y(\omega).$$

Another interesting form of these relations can be written:

$$F(\omega) + i G(\omega) = \left(H_z^{(1)}(\omega) + i H_z^{(2)}(\omega) \right) \left(X(\omega) + i Y(\omega) \right) \quad (8.3-14)$$

Equating powers of unity and $i = \sqrt{-1}$ produces the preceding relations.

The "complex" digital filter of Section (8.1) is obviously

$$H_z^C(z) = H_z^{(1)}(z) + i H_z^{(2)}(z) \quad (8.3-15)$$

In Figure 8.3-1 a block diagram form of the filtering operation can be seen

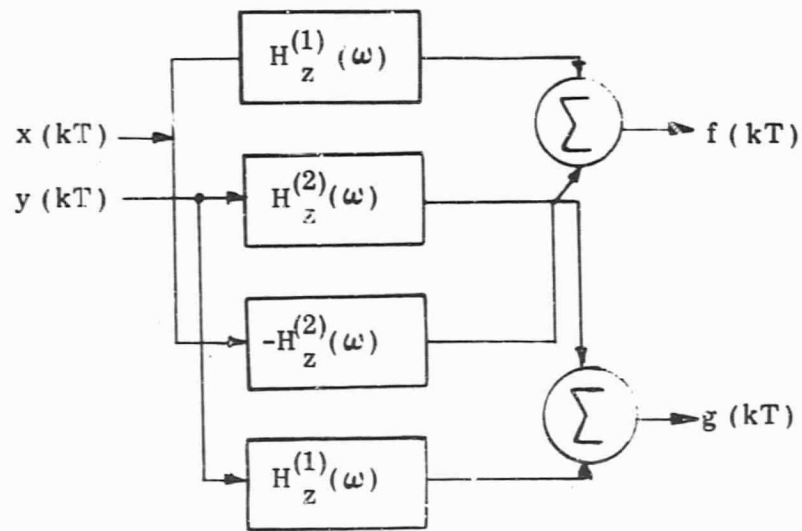


Figure 8.3-1. Implementation of a High Frequency Filter Using Only Low Frequency Digital Narrow-band Filters

An obvious generalization of the simple narrow-band high frequency filter, when the modulating components are given, occurs when it is realized that a linear process can have four independent filters that are different rather than the particular case above. Namely, the most general linear operation is

$$F(\omega) = H_z^{(11)}(\omega) X(\omega) + H_z^{(12)}(\omega) Y(\omega) \quad (8.3-16)$$

$$G(\omega) = H_z^{(21)}(\omega) X(\omega) + H_z^{(22)}(\omega) Y(\omega)$$

The "simple" narrow-band filter implementation occurs when the filter matrix

$$\begin{pmatrix} H_z^{(11)}(\omega) & H_z^{(12)}(\omega) \\ H_z^{(21)}(\omega) & H_z^{(22)}(\omega) \end{pmatrix} \quad (8.3-17)$$

is such that

$$\begin{aligned} H_z^{(11)}(\omega) &= H_z^{(22)}(\omega) \\ &= H_z^{(1)}(\omega) \end{aligned}$$

and

$$\begin{aligned} H_z^{(12)}(\omega) &= -H_z^{(21)}(\omega) \\ &= -H_z^{(2)}(\omega). \end{aligned} \quad (8.3-18)$$

8.3.1 Impulse Response Form of Digital Complex Modulation Filtering

Another useful form of the complex modulation digital filtering formulae expresses the filtering of the modulation inputs directly in the time domain in terms of the impulse responses of the filters $H_z^{(1)}(\omega)$ and $H_z^{(2)}(\omega)$.

By definition, the z transform of the transfer functions $H_z^{(1)}(z)$ and $H_z^{(2)}(z)$ are the impulse responses $w_j^{(1)}$ and $w_j^{(2)}$, respectively. Taking the inverse z transform of (8.3-13) and using the convolution form of the input-output relation for a digital filter:

$$f_k = \sum_{j=0}^{\infty} w_j^{(1)} x_{k-j} - \sum_{j=0}^{\infty} w_j^{(2)} y_{k-j}$$

and

(8.3.1-1)

$$g_k = \sum_{j=0}^{\infty} w_j^{(2)} x_{k-j} + \sum_{j=0}^{\infty} w_j^{(1)} y_{k-j}$$

where functional notation has been replaced by subscripted notation (e.g. $x_k = x(kT)$) for obvious convenience.

Using complex notation the two equations (8.3.1-1) can be written in as

$$u_k = \sum_{j=0}^{\infty} w_j v_{k-j} \quad (8.3.1-2)$$

where

$$w_j = w_j^{(1)} + i w_j^{(2)}$$

$$u_k = f_k + i g_k$$

and

$$v_k = x_k + i y_k$$

as in Section 8.1.

From Section 8.2 it can also be seen that the impulse response of the original narrow-band filter $H_z(\omega)$ can be written as

$$h_j = w_j^{(1)} \cos(\omega_0 jT) - w_j^{(2)} \sin(\omega_0 jT) \quad (8.3.1-3)$$

as $w_j^{(1)}$ and $w_j^{(2)}$ are obviously the low frequency modulations of the impulse response. This form is quite useful in finding recursive implementations $H_z^{(1)}(z)$ and $H_z^{(2)}(z)$ of the complex low frequency equivalents of $H_z(\omega)$ if it is also recursive.

8.3.2 Symmetrical Filter Simplifications

A particularly useful simplification, which is approximately the situation in many applications is that of the symmetrical bandpass filter. As one would expect, a symmetrical filter is defined as being approximately a mirror image in magnitude about the center frequency ω_0 , an even function about ω_0 , and the phase is an odd function about ω_0 .

Mathematically, a symmetrical bandpass filter is defined by

$$H_z(\omega + \omega_0) = H_z^*(\omega_0 - \omega), \quad \omega < \omega_1 < \omega_0 \quad (8.3.2-1a)$$

or as

$$H_z(-\omega) = H_z^*(\omega)$$

the equivalent definition

$$H_z(\omega + \omega_0) = H_z(\omega - \omega_0) \quad (8.3.2-1b)$$

is available.

This immediately implies that the two low frequency complex filters

$$H_z^{(1)}(\omega) = \frac{H_z(\omega + \omega_0) + H_z(\omega - \omega_0)}{2}, \quad (8.3.2-2a)$$

$$= H_z(\omega + \omega_0);$$

and

$$H_z^{(2)}(\omega) = -i \frac{H(\omega + \omega_0) - H(\omega - \omega_0)}{2} \quad (8.3.2-2b)$$

$$= 0.$$

Obviously, the complex modulation filtering operations are simplified as only two identical filters are necessary in this case:

$$F(\omega) = H_z^{(1)}(\omega) X(\omega)$$

and

(8.3.2-3)

$$G(\omega) = H_z^{(1)}(\omega) Y(\omega).$$

Also it can be seen that the impulse response of $H_z(\omega)$ is (equation 8.3.1-3)

$$h_j = w_j^{(1)} \cos (\omega_0 jT), \quad (8.3.2-4)$$

simply an amplitude modulated (casual) carrier.

9. DIGITAL FILTERING OF THE "COMPLED MODULATION" OF ANALOG SIGNALS

In Section 4 a linear, constant coefficient analog system was represented in terms of a transfer function and then used along with a bandlimited signal to express the input-output relationship in terms of samples of the input and output waveforms. The avenues for numerical integration and simulation by digital techniques were thereby opened. In this section the developments are extended to narrow-band analog systems in terms of the modulating components so that numerical integration and simulation may be performed at the modulation component bandwidth rather than at the usually high frequency carrier.

9.1 SAMPLED REPRESENTATION OF ANALOG BANDLIMITED HIGH-FREQUENCY SIGNALS

In a manner parallel to a previous section (8.2) a real, analog, narrow-band, high frequency signal can be represented by

$$s(t) = x(t) \cos \omega_0 t - y(t) \sin \omega_0 t \quad (9.1-1)$$

To represent $s(t)$ in terms of samples, take its Fourier transform:

$$S(\omega) = \int_{-\infty}^{\infty} (x(t) \cos \omega_0 t - y(t) \sin \omega_0 t) e^{-i\omega t} dt \quad (9.1-2)$$

with quantities defined identically as in Section 8.2 except Fourier transforms are used. Using the relations,

$$\cos \omega_0 t = \frac{e^{i\omega_0 t} + e^{-i\omega_0 t}}{2}$$

and

$$\sin \omega_0 t = \frac{e^{i\omega_0 t} - e^{-i\omega_0 t}}{2i}$$

(9.1-3)

$S(\omega)$ can be written

$$S(\omega) = \frac{1}{2} \left[X(\omega - \omega_0) + X(\omega + \omega_0) \right] + \frac{j}{2} \left[Y(\omega - \omega_0) - Y(\omega + \omega_0) \right] \quad (9.1-4)$$

where $X(\omega)$ and $Y(\omega)$ are the Fourier transforms, this time, of $x(t)$ and $y(t)$, respectively.

Now $x(t)$ and $y(t)$ are the "modulating" signals and are "low" frequency waveforms. This follows from the restriction that $S(\omega)$ is a high frequency narrow-band spectrum. Assuming that the bandlimits of $S(\omega)$ are, for positive ω , $\omega_0 - \omega_s/2$ and $\omega_0 + \omega_s/2$ then the signal bandwidth is ω_s . As $x(t)$ and $y(t)$ are real numbers the transforms $X(\omega)$ and $Y(\omega)$, respectively, are completely defined by their values for positive ω as

$$X(-\omega) = X^*(\omega)$$

and

$$Y(-\omega) = Y^*(\omega).$$

(9.1-5)

Both $x(t)$ and $y(t)$ are bandlimited by $\omega_s/2$.

Therefore, both $x(t)$ and $y(t)$ may be represented in terms of their samples taken at intervals, $T = 2\pi/\omega_s$ as in Section 2.1.1.

$$x(t) = \sum_{k=-\infty}^{\infty} x(kT) \frac{\sin\left(\frac{\omega_s}{2}(t-kT)\right)}{\left(\frac{\omega_s}{2}(t-kT)\right)} \quad (9.1-6a)$$

and

$$y(t) = \sum_{k=-\infty}^{\infty} y(kT) \frac{\sin\left(\frac{\omega_s}{2}(t-kT)\right)}{\left(\frac{\omega_s}{2}(t-kT)\right)} \quad (9.1-6b)$$

Their spectrums are given by:

$$X(\omega) = \begin{cases} T \sum_{k=-\infty}^{\infty} x(kT) e^{-j\omega kT}, & |\omega| \leq \frac{\omega_s}{2} \\ 0, & \text{otherwise;} \end{cases} \quad (9.1-7a)$$

and

$$Y(\omega) = \begin{cases} T \sum_{k=-\infty}^{\infty} y(kT) e^{-i\omega kT}, & |\omega| \leq \frac{\omega_s}{2} \\ 0, & \text{otherwise.} \end{cases} \quad (9.1-7b)$$

In terms of the sampled modulating signals the high frequency narrow-band waveform (9.1-1) is

$$s(t) = \sum_{k=-\infty}^{\infty} \left[x(kT) \cos \omega_0 t - y(kT) \sin \omega_0 t \right] \frac{\left(\sin \frac{\omega_s}{2} (t-kT) \right)}{\left(\frac{\omega_s}{2} (t-kT) \right)} \quad (9.1-8)$$

It should be noted that the signal bandwidth for high frequency, narrow-band waveforms can be as high as the sampling frequency, $\omega_s = 2\pi/T$; which is apparently different than that for low frequency signals which cannot have a bandwidth greater than one-half the sampling frequency. However, high frequency, narrow-band waveforms required two values at a given sampling instant: $x(kT)$ and $y(kT)$, the "complex modulation" components of the signal each restricted to a bandwidth of half the sampling frequency.

The spectrum of the signal $S(\omega)$ can be written in terms of the sampled complex modulation as:

$$S(\omega) = \frac{T}{2} \left\{ \sum_{k=-\infty}^{\infty} (x_k + iy_k) e^{-i(\omega - \omega_0)kT}: \left| \omega - \omega_0 \right| \leq \frac{\omega_s}{2} \right. \\ \left. \sum_{k=-\infty}^{\infty} (x_k - iy_k) e^{-i(\omega + \omega_0)kT}: \left| \omega + \omega_0 \right| \leq \frac{\omega_s}{2} \right\} \quad (9.1-9)$$

9.2 NUMERICAL INTEGRATION WITH NARROW-BAND INPUT SPECTRA BY "COMPLEX" DIGITAL FILTERS

Numerical integration and simulation were discussed extensively in a previous section (Section 4) where "low" frequency spectrums were apparently of primary interest. However, in practice, analog signals are found that have "high" frequency spectra which often produce difficult filtering problems. If these signals are narrow-band and decomposed into complex demodulates then they can be filtered either in an analog or digital manner at much lower frequencies and sampling rates. Numerical integration and/or simulation is equivalently performed using the complex modulation.

Developments here follow those of Section 4 which should be consulted if necessary.

Linear filtering of $s(t)$ to produce an output $r(t)$ is expressed in terms of Fourier transforms by the equation

$$R(\omega) = H(\omega) S(\omega) \quad (9.2-1)$$

where

$S(\omega)$ = transform of the input, $s(t)$.

$R(\omega)$ = transform of the output, $r(t)$.

$H(\omega)$ = filter transfer or system function

Upon using the bandlimited representation of $S(\omega)$ in terms of samples:

$$R(\omega) = \frac{T}{2} \left\{ \sum_{k=-\infty}^{\infty} (x_k + iy_k) e^{i\omega_0 kT} H(\omega) e^{-i\omega kT}; \left| \omega - \omega_0 \right| \leq \frac{\omega_s}{2} \right. \\ \left. \sum_{k=-\infty}^{\infty} (x_k - iy_k) e^{-i\omega_0 kT} H(\omega) e^{-\omega kT}; \left| \omega + \omega_0 \right| \leq \frac{\omega_s}{2} \right\} \quad (9.2-2)$$

and taking the inverse Fourier transform, the relation

$$r(t) = \frac{1}{2} \sum_{k=-\infty}^{\infty} (x_k + iy_k) e^{i\omega_0 kT} \frac{T}{2\pi} \int_{\omega_0 - \frac{\omega_s}{2}}^{\omega_0 + \frac{\omega_s}{2}} H(\omega) e^{i\omega(t-kT)} d\omega$$

(9.2-3)

$$+ \frac{1}{2} \sum_{k=-\infty}^{\infty} (x_k - iy_k) e^{-i\omega_0 kT} \frac{T}{2\pi} \int_{-\omega_0 - \frac{\omega_s}{2}}^{-\omega_0 + \frac{\omega_s}{2}} H(\omega) e^{i\omega(t-kT)} d\omega$$

is obtained. The two integrals can be put into another form by a substitution of variables:

$$\frac{1}{\omega_s} \int_{\omega_0 - \frac{\omega_s}{2}}^{\omega_0 + \frac{\omega_s}{2}} H(\omega) e^{i\omega(t-kT)} d\omega = e^{i\omega_0(t-kT)} \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H(\omega + \omega_0) e^{i\omega(t-kT)} d\omega, \quad \omega_s = \frac{2\pi}{T};$$

and

$$\frac{1}{\omega_s} \int_{-\omega_0 - \frac{\omega_s}{2}}^{-\omega_0 + \frac{\omega_s}{2}} H(\omega) e^{i\omega(t-kT)} d\omega = e^{-i\omega_0(t-kT)} \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H^*(\omega + \omega_0) e^{-i\omega(t-kT)} d\omega.$$

(9.2-4)

Letting $t = nT$, a sampling instant, and defining

$$w_j^c = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H(\omega + \omega_0) e^{i\omega_j T} d\omega$$

(9.2-5)

the sampled output can be written as

$$\begin{aligned}
 r(nT) &= f(nT) \cos \omega_0 nT - g(nT) \sin \omega_0 nT \\
 &= f_n \cos \omega_0 nT - g_n \sin \omega_0 nT \\
 &= \frac{1}{2} \sum_{k=-\infty}^{\infty} (x_k + iy_k) w_{n-k}^c e^{i\omega_0 nT} \\
 &\quad + \frac{1}{2} \sum_{k=-\infty}^{\infty} (x_k - iy_k) w_{n-k}^{c*} e^{-i\omega_0 nT}
 \end{aligned} \tag{9.2-6}$$

Writing the output $r(nT)$ in complex form

$$\begin{aligned}
 r(nT) &= f_n \cos \omega_0 nT - g_n \sin \omega_0 nT \\
 &= \frac{1}{2} (f_n + ig_n) e^{i\omega_0 nT} + \frac{1}{2} (f_n - ig_n) e^{-i\omega_0 nT}
 \end{aligned} \tag{9.2-7}$$

it can be seen that

$$f_n + ig_n = \sum_{k=-\infty}^{\infty} w_{n-k}^c (x_k + iy_k) \tag{9.2-8}$$

which allows low frequency numerical integration or simulation as long as the carrier frequency and phase are known. In general the w_j^c 's are complex numbers

$$w_j^c = w_j^R + i w_j^I \tag{9.2-9}$$

so that a filtering operation involves both x_k and y_k to produce f_n (or g_n). To see this the multiplication in equation (9.2-8) must be carried through and the real and imaginary parts equated resulting in

$$f_n = \sum_{k=-\infty}^{\infty} (w_{n-k}^R x_k - w_{n-k}^I y_k),$$

and

$$g_n = \sum_{k=-\infty}^{\infty} (w_{n-k}^I x_k + w_{n-k}^R y_k)$$

(9.2-10)

Two perfectly "real" computing expressions are obtained for the "complex" digital filtering of a bandlimited high frequency signal in terms of the sampled components x_k and y_k in a manner similar to Section 4.5. However, this expression is not realizable as future inputs are required. The w_j^c are generally not zero for $j < 0$ as was the case for low frequency digital filters (Section 4). With delay the filtering operation may be approximated.

By employing the methods of Section 4.5.2 a realizable complex linear digital filtering operation can be obtained. Assuming w_j^c is practically zero for some $j < -M$; which is true for stable, constant coefficient, analog filters; and redefining the time sequence

$$\begin{aligned} f'_{n+M} + ig'_{n+M} &\equiv f'_n + ig_n \\ &\equiv u'_{n+M} \end{aligned} \quad (9.2-11)$$

then

$$u'_n = \sum_{k=0}^{\infty} w_{k-M}^c v_{n-k} \quad (9.2-12)$$

with

$$v_k = x_k + iy_k$$

as usual. A realizable digital filtering expression has been obtained! Delay of MT seconds is obviously still present by the virtue of the redefinition the output sequence.

10. SYNTHESIS OF "COMPLEX" DIGITAL FILTERS FROM NARROW-BAND ANALOG FILTERS

Machinery has been developed to this point where narrow-band "high" frequency signals can theoretically be filtered in terms of their low frequency complex demodulates. This section is devoted to the synthesis of the low frequency "complex" digital filter from a given narrow-band filter in one of two forms:

1. An arbitrary transfer function given in numerical form
2. A constant coefficient analog system as the ratio of two polynomials in "s".

Most of the development is a simple application of synthesis techniques for "real" filters.

The first form will be used to develop non-recursive "complex" filters and be described in barest essentials. It will be recognized that all the theory and techniques of Section 5 are immediately applicable and a brief sketch should be adequate.

Recursive digital "complex" filters can be developed through the use of z transform techniques if a constant coefficient narrow-band analog system is given. Once the necessary forms for the "complex" filters are obtained in Section 10.2, Sections 6 and 7 should be consulted for proper conversion to digital filters.

It should also be pointed out that the method to be discussed in Section 10.2, with obvious modifications, is applicable in producing digital complex filters from narrow-band digital filters.

Synthesis limitations are fundamentally the same as those for "real" digital filters. Non-recursive filters generally require more delay but almost any transfer function can be synthesized to half the sampling frequency. Recursive filters usually require an analog prototype and often are difficult to realize at more than one-third the sampling frequency unless sufficiently bandlimited. Delay with recursive filters is often less than that for non-recursive ones.

10.1 SYNTHESIS OF NON-RECURSIVE "COMPLEX" FILTERS

Computation of the non-recursive complex filter weights is relatively simple and the following is, to some extent, a recapitulation of the work in Section 9.2. Using equation 9.2-5 for the "complex" filter in terms of the narrow-band filter transfer function, $H(\omega)$, allows calculation of the real and imaginary components of the w_j 's.

With these two functions all the methods of Section 5 are immediately applicable along with convergence techniques to minimize the number of w_j 's.

Equation 9.2-5 for the narrow-band filter can be written in terms of the constituent real filters as

$$H(\omega + \omega_0) = H_1(\omega) + i H_2(\omega), \quad |\omega| \leq \omega_s/2;$$

where

$$H_1(\omega) = \frac{H(\omega + \omega_0) + H(\omega - \omega_0)}{2}, \quad |\omega| \leq \omega_s/2; \quad (10.1-1)$$

and

$$H_2(\omega) = -i \frac{H(\omega + \omega_0) - H(\omega - \omega_0)}{2}, \quad |\omega| \leq \omega_s/2.$$

in analogy with equation (8.3-12).

The real and imaginary parts of the non-recursive filter weights w_j are given by

$$w_j^R = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H_1(\omega) e^{i\omega(j-\tau)T} d\omega,$$

and

(10.1-2)

$$w_j^I = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H_2(\omega) e^{i\omega(j-\tau)T} d\omega.$$

With both w_j^R and w_j^I real numbers, and τ the delay, in sampling units necessary to realize the approximation. Numerical techniques of Section 5 should, of course, be used for evaluation of the integrals. Only enough coefficients should be retained for the desired approximation accuracy of the analog transfer functions. The delay, τT , should be determined by the methods of Section 5.3.1. Machine programs used for low frequency filters can be used for computation of the w_j^R and w_j^I .

The filtering operations are obviously

$$f_n = \sum_{k=0}^M (w_k^R x_{n-k} - w_k^I y_{n-k}),$$

and

(10.1-3)

$$g_n = \sum_{k=0}^M (w_k^I x_{n-k} + w_k^R y_{n-k}),$$

where convergence factors may be incorporated in the w_j^I as in Section 5.7.

10.2 SYNTHESIS OF RECURSIVE COMPLEX DIGITAL FILTERS FROM AN ANALOG BANDPASS PROTOTYPE IN RATIONAL FORM

Given an analog prototype bandpass filter, $H(s)$, as the ratio of two polynomials in s that is realizable it is possible to apply all the methods of Section 7 for low pass digital filters to obtain recursive complex digital filters which are also low pass filters. Understanding of that section is necessary for comprehensive design; however as the bandpass filters are usually approximately bandlimited by nature of their utility, straightforward z transformation by the method to be described is usually adequate. Sample and hold and low pass reconstruction methods are also applicable if necessary (Section 7).

Expansion of $H(s)$ in partial functions as in Section 6.2 gives

$$H(s) = \sum_{j=1}^L \sum_{\mu=1}^{\mu(j)} \frac{A_j^{\mu}}{(s+\beta_j)^{\mu}} \quad (10.2-1)$$

with

$$A_j^{\mu(j)-m} = \frac{1}{m!} \frac{d^m}{ds^m} \left[H(s) (s+\beta_j)^{\mu(j)} \right]_{s = -\beta_j} \quad (10.2-2)$$

No constant term is present in the expansion as $H(s)$ is a bandpass function.

The partial impulse response corresponding to a particular term, found by taking the inverse transform, is

$$\mathcal{L}^{-1} \left[\frac{A_j^{\mu}}{(s+\beta_j)^{\mu}} \right] = \begin{cases} \frac{A_j^{\mu}}{(\mu-1)!} t^{\mu-1} e^{-\beta_j t}, & t > 0; \\ 0, & t < 0. \end{cases} \quad (10.2-3)$$

Now for bandpass filters the roots of the transfer function generally occur in complex conjugate pairs, in which case it is also true that the expansion coefficients for the conjugate poles are conjugates. That is, corresponding to a root β_j a conjugate, β_j^* , must exist if β_j is complex. Expansion coefficients of β_j^* , $A_{\beta_j^*}^\mu$ are given by

$$A_{\beta_j^*}^\mu = A_{\beta_j}^{\mu*} \quad (10.2-4)$$

Therefore, consider the partial impulse response of a pair of conjugate poles:

$$\mathcal{L}^{-1} \left[\frac{A_j^\mu}{(s+\beta_j)^\mu} + \frac{A_j^{\mu*}}{(s+\beta_j^*)^\mu} \right] = \begin{cases} \frac{t^{\mu-1}}{(\mu-1)!} \left[A_j^\mu e^{-\beta_j t} + A_j^{\mu*} e^{-\beta_j^* t} \right], & t > 0; \\ 0, & t < 0; \end{cases} \quad (10.2-5)$$

$$= h_{\beta_j}^\mu(t), \text{ the partial impulse response.}$$

Writing the expansion coefficient and the root in complex form

$$\begin{aligned} A_j^\mu &= a_j^\mu + i b_j^\mu, \\ \beta_j &= \alpha_j - i \delta_j. \end{aligned} \quad (10.2-6)$$

then

$$h_{\beta_j}^\mu(t) = \begin{cases} \frac{2 t^{\mu-1} e^{-\alpha_j t}}{(\mu-1)!} \left[a_j^\mu \cos \delta_j t - b_j^\mu \sin \delta_j t \right], & t > 0; \\ 0, & t < 0. \end{cases} \quad (10.2-7)$$

Now the "center" frequency ω_0 , of the bandpass filter is somewhat ill-defined, but any frequency somewhere in the "middle" of the passband will do as long as this frequency is also used when demodulating the bandpass signal that is to be filtered by the digital filter. Defining the difference between the center frequency and the root imaginary part as

$$\gamma_j = \delta_j - \omega_0 \quad (10.2-8)$$

then the partial impulse response may be written after some manipulation as

$$h_{\beta_j}^{\mu}(t) = \frac{2t^{\mu-1} e^{-\alpha_j t}}{(\mu-1)!} \left\{ (a_j^{\mu} \cos \gamma_j t - b_j^{\mu} \sin \gamma_j t) \cos \omega_0 t - (a_j^{\mu} \sin \gamma_j t + b_j^{\mu} \cos \gamma_j t) \sin \omega_0 t \right\}, t > 0. \quad (10.2-9)$$

This equation is of the form

$$h_{\beta_j}^{\mu}(t) = x_j^{\mu}(t) \cos \omega_0 t - y_j^{\mu}(t) \sin \omega_0 t \quad (10.2-10)$$

which is true for all complex roots with obvious definitions for $x_j^{\mu}(t)$ and $y_j^{\mu}(t)$. This partial impulse response is in the narrow-band form of Section 8.2 even though it is in analog form. Therefore, that the total impulse response is

$$h(t) = \mathcal{L}^{-1} H(s) = \begin{cases} w_1(t) \cos \omega_0 t - w_2(t) \sin \omega_0 t, & t > 0; \\ 0, & t < 0; \end{cases} \quad (10.2-11)$$

where

$$w_1(t) = \sum_{j=1}^L \sum_{\mu=1}^{\nu(j)} x_j^{\mu}(t), \quad (10.2-12a)$$

and

$$w_2(t) = \sum_{j=1}^L \sum_{\mu=1}^{\nu(j)} y_j^{\mu}(t). \quad (10.2-12b)$$

At this point the narrow-band analog impulse response has been obtained.

If $w_1(t)$ and $w_2(t)$ are bandlimited by half the sampling frequency then direct z transformation (Section 6.2) will produce the digital filter equivalents within a factor of T . When $w_1(t)$ and $w_2(t)$ are not bandlimited by half the sampling frequency then the techniques of Section 7 must be used if recursive filters are desired.

To simplify computation of the z transform, the impulse responses can be manipulated into an appropriate form which, coincidentally, produces the analog complex filters. This form is identical to that for low frequency filters which means that the same techniques, or machine programs, that are used for low frequency filters can be immediately applied to complex filters.

The quantities $x_j^{\mu}(t)$ and $y_j^{\mu}(t)$ can be recast, using the original partial fraction expansion coefficients, into the form

$$x_j^{\mu}(t) = \frac{t^{\mu-1}}{(\mu-1)!} \left[A_j^{\mu} e^{-\bar{\beta}_j t} + A_j^{\mu*} e^{-\bar{\beta}_j^* t} \right], \quad t > 0; \quad (10.2-13)$$

and

$$y_j^{\mu}(t) = \frac{t^{\mu-1}}{(\mu-1)!} \left[-i A_j^{\mu} e^{-\bar{\beta}_j t} + i A_j^{\mu*} e^{-\bar{\beta}_j^* t} \right], \quad t > 0;$$

where

$$\begin{aligned} \bar{\beta}_j &= \beta_j + i\omega_0, \\ &= \alpha_j - i(\delta - \omega_0). \end{aligned}$$

It can be seen that the Laplace transform of $w_1(t)$, $H_1(s)$ is given by

$$H_1(s) = \sum_{j=1}^L \sum_{\mu=1}^{V(j)} \frac{A_j^\mu}{(s+\beta_j)^\mu} ; \quad (10.2-14)$$

that is, $H_1(s)$ is obtained by simply subtracting ω_0 from the absolute value of the imaginary part of the roots of $H(s)$. The resultant imaginary part may be positive or negative.

Similarly, the transform of $w_2(t)$, $H_2(s)$ is given by

$$H_2(s) = \sum_{j=1}^L \sum_{\mu=1}^{V(j)} \frac{\bar{A}_j^\mu}{(s+\bar{\beta}_j)^\mu} , \quad (10.2-15)$$

where

$$\bar{A}_j^\mu = -i A_j^\mu \quad \text{for a root } \beta_j \text{ with a negative imaginary part:}$$

and

$$\bar{A}_j^\mu = i A_j^\mu \quad \text{for a root } \beta_j \text{ with a positive imaginary part.}$$

These forms for $H_1(s)$ and $H_2(s)$ are obviously identical (not equal) and are those requisite for z transformation, with the modifications of Section 7 to produce the digital filter:

$$H_z(z) = H_z^{(1)}(z) + i H_z^{(2)}(z)$$

$$H_z^{(1)}(z) = \mathcal{Z}(H_1(s)),$$

and

$$H_z^{(2)}(z) = \mathcal{Z}(H_2(s))$$

for complex digital filtering via Sections 8 and 9.

quite simple:

Procedurally, the steps for producing the two component filters are

To produce $H_z^{(1)}(z)$:

1. Reduce the imaginary part of the roots of the partial fraction expansion of the bandpass transfer function $H(s)$ by ω_0 , the "center" frequency used when complex demodulating.
2. z transform the result of step 1 via Section 6.4.
 $H_z^{(1)}(z)$ is the result of the z transformation.

To produce $H_z^{(2)}(z)$

1. Reduce the imaginary part of the roots of the partial fraction expansion of the bandpass transfer function $H(s)$ by ω_0 , the "center" frequency used when complex demodulating.
2. Multiply the partial fraction expansion coefficients (for each multiplicity) by
 - (a) $-i$ for roots with negative imaginary parts
 - (b) i for roots with positive imaginary parts
3. z transform the result of steps 1 and 2 via Section 6.4. $H_z^{(2)}(z)$ is the result of the z transformation.

11. STATISTICAL CONCEPTS

Various statistical aspects enter into the filtering of data. Not only is rejection of clear-cut disturbance in the data (60~ hum) desired, or emphasis of a certain feature wanted but quite often filtering "in" or "out" a "random" signal is desired. It is usually desired to "filter out" noise and "filter in" speech. Hopefully, in some meaningful manner, "random" signals can be characterized. Receiver noise is usually a "random" process in that it is not a deterministic process whose values can be predicted. Similarly speech waveforms may be considered "random" even though they carry information. Speech, unless repeated, appears upon visual or mathematical examination of a record to have very little "correlation" from instant to instant. Statistical analysis often points up the regularity of speech in a "statistical" sense.

A mixture of probability theory, statistics, and communication theory has been found to be exceedingly useful in the description of "random" processes found in everyday life. In fact, the combination of these subjects encompasses much more ground than either of them by themselves.

Covered in the ensuing section are the basic formulae of engineering utility and their development with an apology for brevity and lack of explanation. The material is widely known and covered thoroughly and well in many excellent texts. They are presented for the purposes of utility in further sections and to make the notation used clear and more coherent. Continuous processes are discussed first as they are more easily understood.

11.1 BASIC DESCRIPTION OF "RANDOM" PROCESSES

11.1.1 Probability Densities

Suppose a function $x^{(k)}(t)$ is given which is a member, the k^{th} , of an ensemble of functions; for example, the output noise of an ensemble of identical receivers. Assume that we have N of these member functions ($k = 1, 2, \dots, N$). At a given time t find the number of functions, say $m(N)$, that lie between the limits $x_1(t)$ and $x_2(t)$ and define the probability of $x(t)$ lying between these limits as

$$P(x_1(t) \leq x(t) \leq x_2(t)) = \lim_{N \rightarrow \infty} \frac{m(N)}{N} \quad (11.1.1-1)$$

Further define the probability density $p(x, t)$ of $x(t)$ at time t by the integral

$$P(-\infty \leq x(t) \leq x(t)) = \int_{-\infty}^{x(t)} p(x, t) dx \quad (11.1.1-2)$$

Obviously,

$$P(-\infty \leq x \leq \infty) = \int_{-\infty}^{\infty} p(x, t) dx = 1. \quad (11.1.1-3)$$

The probability that the process has some value is unity. It is immediately seen that

$$\begin{aligned} P(x(t) \leq x(t) \leq x(t+\Delta)) &= P(-\infty \leq x(t) \leq x(t) + \Delta) - P(-\infty \leq x(t) \leq x(t)) \\ &= \int_{x(t)}^{x(t) + \Delta} p(x, t) dx \end{aligned} \quad (11.1.1-4)$$

For small Δ it is seen that, approximately,

$$P(-\infty \leq x(t) \leq x(t) + \Delta) - P(-\infty \leq x(t) \leq x(t)) \simeq p(x, t) \Delta \quad (11.1.1-5)$$

Taking the limit as Δ approaches zero it is seen that

$$\frac{\partial P}{\partial x} = p(x, t) \quad (11.1.1-6)$$

It is up to the analyst or experimentalist to discover, or approximate, the probability or probability density for a given process. Intuitively, the probability of a function taking on a value between limits defined in this manner is natural. If physical reality is repeatable, in an "average" sense, then a "random" phenomena can be expected to have a certain definite probability of occurring, which is taken as the ratio of the occurrences of the event to the total number of trials as the number of trials approaches infinity. The probability density $p(x, t)$ is a natural generalization to include the idea of the distribution of the values of a variable $x(t)$ (t fixed) which can take on values continuously. To reiterate $x(t)$ is not being spoken of as a function of t but as a member of an ensemble $x^{(k)}(t)$ ($k = 1, 2, \dots, N$) of functions of fixed t . The notation $p(x, t)$ is designed to include the fact that the distribution of members of the ensemble $x^{(k)}(t)$ depends upon the time t at which the probability is evaluated.

Higher order probability densities can be defined. The most useful of these is the probability of $x(t)$ lying between x_1 and x_2 at time t_1 and between x_1 and x_2 at time t_2 :

$$P(x_1 \leq x(t_1) \leq x_2; x_1' \leq x(t_2) \leq x_2') = \int_{x_1}^{x_2} \int_{x_1'}^{x_2'} p(x, t_1; x', t_2) dx dx' \quad (11.1.1-7)$$

The original probability density may be obtained from this second order density.

The probability that $x(t)$ lies between two limits: $x_1(t) \leq x(t) \leq x_2(t)$, must be equal to the probability that

$$P(x_1 \leq x(t) \leq x_2; -\infty \leq x(t_2) \leq \infty) = \int_{x_1}^{x_2} \int_{-\infty}^{\infty} p(x, t; x', t_2) dx dx' \quad (11.1.1-8)$$

$$= P(x_1 \leq x(t) \leq x_2)$$

Namely, the probability of $x(t)$ lying between the limits x_1 and x_2 at time t and taking on any value at time t_2 is just $P(x_1 \leq x(t) \leq x_2)$. These statements imply the relation among density functions:

$$p(x, t) = \int_{-\infty}^{\infty} p(x, t; x', t) dx'. \quad (11.1.1-9)$$

Similarly, higher order density functions may be defined, but for present purposes the two defined will suffice.

It is also possible to define joint probabilities and densities of two (or more) random ensembles $x^{(k)}(t)$ and $y^{(k)}(t)$ at the same, or different, times. The joint probability function is

$$P(x_1(t_1) \leq x(t_1) \leq x_2(t_1); y_1(t_2) \leq y(t_2) \leq y_2(t_2)) = \int_{x_1(t_1)}^{x_2(t_1)} \int_{y_1(t_2)}^{y_2(t_2)} p(x, t_1; y, t_2) dx dy; \quad (11.1.1-10)$$

the joint density is given by

$$\frac{\partial^2 P}{\partial x \partial y} = p(x, t_1; y, t_2) \quad (11.1.1-11)$$

11.1.2 Statistical Independence

If two processes are statistically independent it is meant that

$$p(x, t_1; y, t_2) = f(x, t_1) g(y, t_2) \quad (11.1.2-1)$$

11.1.3 Averages

The average over the ensemble $x^{(k)}(t)$ at a fixed time t of some function of $x^{(k)}(t)$, $f(x^{(k)}(t))$ is defined in an expected manner. Take the number of times a given value $f(x^{(k)}(t))$ is found out of N trials, say, m_k ($m_1 + m_2 + \dots + m_r = N$) and form

$$m_k f(x^{(k)}(t)) \quad (11.1.3-1)$$

Add all possible values found, divide by the total number of trials, and define the "ensemble" average (denoted by $\langle \dots \rangle$)

$$\langle f^{(k)}(t) \rangle = \lim_{N \rightarrow \infty} \sum_{k=1}^{r(N)} \frac{m_k f(x^{(k)}(t))}{N} \quad (11.1.3-2)$$

which in the limit of continuous $x(t)$, fixed t , becomes

$$\langle f^{(k)}(t) \rangle = \int_{-\infty}^{\infty} f(x(t)) p(x, t) dx. \quad (11.1.3-3)$$

For the simple case

$$f^{(k)}(t) = x^{(k)}(t) \quad (11.1.3-4)$$

the ensemble average or "mean" of the random process $x(t)$, itself, at time t is obtained

$$\langle x^{(k)}(t) \rangle = \int_{-\infty}^{\infty} x(t) p(x, t) dx \quad (11.1.3-5)$$

Second order ensemble averages can be defined in a similar way. The average of $F(x^{(k)}(t), x^{(k)}(t_2))$ is given by

$$\langle F(x^{(k)}(t_1), x^{(k)}(t_2)) \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(x(t_1), x'(t_2)) p(x, t_1; x', t_2) dx dx' \quad (11.1.3-6)$$

Two special cases are the autocorrelation function

$$\langle x^{(k)}(t_1) x^{(k)}(t_2) \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(t_1) x'(t_2) p(x, t_1; x', t_2) dx dx' \quad (11.1.3-7)$$

which measures the degree of correlation between the function $x^{(k)}(t)$ at the times t_1 and t_2 ; and the mean-square of a function $x^{(k)}(t)$

$$\langle (x^{(k)}(t))^2 \rangle = \langle x^{(k)}(t) x^{(k)}(t) \rangle \quad (11.1.3-8)$$

or

$$\begin{aligned} \langle (x^{(k)}(t))^2 \rangle &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(t)^2 p(x, t; x', t) dx dx' \\ &= \int_{-\infty}^{\infty} x(t)^2 p(x, t) dx \end{aligned} \quad (11.1.3-9)$$

The mean square is usually called the instantaneous average power. If $\langle x^2(t) \rangle$ is not a function of the time of evaluation t then it is called the average power:

Besides the mean, one of the most important averages of a single ensemble is the variance σ^2 , defined as

$$\begin{aligned} \sigma^2 &= \langle (x^{(k)}(t) - \langle x^{(k)}(t) \rangle)^2 \rangle \\ &= \langle x^{(k)}(t)^2 - 2x^{(k)}(t) \langle x^{(k)}(t) \rangle + \langle x^{(k)}(t) \rangle^2 \rangle \end{aligned} \quad (11.1.3-10)$$

or

$$\sigma^2 = \langle x^{(k)}(t)^2 \rangle - \langle x^{(k)}(t) \rangle^2 \quad (11.1.3-11)$$

where these facts are used: the average of $\langle x(t) \rangle x(t)$ is simply $\langle x(t) \rangle^2$ and the average of a constant is that constant, which produces $\langle \langle x(t) \rangle^2 \rangle = \langle x(t) \rangle^2$.

The significance of the variance lies in its property of measuring the "spread" of a random process about its mean. It takes the values of the difference of the random variable from the mean (squared to avoid cancellation from positive and negative differences), weights them by the probability of occurrence, and averages them. Fortunately, many processes have peaked distributions so that the degree of "randomness" of a variable is characterized by the square root of the variance, σ , commonly called the "root-mean-square" error (r. m. s error). In fact as σ approaches zero, the process becomes less and less "random" and takes on only its mean value (with the exception of values that have probability zero).

Averages of functions of two (or more) ensembles of random processes may be defined. Given the function $F(x^{(k)}(t_1), y^{(k)}(t_2))$ of the random processes $x^{(k)}(t_1)$ and $y^{(k)}(t_2)$ form the ensemble average:

$$F(x^{(k)}(t_1), y^{(k)}(t_2)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(x(t_1), y(t_2)) p(x, t_1; y, t_2) dx dy \quad (11.1.3-12)$$

An important example is

$$F(x(t_1), y(t_2)) = x(t_1) + y(t_2) \quad (11.1.3-13)$$

with

$$\begin{aligned} \langle x^{(k)}(t_1) + y^{(k)}(t_2) \rangle &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x(t_1) + y(t_2)) p(x, t_1; y, t_2) dx dy \\ &= \int_{-\infty}^{\infty} x(t_1) p(x, t_1) dx + \int_{-\infty}^{\infty} y(t_2) p(y, t_2) dy \end{aligned} \quad (11.1.3-14)$$

or finally

$$\langle x^{(k)}(t_1) + y^{(k)}(t_2) \rangle = \langle x^{(k)}(t_1) \rangle + \langle y^{(k)}(t_2) \rangle \quad (11.1.3-15)$$

It is found, as expected, that the average of a sum is the sum of the averages.

The cross-correlation of two random ensembles is defined as the average of the product

$$\langle x^{(k)}(t_1) y^{(k)}(t_2) \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(t_1) y(t_2) p(x, t_1; y, t_2) dx dy \quad (11.1.3-16)$$

If the two processes are statistically independent then, simply,

$$\langle x^{(k)}(t_1) y^{(k)}(t_2) \rangle = \langle x^{(k)}(t_1) \rangle \langle y^{(k)}(t_2) \rangle ; \quad (11.1.3-17)$$

the cross-correlation is the product of the means. If either mean is zero, for statistically independent processes, then the cross correlation is zero.

It has been shown that the ensemble average of a sum of two random processes is the sum of the averages of the processes. By use of higher order density functions it is simple to show that the ensemble average of the sum of n processes is the sum of ensemble averages for the processes:

$$\left\langle \sum_{i=1}^n x_i(t_i) \right\rangle = \sum_{i=1}^n \langle x_i(t_i) \rangle \quad (11.1.3-18)$$

Another average of interest is the average of the square of a sum of random processes

$$\begin{aligned} \left\langle \left(\sum_{i=1}^n x_i(t_i) \right)^2 \right\rangle &= \left\langle \sum_{i=1}^n x_i(t_i)^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n x_i(t_i) x_j(t_j) \right\rangle, \\ &= \sum_{i=1}^n \langle x_i(t_i)^2 \rangle + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \langle x_i(t_i) x_j(t_j) \rangle. \end{aligned} \quad (11.1.3-19)$$

If the cross-correlation between the various processes vanishes; then,

$$\left\langle \left(\sum_{i=1}^n x_i(t_i) \right)^2 \right\rangle = \sum_{i=1}^n \langle x_i(t_i)^2 \rangle. \quad (11.1.3-20)$$

Therefore, when the cross-correlation among the various processes vanishes, the average of the square of a sum is the sum of the averages of the squares.

Also defined is a central correlation, the "covariance" between two random process (which can be the same process at the same or different times):

$$\mu = \langle (x(t_1) - \langle x(t_1) \rangle) (y(t_2) - \langle y(t_2) \rangle) \rangle$$

or

$$\mu = \langle x(t_1) y(t_2) \rangle - \langle x(t_1) \rangle \langle y(t_2) \rangle$$

(11.1.3-21)

Normalization of μ produces what is known as the correlation coefficient of two random processes:

$$\rho = \frac{\mu}{\sigma_x \sigma_y} = \frac{\langle x(t_1) y(t_2) \rangle - \langle x(t_1) \rangle \langle y(t_2) \rangle}{\sqrt{\langle (x(t_1) - \langle x(t_1) \rangle)^2 \rangle \langle (y(t_1) - \langle y(t_1) \rangle)^2 \rangle}} \quad (11.1.3-22)$$

It is seen, for statistically independent processes where $\langle x(t_1) y(t_2) \rangle = \langle x(t_1) \rangle \langle y(t_2) \rangle$ that $\mu = \rho = 0$. The correlation coefficient has the property that

$$-1 \leq \rho \leq 1 \quad (11.1.3-23)$$

If $y(t_2) = x(t_2)$ then the covariance and correlation coefficient express the degree of correlation between the ensemble $x^{(k)}(t)$ at time t_1 and itself at time t_2 .

It should be mentioned that a value of zero for the covariance does not imply statistical independence even though statistical independence does imply a zero covariance.

11.1.4 Stationary and Ergodic Processes

Computation of the various averages are complicated by the fact that they depend on the time at which the ensemble is averaged. Most physical processes require these complications. The statistics of receivers as they warm up certainly depend on the time from turn-on.

Fortunately, a large class of processes that are encountered possess, either exactly or approximately, the property of being "stationary". That is, their statistics do not depend on translations in time, or the time origin. Mathematically, this is expressed by the statement that

$$\langle F(x(t_1)) \rangle = \langle F(x(t_1 + \tau)) \rangle \quad \text{for all } \tau, \quad (11.1.4-1)$$

and

$$\langle F(x(t_1), x(t_2)) \rangle = \langle F(x(t_1 + \tau), x(t_2 + \tau)) \rangle \quad \text{for all } \tau, \quad (11.1.4-2)$$

and so on for higher order averages; for arbitrary functions F .

In particular these equations imply

$$\langle F(x(t_1)) \rangle = \langle F(x(0)) \rangle \quad (11.1.4-3)$$

and

$$\langle F(x(t_1), x(t_2)) \rangle = \langle F(x(0), x(t_2 - t_1)) \rangle. \quad (11.1.4-4)$$

The mean at an arbitrary time is the same as that for $t = 0$, and second means depend only on time differences.

It will be seen that stationary processes afford a considerable computational savings.

A further restriction on stationary random process known as "ergodicity" reduces the problems farther. Ergodicity entails:

1. Each record of an ensemble is statistically equivalent (identical means, variances, etc.) to all others.
2. Ensemble averages over records may be equated to time averages over a single record.

11.1.5 Definition of the Fourier Transform And the Energy Spectrum for a Single Record

The Fourier transform that has been used throughout this document is

$$\mathcal{F}\{x(t)\} = F(\omega) = \int_{-\infty}^{\infty} x(t) e^{-i\omega t} dt \quad (11.1.5-1a)$$

where $x(t)$ is a record considered to be magnitude integrable so that $F(\omega)$ will exist. To recover $x(t)$ from its transform use the relation

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega \quad (11.1.5-1b)$$

As an extension from Fourier series for periodic functions the quantity

$$S(\omega) = |F(\omega)|^2 \quad (11.1.5-2)$$

is called the energy spectrum of the function $x(t)$. A justification for this nomenclature for a non-periodic function follows from Parseval's theorem, * namely

$$\int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} \frac{|F(\omega)|^2}{2\pi} d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) d\omega \quad (11.1.5-3)$$

That is; the total signal energy is equal to the integrated "energy spectrum". If the signal is passed through a narrow bandpass filter at $\omega = \omega_0$, then only $|F(\omega)|^2/2\pi$ will contribute to the output energy (a realizable filter also has a peak at $-\omega_0$ contributing energy at $-\omega_0$). Therefore, the signal can be thought of as composed of energy distributed over the frequency domain by its energy spectrum.

A heuristic proof, of a specialization of Parseval's* theorem, will be given in a general fashion for use in subsequent work. The integral

$$C(\tau) = \int_{-\infty}^{\infty} x^*(t) x(t + \tau) dt \quad (11.1.5-4)$$

Generally Parseval's Theorem states $\int_{-\infty}^{\infty} x(t) y^(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} A_x(\omega) A_y^*(\omega) d\omega$

where $A_x(\omega)$ and $A_y(\omega)$ are the Fourier transforms of $x(t)$ and $y(t)$, respectively.

in terms of its Fourier transform, using the defining relation 11.1.5-2:

$$\int_{-\infty}^{\infty} x(t)^* x(t + \tau) dt = \int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} d\omega' \frac{F(\omega) F^*(\omega')}{2\pi} e^{i\omega\tau} \frac{e^{-i(\omega - \omega')t}}{2\pi} \quad (11.1.5-5)$$

as

$$\begin{aligned} F(x(t + \tau)) &= e^{i\omega\tau} F(x(t)) \\ &= e^{i\omega\tau} F(\omega). \end{aligned} \quad (11.1.5-6)$$

Interchanging orders of integration and using the relation

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i(\omega - \omega')t} dt = \delta(\omega - \omega'), \quad (11.1.5-7)$$

where $\delta(\omega - \omega')$ is the "Dirac delta function" or the "impulse function" with the property

$$\int_{-\epsilon}^{\epsilon} \delta(\omega) g(\omega) d\omega = g(0)$$

and $g(\omega)$ continuous, equations (11.1.5-4) and (11.1.5-5) become

$$\begin{aligned} C(\tau) &= \int_{-\infty}^{\infty} x(t)^* x(t + \tau) dt \\ &= \int_{-\infty}^{\infty} \frac{|F(\omega)|^2}{2\pi} e^{i\omega\tau} d\omega \end{aligned} \quad (11.1.5-8)$$

which is a specialization of Parseval's Theorem. A result above (equation (11.1.5-3)) is a further specialization to $\tau = 0$.

Equation (11.1.5-8) is, in terms of the energy spectrum,

$$C(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) e^{i\omega\tau} d\omega \quad (11.1.5-9)$$

and it can be seen that $C(\tau)$ and $S(\omega)$ are Fourier transform pairs (equation (11.1.5-1)).

Therefore,

$$S(\omega) = \int_{-\infty}^{\infty} C(\tau) e^{-i\omega\tau} d\tau \quad (11.1.5-10)$$

A further note of interest is that

$$S(-\omega) = S^*(\omega) \quad (11.1.5-11)$$

from equation (11.1.5-2) which implies the relations:

$$(a) \quad \int_{-\infty}^{\infty} |x(t)|^2 dt = \frac{1}{\pi} \int_0^{\infty} S(\omega) d\omega \quad (11.1.5-12)$$

The energy from both negative and positive frequencies is the same as twice that from only the positive frequencies;

(b) and as $C(-\tau) = C^*(\tau) = C(\tau)$ for real $x(t)$

$$S(\omega) = 2 \int_0^{\infty} C(\tau) \cos \omega\tau d\tau \quad (11.1.5-13a)$$

and

$$C(\tau) = \frac{1}{2\pi} \int_0^{\infty} S(\omega) \cos \omega\tau d\omega \quad (11.1.5-13b)$$

which means that $S(\omega)$ and $C(\tau)$ are cosine transforms of each other for real functions $x(t)$.

11.1.6 Definition of the Power Spectrum and Auto-Correlation Function for a Single Record

While the energy spectrum $S(\omega)$ and, its transform $C(\tau)$ discussed in the previous section are often useful in discussions of signal processing, when a short term signal is present, their utility is limited in the discussions of long term random processes. This fact occurs due to the lack of existence of the Fourier integrals calculated there. For a transform to exist it must be true, in general, that a function be magnitude integrable; that is,

$$\int_{-\infty}^{\infty} |x(t)| dt < \infty \quad (11.1.6-1)$$

This limit does not exist for most random processes and, in fact, does not exist for familiar sines and cosines.

To avoid these difficulties the "power spectrum" is used rather than the "energy spectrum". The method of getting around the difficulties involves the "average power" of a process which is a familiar idea. It is defined by the relation

$$\begin{aligned}\bar{P} &= \text{Average Power} \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T |x(t)|^2 dt\end{aligned}\quad (11.1.6-2)$$

The statement that a direct current voltage of $\sqrt{\bar{P}}$ across a one ohm resistor would have an average power of $\bar{P}$ is a typical way of giving physical meaning to the idea of the "average power".

The concept of power spectrum is obtained through the artifice of defining a truncated version $x_T(t)$ of the random process $x(t)$; namely

$$x_T(t) = \begin{cases} x(t), & |t| \leq T \\ 0, & \text{otherwise} \end{cases}\quad (11.1.6-3)$$

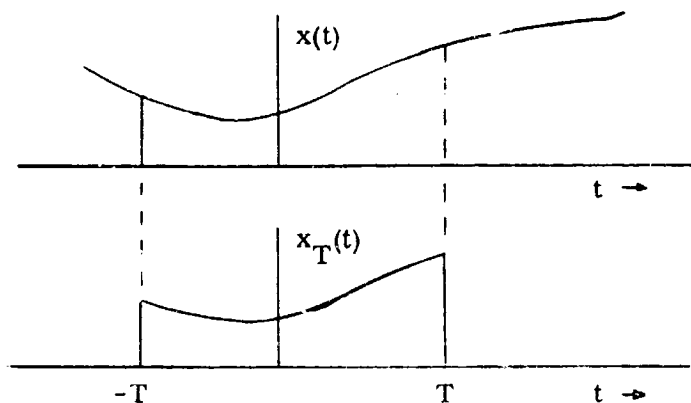


Figure 11.1.6-4. Truncation to a Finite Time Process

Taking the Fourier transform of $x_T(t)$ which exists as $x_T(t)$ is certainly square integrable;

$$\begin{aligned}F_T(\omega) &= \int_{-T}^T x(t) e^{-i\omega t} dt \\ &= \int_{-\infty}^{\infty} x_T(t) e^{-i\omega t} dt\end{aligned}\quad (11.1.6-4)$$

where $F_T(\omega)$ has been subscripted with T to indicate its explicit dependence on T .

Now compute the quantity

$$\frac{1}{2T} \int_{-T}^T |x(t)|^2 dt = \int_{-T}^T dt \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} d\omega' \frac{F_T(\omega) F_T^*(\omega')}{4\pi T} \frac{e^{i(\omega-\omega')t}}{2\pi}$$

as

(11.1.6-5)

$$x_T(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F_T(\omega) e^{i\omega t} d\omega$$

Assuming orders of integration can be changed:

$$\frac{1}{2T} \int_{-T}^T |x(t)|^2 dt = \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} d\omega' \frac{F_T(\omega) F_T^*(\omega')}{4\pi T} \frac{1}{2\pi} \int_{-T}^T dt e^{i(\omega-\omega')t}$$

(11.1.6-6)

Taking the limit at $T \rightarrow \infty$ and assuming the limit

$$P(\omega) = \lim_{T \rightarrow \infty} \frac{|F_T(\omega)|^2}{2T}$$

(11.1.6-7)

exists, plus recognizing the time integral as the Dirac delta function, gives

$$\bar{P} = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-\infty}^{\infty} |x_T(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} P(\omega) d\omega$$

(11.1.6-8)

As in the case of the energy spectrum, $P(\omega)$ is considered as the distribution of "power" over the frequencies. Because $P(\omega) = P(-\omega)$, an even function, the average power may be looked upon as being distributed over any positive frequencies as

$$\bar{P} = \frac{1}{2\pi} \int_{-\infty}^{\infty} P(\omega) d\omega = \frac{1}{\pi} \int_0^{\infty} P(\omega) d\omega$$

(11.1.6-9)

In the preceding section Parseval's Theorem for energy spectrums was derived. An equivalent relation for power spectrums can also be developed. Simply generalize the above by considering the time average:

$$R(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x^*(t) x(t + \tau) dt \quad (11.1.6-10)$$

which is called the auto-correlation function for a single record. Note that

$$R(0) = \overline{P} \quad , \text{ the average power,} \quad (11.1.6-11)$$

and that

$$R(-\tau) = R^*(\tau), \quad (11.1.6-12a)$$

which for real $x(t)$ is

$$R(-\tau) = R(\tau) \quad (11.1.6-12b)$$

Namely; $R(\tau)$ is an even function for real $x(t)$. It also can easily be shown that

$$R(0) \geq |R(\tau)| \quad (11.1.6-13)$$

Inserting the expression in equation (11.1.6-10) for $x(t)$ in terms of its transform it also can be found that

$$\begin{aligned} R(\tau) &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x^*(t) x(t + \tau) dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} P(\omega) e^{i\omega\tau} d\omega \end{aligned} \quad (11.1.6-14)$$

and as before, for energy spectra, the power spectrum and the auto-correlation function for a single record are Fourier transform pairs; that is

$$R(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} P(\omega) e^{j\omega\tau} d\omega \quad (11.1.6-15a)$$

and

$$P(\omega) = \int_{-\infty}^{\infty} R(\tau) e^{-j\omega\tau} d\tau \quad (11.1.6-15b)$$

11.1.7 Power Spectrum and Auto-Correlation Functions of Random Processes

In the preceding sections a single record was considered. Discussed here is the spectrum and auto-correlation function of an ensemble of records of a random process (Section 11.1.3).

Form the ensemble average of the power spectrums of the member functions and explicitly denote the time and ensemble member dependence of a truncated single record spectrum by $P_T(\omega, x^{(k)})$; the ensemble averaged power spectrum is (if the limit exists)

$$\begin{aligned} P(\omega) &= \lim_{T \rightarrow \infty} \langle P_T(\omega, x^{(k)}) \rangle \\ &= \lim_{T \rightarrow \infty} \left\langle \frac{|F_T(\omega, x^{(k)})|^2}{2T} \right\rangle \end{aligned} \quad (11.1.7-1)$$

Similarly, define the ensemble averaged auto-correlation function as

$$\Phi_{xx}(t, t+\tau) = \langle x^{(k)}(t) x^{(k)}(t+\tau) \rangle \quad (11.1.7-2)$$

Consider the quantity

$$\frac{1}{2T} \int_{-T}^T x_T^{(k)*}(t) x_T^{(k)}(t+\tau) dt \quad (11.1.7-3)$$

upon which the ensemble and time averages will eventually be taken. From previous work

$$\begin{aligned} \frac{1}{2T} \int_{-T}^T x_T^{(k)*}(t) x_T^{(k)}(t+\tau) dt &= \frac{1}{2T} \int_{-T}^T dt \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} d\omega' \frac{F_T(\omega, x^{(k)}) F_T^*(\omega', x^{(k)})}{2\pi} \\ &\quad \cdot \frac{e^{-i(\omega-\omega')t}}{2\pi} e^{i\omega\tau} = \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} d\omega' \frac{F_T(\omega, x^{(k)}) F_T^*(\omega', x^{(k)})}{4\pi T} e^{i\omega\tau} \cdot \frac{1}{2\pi} \int_{-T}^T e^{-i(\omega-\omega')t} dt \end{aligned} \quad (11.1.7-4)$$

Obviously, things have been arranged in equation (11.1.7-4) so that ensemble averaging and limiting will give previously defined quantities.

Performing the ensemble average:

$$\begin{aligned} \frac{1}{2T} \int_{-T}^T \langle x_T^{(k)*}(t) x_T^{(k)}(t+\tau) \rangle dt &= \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} d\omega' \left\langle \frac{F_T(\omega, x^{(k)}) F_T^*(\omega', x^{(k)})}{4\pi T} \right\rangle e^{i\omega\tau} \\ &\quad \cdot \frac{1}{2\pi} \int_{-T}^T e^{-i(\omega-\omega')t} dt \end{aligned} \quad (11.1.7-5)$$

Taking the limit as $T \rightarrow \infty$ and identifying the various quantities as previously defined:

$$\begin{aligned}
 \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \langle x_T^{(k)}(t) x_T^{(k)}(t+\tau) \rangle dt &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-\infty}^{\infty} \Phi_{xx}(t, t+\tau) dt \\
 &= \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} d\omega' \left[\lim_{T \rightarrow \infty} \left\langle \frac{F_T(\omega, x^{(k)}) F_T^*(\omega', x^{(k)})}{4\pi T} \right\rangle \right] e^{i\omega\tau} \delta(\omega - \omega') \\
 &= \int_{-\infty}^{\infty} \left[\lim_{T \rightarrow \infty} \left\langle \frac{|F_T(\omega, x^{(k)})|^2}{4\pi T} \right\rangle \right] e^{i\omega\tau} d\omega
 \end{aligned} \tag{11.1.7-6}$$

or

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \Phi_{xx}(t, t+\tau) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} P(\omega) e^{i\omega\tau} d\omega \tag{11.1.7-7}$$

In words, the time averaged ensemble auto-correlation function is the Fourier transform of the ensemble spectrum. As the quantities are Fourier transform pairs the relation

$$P(\omega) = \int_{-\infty}^{\infty} \left[\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \Phi_{xx}(t, t+\tau) dt \right] e^{i\omega\tau} d\tau \tag{11.1.7-8}$$

is also obtained.

Now, equation(11.1.7-8) is reasonably complicated to use for calculations; but in the special case of stationary processes (Section 11.1.4):

$$\Phi_{xx}(t, t+\tau) = \Phi_{xx}(0, \tau) \equiv \phi_{xx}(\tau) \tag{11.1.7-9}$$

giving

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \Phi_{xx}(t, t+\tau) dt = \phi_{xx}(\tau) \tag{11.1.7-10}$$

which produces

$$P(\omega) = \int_{-\infty}^{\infty} \phi_{xx}(\tau) e^{-i\omega\tau} d\tau \tag{11.1.7-11}$$

and, also

$$\phi_{xx}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} P(\omega) e^{i\omega\tau} d\omega. \quad (11.1.7-12)$$

These are fundamental results. Knowledge of the spectrum of a stationary random process allows computation of the auto-correlation function through Fourier transformation, and conversely. In particular, if the process is Gaussian with mean zero, then $\phi_{xx}(\tau)$, normalized to $\phi_{xx}(0)$, is the correlation coefficient $\rho(\tau) = \phi_{xx}(\tau)/\phi_{xx}(0)$.

The formulae can be specialized a bit farther. If the random process is not only stationary but ergodic, such that time averages can be equated over a single record (k) to ensemble averages; then

$$\begin{aligned} \Phi_{xx}(t, t+\tau) &= \phi_{xx}(\tau) = R_{xx}(\tau) \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x^{(k)}(t) x^{(k)}(t+\tau) dt \end{aligned} \quad (11.1.7-13)$$

and

$$P(\omega) = \lim_{T \rightarrow \infty} \frac{|F_T(\omega, x^{(k)})|^2}{2T} \quad (11.1.7-14)$$

For ergodic processes only a single record need be studied to gather information about the random process. Of course, the finiteness of the record develops errors in calculation of the process statistics. Section 11.1.16 discusses the effects of finite time records upon spectra and auto-correlation functions.

Often when discussing a single process the subscripts will be removed from $\phi_{xx}(\tau)$ ($\phi_{xx}(\tau) = \phi(\tau)$) when confusion will not arise.

11.1.8 Cross-Power Spectra and Cross-Correlation Functions for Random Processes

A development, almost identical to that for a single random process, produces relations between two random processes that are quite similar to the single process case.

As before, define the truncated Fourier transforms of two random process members $x^{(k)}(t)$ and $y^{(k)}(t)$

$$\begin{aligned} A_T(\omega, x^{(k)}) &= \int_{-T}^T x^{(k)}(t) e^{-i\omega t} dt \\ &= \int_{-\infty}^{\infty} x_T^{(k)}(t) e^{-i\omega t} dt \end{aligned} \quad (11.1.8-1)$$

and

$$B_T(\omega, x^{(k)}) = \int_{-\infty}^{\infty} y_T^{(k)}(t) e^{-i\omega t} dt \quad (11.1.3-2)$$

Per record pair the truncated cross-power spectral density is defined by

$$P_T(\omega, x^{(k)}, y^{(k)}) = \frac{A_T^*(\omega, x^{(k)}) B_T(\omega, y^{(k)})}{2T} \quad (11.1.8-3)$$

where the motivation for this definition comes from Parseval's Theorem (modified).

Next form the ensemble average of the truncated cross-power spectral density and take its limit as $T \rightarrow \infty$:

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \langle x_T^{(k)}(t) * y_T^{(k)}(t) \rangle dt = \int_{-\infty}^{\infty} \lim_{T \rightarrow \infty} \left\langle \frac{A_T^*(\omega, x^{(k)}) B_T(\omega, y^{(k)})}{4\pi T} \right\rangle d\omega \quad (11.1.8-4)$$

where the proof of this remark is analogous to that for single random processes.

The cross-power spectrum is defined to be the limit:

$$\begin{aligned} P(\omega, x, y) &= \lim_{T \rightarrow \infty} \langle P_T(\omega, x^{(k)}, y^{(k)}) \rangle \\ &= \lim_{T \rightarrow \infty} \left\langle \frac{A_T^*(\omega, x^{(k)}) B_T(\omega, y^{(k)})}{2T} \right\rangle \end{aligned} \quad (11.1.8-5)$$

In a manner parallel to the preceding section, a relation between the cross-power spectrum and the cross-correlation function can also be found.

The truncated cross-power spectrum for a pair of records $x^{(k)}(t)$, $y^{(k)}(t)$ of length T is found to be

$$P_T(\omega, x^{(k)}, y^{(k)}) = \int_{-\infty}^{\infty} \left[\frac{1}{2T} \int_{-T}^T x_T^{(k)}(t) * y_T^{(k)}(t+\tau) dt \right] e^{-i\omega\tau} d\tau \quad (11.1.8-6)$$

Taking the ensemble average and limiting on T :

$$\begin{aligned} P(\omega, x, y) &= \lim_{T \rightarrow \infty} \int_{-\infty}^{\infty} \left[\frac{1}{2T} \int_{-T}^T \Phi_{xy}(t, t+\tau) dt \right] e^{-i\omega\tau} d\tau \\ &= \lim_{T \rightarrow \infty} \langle P_T(\omega, x^{(k)}, y^{(k)}) \rangle \end{aligned} \quad (11.1.8-7)$$

where $\Phi_{xy}(t, t+\tau)$ is the cross-correlation function.

As before, $P(\omega, x, y)$ and

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \Phi_{xy}(t, t+\tau) dt \quad (11.1.8-8)$$

are Fourier transform pairs, and therefore,

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \Phi_{xy}(t, t+\tau) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} P(\omega, x, y) e^{i\omega\tau} d\omega \quad (11.1.8-9)$$

For stationary processes, as before,

$$\Phi_{xy}(t, t+\tau) \equiv \phi_{xy}(\tau) \quad (11.1.8-10)$$

and

$$\begin{aligned} \frac{1}{2\pi} \int_{-\infty}^{\infty} P(\omega, x, y) e^{i\omega\tau} d\omega &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \Phi_{xy}(t, t+\tau) dt \\ &= \phi_{xy}(\tau) \end{aligned} \quad (11.1.8-11)$$

For ergodic processes

$$\phi_{xy}(\tau) = R_{xy}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x^*(t) y(t+\tau) dt \quad (11.1.8-12)$$

11.1.9 Examples of Auto-Correlation Functions and Their Spectrums (and vice versa).

$$1. \quad \phi_{xx}(\tau) = \alpha e^{-\alpha|\tau|} \cos \beta\tau \quad (11.1.9-1)$$

Its corresponding spectrum is

$$P(\omega) = 4 \left[\frac{\alpha^2}{\alpha^2 + (\omega - \beta)^2} + \frac{\alpha^2}{\alpha^2 + (\omega + \beta)^2} \right] \quad (11.1.9-2)$$

which has peaks at $\omega = \pm\beta$

$$2. \quad P(\omega) = \begin{cases} 2\pi, & |\omega| \leq \omega_0 \\ 0, & \text{otherwise} \end{cases} \quad (11.1.9-3)$$

The corresponding auto-correlation function is (assuming a stationary process)

$$\phi_{xx}(\tau) = 2\omega_0 \frac{\sin \omega_0 \tau}{\omega_0 \tau} \quad (11.1.9-4)$$

which is a typical $\sin(x)/x$ function that falls off rapidly as $x = \omega_0 \tau$ increases. The function decreases markedly when $\omega_0 \tau$ is of the order of π or when

$$\tau = \frac{1}{2f_0} ; f_0 = \frac{\omega_0}{2\pi} \quad (11.1.9-5)$$

The correlation time is of the order of the reciprocal of $f_0 = \frac{1}{2\tau}$

11.1.10 Relationship of Finite Interval Spectrum to Infinite Interval Spectrum

When defining the various spectra: energy, single record power, and random process power, the record(s) was assumed to be of infinite length. In reality, a record can be made only for a finite time T' . Suppose this to be a portion of the infinite record which is assumed to exist.

Calculating the effect of a finite record upon the energy spectrum shows that it affects the resolution with which the "true" (infinite record) spectrum can be measured. Identical results hold for the single record and random process spectrum. Statistical variations also limit spectrum measurement accuracy. Their discussion is deferred to Section 12.

The resolution in the spectrum is approximately given by $1/T'$. That is, the spectrum at two frequencies (hertz, not radians per second) can be obtained no closer than at frequencies $\Delta f = 1/T'$ apart.

This remark is proven by writing

$$x_{T'/2}(t) = u(t) x(t) \quad (11.1.10-1)$$

where

$$x_{T'/2}(t) = \begin{cases} x(t), & t \leq \frac{T'}{2} \\ 0, & \text{otherwise} \end{cases} \quad (11.1.10-2)$$

$$\begin{aligned} x(t) &= \text{the infinite (in time) record,} \\ u(t) &= \begin{cases} 1, & t \leq T'/2 \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Taking the Fourier transform of $x_{T'/2}(t)$

$$\begin{aligned} F(x_{T'/2}(t)) &= \int_{-\infty}^{\infty} u(t) x(t) e^{-i\omega t} dt \\ &\equiv F_{T'/2}(\omega) \end{aligned} \quad (11.1.10-3)$$

Now Parseval's Theorem states (Section 11.1. 5)

$$\int_{-\infty}^{\infty} x(t) y(t)^* dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega') G(\omega')^* d\omega' \quad (11.1.10-4)$$

where

$$\mathcal{F}(y(t)) = G(\omega') \quad (11.1.10-5)$$

If $y(t)$ is replaced by $e^{-i\omega t} y(t)^*$:

$$\int_{-\infty}^{\infty} x(t) y(t)^* e^{-i\omega t} dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega') G(\omega + \omega')^* d\omega' \quad (11.1.10-6)$$

That is, the Fourier transform of a product of time functions is the "convolution" of the Fourier transforms of the individual functions. Now the Fourier transform of $u(t)$ is

$$\begin{aligned} U(\omega) &= \int_{-T'/2}^{T'/2} e^{-i\omega t} dt \\ &= T' \frac{\sin(\frac{\omega T'}{2})}{(\frac{\omega T'}{2})} \end{aligned} \quad (11.1.10-7)$$

Therefore, the transform of $x_{T'/2}(t)$ is given by

$$F_{T'/2}(\omega) = \frac{T'}{2\pi} \int_{-\infty}^{\infty} F(\omega') \frac{\sin((\omega - \omega') T'/2)}{(\omega - \omega') T'/2} d\omega' \quad (11.1.10-8)$$

As the energy spectrum is simply the magnitude, squared, of $F_{T'/2}(\omega)$ almost any statement about $F_{T'/2}(\omega)$ can be interpreted as a statement about the spectrum.

The function

$$\text{Samp } (\omega - \omega') = \frac{T'}{2\pi} \frac{\sin((\omega - \omega') T'/2)}{(\omega - \omega') T'/2} \quad (11.1.10-9)$$

is sharply peaked at $\omega = \omega'$ for T' large. The width of the peak is of the order of $2\pi/T'$. The ripples of its tails away from $\omega = \omega'$ tend to cancel the contributions to the integral from $\omega - \omega' > \frac{2\pi}{T'}$. In the limit as $T' \rightarrow \infty$, this function tends toward the Dirac Delta function and (as $T' \rightarrow \infty$)

$$F_{\infty}(\omega) = \int_{-\infty}^{\infty} F(\omega') \delta(\omega - \omega') d\omega' \quad (11.1.10-10)$$

or

$$F_{\infty}(\omega) = F(\omega)$$

as expected. For finite T' the effect of the function $\text{Sample}(\omega - \omega')$ (as a bit of thought will show) is to sample the "true" spectrum $F(\omega)$ at the point $\omega = \omega'$ while "smearing" in other values of $F(\omega')$ in a region of approximately $\pm\pi/T'$ about ω . Therefore, the spectrum can be resolved only at points approximately $2\pi/T'$ distant from each other in ω , or in terms of frequency $f = \omega/2\pi$, at points spaced approximately $\Delta f = 1/T'$ from each other if there are only a record T seconds in length.

As mentioned, statistical spectrums have other problems when they are measured which are discussed in Section 12.

11.1.11 The Effect of a Linear Time Invariant System on the Input Power Spectrum

The output of a linear time invariant filter in terms of the Fourier transform of the output say, $y(t)$ is given by

$$B_T(\omega) = H(\omega) A_T(\omega) = \int_{-\infty}^{\infty} y_T(t) e^{-i\omega t} dt, \quad (11.1.11-1)$$

where $H(\omega)$ is the transfer function of the filter and $A_T(\omega)$ is the Fourier transform of the truncated input $x(t)$, namely; $x_T(t)$ (see Section 11.1.6). The output average power is

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T |y_T(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |H(\omega)|^2 \lim_{T \rightarrow \infty} \frac{|A_T(\omega)|^2}{2T} d\omega, \quad (11.1.11-2)$$

from Parseval's Theorem, and comparison with equation (11.1.6-8), the output spectrum is

$$P_{\text{out}}(\omega) = |H(\omega)|^2 P_{\text{in}}(\omega) \quad (11.1.11-3)$$

While this result was derived for a single record, an identical development using ensemble averaging for random processes as in Section 11.1.7 produces the same result for statistical spectrums.

If the input to a time invariant, infinite operating time network is stationary then the output is stationary. This follows immediately as the output from the network for an input translated in time is simply the translated output. All statistical quantities calculated upon the output will be independent of the translation if it is true of the input.

In terms of frequencies alone the output power of the filtered input may be written

$$\begin{aligned} \bar{P}_{\text{out}} &= \int_{-\infty}^{\infty} P_{\text{out}}(\omega) d\omega = 2 \int_0^{\infty} P_{\text{out}}(\omega) d\omega, \\ &= \int_{-\infty}^{\infty} |H(\omega)|^2 P(\omega) d\omega = 2 \int_0^{\infty} |H(\omega)|^2 P(\omega) d\omega. \end{aligned} \quad (11.1.11-4)$$

11.2 SAMPLED DATA SPECTRUMS AND AUTO-CORRELATION FUNCTIONS

The concept of a power spectrum and an auto-correlation function for sampled data or discrete sequences is quite similar to that for continuous functions of time. Most developments for continuous functions of time can be extended immediately to sequences. Relations also can be developed relating the sampled sequence spectrum to the continuous spectrum if the sampled sequence is derived from a continuous function. It is not, however, necessary that the sequence be derived from a continuous waveform to use the various ideas applicable to continuous functions of time.

Subsequent developments in this section require the use of the two-sided z transform that is, both positive and negative powers of z are required. The necessity for this two-sided transform are its use in the description of statistical processes and the relative simplicity of power spectrum/auto-correlation function relations. The two-sided z transform is defined by

$$X(z) = \sum_{k=-\infty}^{\infty} x_k z^{-k} \quad (11.2-1)$$

Splitting the sum into one for positive and one for negative k it can be seen from the considerations of Section 3.4.2 that bounded sequences for positive time (k) have poles within the unit circle and those for negative time have poles outside the unit circle. Therefore if this relation is to be inverted, with x_k given in terms of $X(z)$, then the unit circle can be considered as the path of integration if bounded sequences are desired. Deformations of the contour that do not disturb the pole content within or without the unit circle are certainly allowed.

Equation (11.2-1) is a Taylor-Laurent series* and may be inverted, within the foregoing considerations, by the same procedure as for single sided transforms:

$$x_k = \frac{1}{2\pi i} \oint X(z) z^k \frac{dz}{z} \quad (11.2-2)$$

As the unit circle is the path of integration the relation

$$x_k = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} X(\omega) e^{j\omega k T_d} d\omega \quad (11.2-3)$$

* Introduction to Complex Variable Theory and Applications, R. V. Churchill, McGraw-Hill

may be written where

$$z = e^{j\omega_s T}, \quad \omega_s = \frac{2\pi}{T}$$

and notation has been understandably abridged: $X(\omega) = X(\dots)$

Often the complex conjugate sequence x_k^* and a certain form of its transform will be needed in subsequent sections. The quantity of interest is

$$\begin{aligned} X(z) &= \sum_{k=-\infty}^{\infty} x_k^* z^k \\ &= \left[\sum_{k=-\infty}^{\infty} x_k (z^*)^k \right]^* \\ &= X^* \left(\frac{1}{z^*} \right) \end{aligned} \tag{11.2-4}$$

If x_k is real then, obviously

$$X(z) = X \left(\frac{1}{z} \right) \tag{11.2-5}$$

Evaluation along the unit circle is of interest: in this case

$$X(z) = X(\omega),$$

in abridged notation and

$$X(\omega) = X^*(\omega). \tag{11.2-6}$$

When the sequence x_k is not real, and the z notation is desired, the conjugation (11.2-4) must be used. If the unit circle is acceptable then (11.2-6) may be used. Real sequences allow simplification of formula appearance by (11.2-5).

Most relations will be written in both the frequency, ω , notation and the more general z notation. However, it will be seen that physical understanding will be most easily be obtained in terms of power distribution versus frequency via the spectrum.

Actual computations will be most easily computed using the residue method of evaluating contour integrals* as $X(z)$ will usually be a rational function of z (a ratio of two polynomials) in practice.

11.2-1 Definition of the Sequence Energy Spectrum and Correlation Function

The energy spectrum for sequences is defined by manipulating the sequence energy

$$E = \sum_{k=-\infty}^{\infty} |x_k|^2 \quad (11.2.1-1)$$

if the limit exists.

It is easy to show that

$$\begin{aligned} E &= \sum_{k=-\infty}^{\infty} |x_k|^2 \\ &= \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |X(\omega)|^2 d\omega, \end{aligned} \quad (11.2.1-2a)$$

with notation abridged so that $X(\omega) = X(z)$ and also that

$$E = \frac{1}{2\pi i} \oint X(z) X^*\left(\frac{1}{z^*}\right) \frac{dz}{z} \quad (11.2.1-2b)$$

where the contour is the unit circle. The pertinent relations for developing this result are the "orthogonality" properties

$$\frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} e^{i\omega(j-k)T} d\omega = \delta_{jk} \quad (11.2.1-3)$$

* IBID

or

$$\frac{1}{2\pi i} \oint \frac{dz}{z^{j-k+1}} = \delta_{jk}$$

with the contour enclosing the origin. Inserting (11.2-3) into (11.2.1-2b)

$$\frac{1}{2\pi i} \oint_{\text{unit circle}} X(z) X^* \left(\frac{1}{z^*} \right) \frac{dz}{z}$$

$$= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} x_j x_k^* \frac{1}{2\pi i} \oint_{\text{unit circle}} \frac{dz}{z^{j-k+1}} \quad (11.2.1-4)$$

$$= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} x_j x_k^* \delta_{jk}$$

$$= \sum_{k=-\infty}^{\infty} |x_k|^2$$

The frequency form (11.2.1-2a) of the relation is found in a similar manner.

As in the analog situation the quantity

$$S(\omega) = |X(\omega)|^2 \quad (11.2.1-5)$$

is called the energy spectrum. It can be related relatively easily to the energy correlation sequence. The energy spectrum will be found to be the z transform (two-sided) of the energy correlation sequence.

Define the correlation sequence by

$$C_{\tau} = \sum_{k=-\infty}^{\infty} x_k^* x_{k+\tau} \quad (11.2.1-6)$$

with τ an integer and assuming that the summation converges. Using the orthogonality relations above it may be shown in developments similar to those in (11.1-4) that

$$\begin{aligned} C_{\tau} &= \sum_{k=-\infty}^{\infty} x_k^* x_{k+\tau} \\ &= \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} S(\omega) e^{i\omega\tau T} d\omega \end{aligned} \quad (11.2.1-7)$$

and also

$$= \frac{1}{2\pi i} \oint X(z) X^* \left(\frac{1}{z^*} \right) z^{\tau} \frac{dz}{z}$$

with the unit circle as the contour. Therefore as the z transform of a sequence is unique it can also be seen that $S(\omega)$ is the z transform of τ .

$$S(\omega) = \sum_{\tau=-\infty}^{\infty} C_{\tau} e^{-i\omega\tau T} \quad (11.2.1-8a)$$

or aborting notation.

$$S(z) = \sum_{\tau=-\infty}^{\infty} C_{\tau} z^{-\tau} \quad (11.2.1-8b)$$

Obviously the total energy, E , is related to the energy correlation sequence by

$$E = C_0 \quad (11.2.1-9)$$

11.2.2 Definition of the Sequence Power Spectrum and Auto-correlation Function

Often in theoretical and practical situations the sequence is infinitely, or nearly so, long with the sequence members large enough relatively so that the energy spectrum does not tend to converge. In this case the sequence power spectrum is defined along with the power auto-correlation function.

The sequence power is defined by

$$\bar{P} \equiv \lim_{N \rightarrow \infty} \frac{\sum_{k=-N}^N |x_k|^2}{2N+1} \quad (11.2.2-1)$$

Using the artifice of a truncated transform

$$X_N(\omega) = X_N(z) \quad (11.2.2-2)$$

$$= \sum_{j=-N}^N x_j z^{-j}$$

then

$$\bar{P} \equiv \lim_{N \rightarrow \infty} \sum_{k=-N}^N |x_k|^2$$

$$= \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} \bar{P}_s(\omega) d\omega \quad (11.2.2-3)$$

where

$$\bar{P}_s(\omega) = \lim_{N \rightarrow \infty} \frac{|X_N(\omega)|^2}{2N+1}$$

if the limit exists (with "bent" notation as usual) or in the alternative form

$$\bar{P} = \frac{1}{2\pi i} \oint \bar{P}_s(z) \frac{dz}{z} \quad (11.2.2-4)$$

with integration about the unit circle and

$$P_s(z) = \lim_{N \rightarrow \infty} \left[\frac{X_N(z) X_N^*\left(\frac{1}{z^*}\right)}{2N+1} \right]$$

$P(\omega)$ can be interpreted as the frequency distribution of the sequence power as in the analog case.

A sequence power auto-correlation function may be defined as for the energy spectrum and a transform relation found. The power auto-correlation function is defined as

$$R_\tau = \lim_{N \rightarrow \infty} \frac{\sum_{k=-N}^N x_k^* x_{k+\tau}}{2N+1} \quad (11.2.2-5)$$

Using the truncated transform $X_N(z)$ it may be seen upon taking the limit at

$$R_\tau = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} \bar{P}_s(\omega) e^{i\omega\tau T} d\omega \quad (11.2.2-6)$$

or

$$R_\tau = \frac{1}{2\pi i} \oint \bar{P}_s(z) z^\tau \frac{dz}{z}$$

Obviously, the power distribution is the z transform of the power auto-correlation function:

$$\bar{P}_s(z) = \sum_{\tau=-\infty}^{\infty} R_{\tau} z^{-\tau} \quad (11.2.2-7)$$

11.2.3 Sequence Power Spectrum and Auto-correlation Function For Random Processes

In the preceding section a single sequence was considered. The ensemble averages are of interest (Section 11.1.3).

As for continuous processes the ensemble averaged auto-correlation function is defined.

$$\Phi_{xx}(kT, (k+\tau)T) = \langle x_k x_{k+\tau} \rangle \quad (11.2.3-1)$$

and if the sequence is stationary

$$\Phi_{xx}(kT, (k+\tau)T) \equiv \varphi_{xx}(\tau T); \quad (11.2.3-2)$$

namely, the statistical auto-correlation function is independent of the origin.

The ensemble averaged sequence power spectrum is defined also in a manner similar to continuous processes

$$P_s(\omega) = \lim_{N \rightarrow \infty} \left\langle \frac{|X_N(\omega)|^2}{2N+1} \right\rangle \quad (11.2.3-3)$$

if the limit exists, where $X_N(\omega)$ was defined in the previous section.

To find the relation between the statistical ensemble averaged sequence auto-correlation function and the averaged power spectrum the relation

$$\sum_{k=-N}^N \frac{x_k^* x_{k+\tau}}{2N+1} = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} \frac{X_N(\omega) X_N^*(\omega)}{2N+1} e^{i\omega\tau T} d\omega \quad (11.2.3-4)$$

may be written. Taking the ensemble average and identifying the various quantities:

$$\lim_{N \rightarrow \infty} \sum_{k=-N}^N \frac{\Phi_{XX}(kT, (k+\tau)T)}{2N+1} = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_S(\omega) e^{i\omega\tau T} d\omega \quad (11.2.3-5)$$

Therefore, in general $P_S(\omega)$, the statistical power spectrum, is the transform of the sequence

$$\lim_{N \rightarrow \infty} \sum_{k=-N}^N \frac{\Phi_{XX}(kT, (k+\tau)T)}{2N+1} \quad (11.2.3-6)$$

In transform notation (abridged: $P_S(\omega) = P_S(z)$)

$$P_S(z) = \sum_{\tau=-\infty}^{\infty} \left[\lim_{N \rightarrow \infty} \sum_{k=-N}^N \Phi_{XX}(kT, (k+\tau)T) \right] z^{-\tau} \quad (11.2.3-7)$$

Expressed in words, the transform of the time averaged ensemble auto-correlation function is the ensemble averaged power spectrum.

When the process is stationary, relations simplify considerably (11.2.3-2):

$$\lim_{N \rightarrow \infty} \sum_{k=-N}^N \Phi_{XX}(kT, (k+\tau)T) = \varphi_{XX}(\tau T) \quad (11.2.3-8)$$

and

$$P_S(\omega) = \sum_{\tau=-\infty}^{\infty} \varphi_{XX}(\tau T) e^{-i\omega\tau T} \quad (11.2.3-9a)$$

or

$$P_s(z) = \sum_{\tau=-\infty}^{\infty} \varphi_{xx}(\tau T) z^{-\tau} \quad (11.2.3-9b)$$

Inverting these relations for the stationary case

$$\varphi_{xx}(\tau T) = \frac{1}{2\pi i} \oint P_s(z) z^{\tau} \frac{dz}{z} \quad (11.2.3-10a)$$

or

$$\varphi_{xx}(\tau T) = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s(\omega) e^{-i\omega\tau T} d\omega \quad (11.2.3-10b)$$

If the process is not only stationary but ergodic then

$$\varphi_{xx}(\tau T) = R_{\tau} \quad (11.2.3-11)$$

where R_{τ} was defined previously (Section 11.2.2) as the sequence time average

$$R_{\tau} = \lim_{N \rightarrow \infty} \sum_{k=-N}^N \frac{x_k^* x_{k+\tau}}{2N+1} \quad (11.2.2-5)$$

11.2.4 Relation Between Analog and Sampled Spectra and Auto-Correlation Functions

Of quite a bit of interest is the relation between sampled and continuous spectra and auto-correlation functions when the sampled sequence is derived from a continuous process. The subsequent derivation is applicable to energy and power spectra, deterministic and statistical. Stationarity is necessary for these relations to hold in the statistical case.

Examination of (11.1.7-2) and (11.2.3-1) indicate that the auto-correlation functions for both the analog and sampled situation are identical when the process is stationary. Therefore, as the auto-correlation is related to the analog spectrum, $P(\omega)$ by

$$\varphi_{xx}(\tau T) = \frac{1}{2\pi} \int_{-\infty}^{\infty} P(\omega) e^{i\omega\tau T} d\omega \quad (11.1.7-12)$$

and the sampled spectrum by

$$\varphi_{xx}(\tau T) = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s(\omega) e^{i\omega\tau T} d\omega \quad (11.2.3-10b)$$

then insertion of (11.1.7-12) into (11.2.3-9a) relates the two spectra using the Dirac delta function representation (2.1.2-6).

$$\begin{aligned} P_s(\omega) &= \frac{1}{T} \int_{-\infty}^{\infty} P(\omega') \left[\frac{T}{2\pi} \sum_{\tau=-\infty}^{\infty} e^{i(\omega - \omega')\tau T} \right] d\omega', \\ &= \frac{1}{T} \sum_{k=-\infty}^{\infty} P(\omega - k\omega_s) \end{aligned} \quad (11.2.4-1)$$

As was the case for the "simple" transform of a sequence, the function was "aliased", the energy, deterministic power, and stationary statistical power spectrums are also aliased. If the analog spectrums are bandlimited within half the sampling frequency then the sampled spectrum and analog spectrums are equal within a factor of $1/T$ over the range $|\omega| \leq \omega_s/2$.

It can be seen that spectrums that are "white", flat or constant, over a large region have auto-correlations that peak sharply. Section 3.10 discusses sequence (auto-correlation) widths and their related transform (spectrum) widths in general. A sharp correlation function produces a flat spectrum and, conversely, a sharp spectrum (narrow-band) produces a long correlation time or sequence.

11.2.5 Sequence Cross Spectrums and Cross-Correlation Functions

Cross-spectra and cross-correlation functions are often used in statistical studies and vibration analyses. Their definitions and relations follow closely those for single processes and will be simply itemized here. The cross functions require two process sequences, x_k and y_k .

I Cross Energy

Correlation

$$C_{\tau}^{xy} = \sum_{k=-\infty}^{\infty} x_k^* y_{k+\tau} \quad (11.2.5-1)$$

Power Spectrum

$$S(\omega, x, y) = \sum_{\tau=-\infty}^{\infty} C_{\tau}^{xy} e^{i\omega\tau T} \quad (11.2.5-2)$$

II Cross Power (Deterministic)

Correlation

$$R_{\tau}^{xy} = \lim_{N \rightarrow \infty} \sum_{k=-N}^N \frac{x_k^* x_{k+\tau}}{2N+1} \quad (11.2.5-3)$$

Power Spectrum

$$\bar{P}_S(\omega, x, y) = \lim_{N \rightarrow \infty} \left[\frac{X_N^*(\omega) Y_N(\omega)}{2N+1} \right] \quad (11.2.5-4)$$

and

$$\bar{P}_S(\omega, x, y) = \sum_{\tau=-\infty}^{\infty} R_{\tau}^{xy} e^{-i\omega\tau T} \quad (11.2.5-6)$$

III Cross Power (Statistical)

Correlation

$$\Phi_{xy}(kT, (k+\tau)T) = \langle x_k, y_{k+\tau} \rangle \quad (11.2.5-7)$$

and if the processes are stationary

$$\Phi_{xy}(kT, (k+\tau)T) \equiv \varphi_{xy}(\tau) \quad (11.2.5-8)$$

Power Spectrum

$$P_s(\omega, x, y) = \lim_{N \rightarrow \infty} \left\langle \frac{X_N^*(\omega) Y_N(\omega)}{2N+1} \right\rangle \quad (11.2.5-9)$$

$$= \sum_{\tau=-\infty}^{\infty} \left[\frac{1}{N} \sum_{k=-n}^N \frac{\Phi_{xy}(kT, (k+\tau)T)}{2N+1} \right] e^{-i\omega\tau T}$$

and if the processes are stationary

$$P_s(\omega, x, y) = \sum_{\tau=-\infty}^{\infty} \varphi_{xy}(\tau) e^{-i\omega\tau T} \quad (11.2.5-10)$$

All the various cross-spectra are related to their continuous counterparts if x_k and y_k are derived from stationary continuous process $x(t)$ and $y(t)$ by the "aliasing" relation:

$$P_s(\omega, x, y) = \frac{1}{T} \sum_{m=-\infty}^{\infty} P(\omega - m\omega_s, x, y) \quad (11.2.5-11)$$

where P can mean energy or power spectrums.

11.2.6 Modification of a Stationary Sampled Sequence Power Spectrum by a Digital Filter $H_z(\omega)$, and the Effect upon "White" Noise as an Example

The operation upon a finite sequence with a digital filter in terms of the filter frequency response and the signal spectral content is certainly of interest. Only the statistical case will be presented as the other spectra are modified in an identical manner which can be shown by a similar method.

If the truncated sequence of Section 11.2.2 is filtered then the output sequence transform would be

$$Y_N(\omega) = H_z(\omega) X_N(\omega) \quad (11.2.6-1)$$

where $H_z(\omega)$ is the digital filter transfer function in abridged notation and transient response has been neglected. The output spectrum $P_{sy}(\omega)$ is then

$$\begin{aligned} P_{sy}(\omega) &= \lim_{N \rightarrow \infty} \left\langle \left| \frac{Y_N(\omega)}{2N+1} \right|^2 \right\rangle \\ &= |H_z(\omega)|^2 P_{sx}(\omega) \end{aligned} \quad (11.2.6-2)$$

with $P_{sx}(\omega)$ the input spectrum.

The output sequence power is given, for stationary sequences, by

$$\varphi_{yy}(0) = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |H_z(\omega)|^2 P_{sx}(\omega) d\omega \quad (11.2.6-3)$$

If the input spectrum $P_{sx}(x)$ is flat ("white"); namely,

$$P_{sx}(\omega) = 1; \quad (11.2.6-4)$$

then, the input average sequence power is:

$$\varphi_{xx}(0) = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega \quad (11.2.6-5)$$

$$= 1.$$

The variance of the input noise is the average input sequence power:

$$\sigma^2 = \varphi_{xx}(0) \quad (11.2.6-6)$$

as the average is zero. Therefore the r.m.s. noise voltage, N_{rms} , is

$$N_{\text{rms}} = \sigma = 1. \quad (11.2.6-7)$$

Also, the auto-correlation sequence is found to be

$$\varphi_{xx}(kT) = \begin{cases} 1; & k = 0; \\ 0, & k \neq 0; \end{cases} \quad (11.2.6-8)$$

a flat spectrum means that successive samples are uncorrelated. An auto-correlation sequence of the form (11.6-8) also implies a flat spectrum.

Finally, the digital filter output power is found to be (equation 11.2.6-3)

$$\varphi_{yy}(0) = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |H_z(\omega)|^2 d\omega \quad (11.2.6-9)$$

11.2.7 Modification of the Sampled Sequence Average by a Digital Filter $H_z(z)$

From Section 3.4.1 the impulse response sequence of a time invariant digital filter was found to be related to the transfer function, $H_z(z)$, by

$$H_z(z) = \sum_{k=0}^{\infty} w_k z^{-k}$$

where w_k is the impulse response sequence. At $\omega = 0$

$$\begin{aligned} z &= e^{i\omega T} \Big|_{\omega=0} \\ &= 1 \end{aligned} \quad (11.2.7-1)$$

Therefore:

$$H_z(1) = \sum_{k=0}^{\infty} w_k; \quad (11.2.7-2)$$

the zero frequency response (DC) is equal to the summed impulse response. The notation is abridged, as before, by writing $H_z(0)$ rather than $H_z(1)$ to denote the DC response.

This result is used to determine the relation between input and output sequence averages. As before, assume the input is stationary and that transient effects have decayed, so that the output will be stationary. The result will be that the output average, $\langle y \rangle$, is simply the input average, $\langle x \rangle$, multiplied by the DC response $H_z(0)$:

$$\langle y \rangle = H_z(0) \langle x \rangle \quad (11.2.7-3)$$

From Section 3.4.1, the relation between input and output sequences can be written as

$$y_n = \sum_{j=0}^{\infty} w_{n-j} x_j.$$

Taking ensemble averages and assuming the sequences y_n and x_j are stationary:

$$\langle y \rangle = \langle y_n \rangle, \quad n = 0, \pm 1, \pm 2, \dots,$$

and (11.2.7-4)

$$\langle x \rangle = \langle x_j \rangle, \quad j = 0, \pm 1, \pm 2, \dots,$$

then

$$\langle y \rangle = \sum_{j=0}^{\infty} w_{n-j} \langle x \rangle \quad (11.2.7-5)$$

As $w_k = 0$, for $k < 0$ the result, using equation (11.2.7-2) is:

$$\langle y \rangle = H_z(0) \langle x \rangle. \quad (11.2.7-6)$$

Intuitively, the result is expected. The average input behaves as the DC component and is treated as such by the filter

11.2.8 Use of the Discrete Fourier Transform to Evaluate the Sampled Spectrum For Deterministic Waveforms

In practical situations the energy and power spectrums of deterministic signals often must be calculated. Statistical spectra computations are more complicated due to the spectral estimates depending on statistical data and hence being statistical quantities themselves. Considerations for statistical data spectrum analysis are discussed in Section 12

Computation of deterministic spectra are most easily performed using the discrete Fourier transform of Section 2.6. That is, the truncated transform

$$X(\omega) = \sum_{k=-(N-1)}^N x_k e^{-i\omega k T}$$

is computed at the discrete frequencies

$$\omega_n = \frac{2\pi n}{(2N+1)T}, \quad n = 0, \pm 1, \dots$$

Or defining

$$X_n = X(\omega_n) \tag{11.2.8-1}$$

then

$$X_n = \sum_{j=-(N-1)}^N x_k e^{-\frac{2\pi i}{2N+1} kn} \tag{11.2.8-2}$$

which is the discrete transform DFT2 of Section 2.6.1. The DFT1 of that Section may be used equally as well, though the different range of n should be carefully observed.

The fact that a finite time record is being used to compute the spectra means that a resolution in the spectra less than

$$\Delta f = \frac{1}{2NT} \quad (11.2.8-3)$$

may not be obtained. Section 11.1.10 should be consulted where it may be seen that the effect of a truncated record is to convolve a $\sin x/x$ waveform with the "true" spectrum limiting the resolution. If the high leakage side lobes of the $\sin x/x$ function are of concern in spectrum computation the weighting methods of Section 5.7 may be used with a concomitant loss in resolution by a factor of about 2.

That is the spectrum computed by the discrete Fourier transform would be

$$X^{\sigma}(\omega_n) = \sum_{k=-(N-1)}^N x_k \sigma_k e^{-i \frac{2\pi}{2N+1} k n} \quad (11.2.8-4)$$

where the σ_k are the Lanczos, cosine-squared, or Fejer weights of Section 5.7.

Of course, the discrete Fourier transform should be computed via the fast Fourier transform algorithm of Section 14 if at all possible to economize on computation time.

12. DIGITAL SPECTRUM ESTIMATION

A major problem confronting engineers and researchers is the computation of spectra and cross-spectra that arise from various noise-like sources. Usually certain basic assumptions are used such as stationarity of the time sequence(s) and that sampling rate of an analog record(s), if the sequence(s) is derived by this means, is high enough to avoid aliasing. However, in many applications the stationarity of the process(es) is in question which limits the length of time over which meaningful records may be taken and hence the frequency resolution and statistical stability of estimates. Also, certain "signal-like" waveforms are often contained within a noise spectra and it is the purpose of the spectrum analysis to detect this "signal" and ascertain its characteristics; length of the data record and resolution bandwidth are then often determined by what the experimenter believes is signal duration. Finally, in some modern applications such as speech processing and the study of time varying communications links certain other functions of the Fourier transforms and spectrums have taken on meaning.

Irrespective of all the conflicting aims, meanings, and applications of Fourier transforms and spectrums an understanding of spectrum estimation for noise-like signals is of basic interest and has quite general application.

Subsequent discussion will encompass basic approaches and formulae for spectrum resolution, statistical stability in terms of relative variation of the estimation procedure, and the necessary data record length to achieve these ends. The fast discrete Fourier transform algorithm (Section 14) will be used where necessary for reduction in the computational volume.

It should be pointed out that a simple "periodogram", a conversion of the data directly to the Fourier transform is not a good method of spectrum estimation as estimates are usually exceedingly unstable statistically. Section 12.4 discusses this matter at length.

Estimation of cross-spectra differs only in minor details which will be pointed out where pertinent.

If a more rigorous and detailed discussion of spectrum estimation is desired, the reader is referred to an excellent treatment by Grenander and Rosenblatt, Statistical Analysis of Time Series, Wiley, 1957. However, the treatment here is probably a prerequisite.

12.1 ESTIMATION THROUGH THE USE OF "LAGGED" PRODUCTS

A popular method of implementing estimates of the sampled data spectrum and cross-spectra is through the use of "lagged" products. It is basically an "indirect" method of computing the estimates as the data is not directly Fourier transformed even assuming possible linear weighting of the data before transformation. Estimates of the correlation function, the "lags", are first formed, the amount necessary determined by the desired frequency resolution and stability, then weighted and the transform taken.

Procedurally, it will be found that the fast transform should be used to compute the correlation "lagged" products as was done for discrete convolutions (Section 3.7.5) when a significant number of "lagged" products are needed.

As the lagged products are linearly combined to form spectrum estimates it should not be unexpected that the weights form a "filter" or "window" through which spectrum estimates are made.

Equation (11.2.2-5) suggests taking the lagged products defined by

$$\hat{\rho}_k \approx \frac{1}{N-|k|} \sum_{r=0}^{N-1-|k|} x_r x_{r+|k|}, \quad |k| \leq N-1 \quad (12.1-1)$$

as an estimate of the auto-correlation function. N , of course, is the number of data points. If cross spectra between two processes x_k and y_k are of interest, then the only modification here and in future developments is the replacement

$$\begin{aligned} \hat{\rho}_k &\Leftarrow \hat{\rho}_k^{xy} \\ &\equiv \frac{1}{N-|k|} \sum_{r=0}^{N-1-|k|} x_r y_{r+|k|} \end{aligned} \quad (12.1-2)$$

When actual computations are made, possibly using the fast Fourier transform algorithm, then the calculation of $\hat{\rho}_k^{xy}$ may involve individual transforms of both sequences (Section 12.1.1) which is not as much additional labor as it might seem.

The ensemble average of $\hat{\rho}_k$ is equal to the sampled auto-correlation function assuming a stationary process:

$$\langle \rho_k \rangle = \varphi(kT) \quad (12.1-3)$$

An estimate whose average is equal to the actual value is an "unbiased" estimate. If the two-sided transform of $\hat{\rho}_k$ is taken and considered to be an approximation to the spectrum, which seems quite rational from the various definitions of spectra and correlation functions in Section 11.2, then a spectrum estimate might be taken to be

$$\hat{P}_s(\omega) = \sum_{k=-M}^M \rho_k e^{-i\omega kT} \quad (12.1-4)$$

where M is some integer less than $N-1$. The ensemble average of this spectrum estimate does, in fact, approach the actual spectrum when the number of data points and M becomes large. This uniform weighting has undesirable "leakage" from other frequencies.

A modified estimation procedure which produces a great deal of information about the various aspects of spectrum measurement is a simple linear weighting:

$$\hat{P}_s(\omega) = \sum_{k=-M}^M d_k \hat{p}_k e^{-i\omega kT} \quad (12.1-5)$$

where the weights d_k possess a useful property $d_{-k} = d_k$. It should be noticed that the preceding estimation formula (12.1-4) is contained within this new procedure when all the d_k are considered equal. In z transform notation the estimate may be written as

$$\hat{P}_s(z) = \sum_{k=-M}^M d_k \hat{p}_k z^{-k} \quad (12.1-6)$$

where notation has again been abridged: $P_s(\omega) = P_s(z)$ for notational convenience.

Ensemble averaging the estimate produces

$$\langle \hat{P}_s(\omega) \rangle = \sum_{k=-M}^M d_k \varphi(kT) e^{-i\omega kT} \quad (12.1-7)$$

which becomes

$$\langle \hat{P}_s(\omega) \rangle = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s(\omega') D(\omega - \omega') d\omega' \quad (12.1-8)$$

where relation (11.2.3-10) has been used and the definition

$$D(\omega) = \sum_{k=-M}^M d_k e^{-i\omega kT} \quad (12.1-9)$$

made. In this form the averaged estimate is capable of some interpretation. The function $D(\omega)$ is in the form of a central non-recursive digital filter and may be looked upon as a frequency "window" or aperture through which the actual spectrum contributes to the estimate. All the methods of section 5 may be considered for synthesis of $D(\omega)$, if necessary. Future arguments show that $D(\omega)$ must apparently be a sharply peaked function if the spectrum estimate average (12.1-8) is to be composed mainly of frequencies about the frequency of estimation. In fact; ideally it would approach the Dirac delta function.

$$D(\omega) \Rightarrow \omega_B \delta(\omega) \quad (12.1-10)$$

in the limit of large M if an "asymptotically unbiased" estimate

$$\langle \hat{P}_B(\omega) \rangle \simeq P_B(\omega) \quad (12.1-11)$$

is desired. However, statistical stability considerations indicate that the width of $D(\omega)$ measured by some practical criteria (e.g. Section 3.10) should be finite, and increases as statistical stability increases. Another measure of effective width for $D(\omega)$ more pertinent to spectrum estimation will be introduced subsequently.

An apparent defect of this estimation procedure may be the possibility of negative estimates due to $D(\omega)$ being possibly negative. However, when actual forms for $D(\omega)$ are considered negative estimates can be used to indicate large leakage from side-lobes of $D(\omega)$ as in Figure 12.1-1.

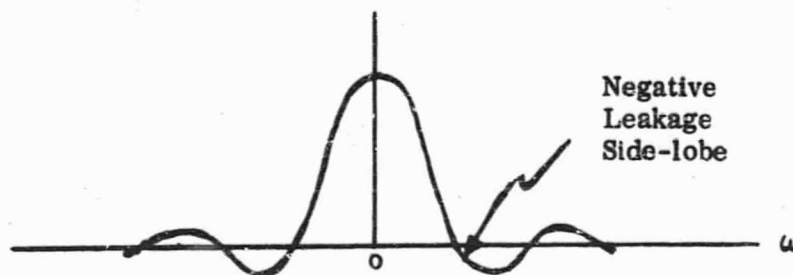


Figure 12.1-1. Typical Frequency Window ($D(\omega)$) Shape

Simple arguments demonstrate, in the limit of large M , what properties $D(\omega)$ must possess. One of these is

$$\frac{1}{\omega_B} \int_{-\frac{\omega_B}{2}}^{\frac{\omega_B}{2}} D(\omega') d\omega' = 1 \quad (12.1-12)$$

To show that this is a desirable property consider first $P_B(\omega)$ as being relatively smooth over the interval of peakedness of $D(\omega')$; then

$$\langle \hat{P}_s(\omega) \rangle = P_s(\omega) \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} D(\omega') d\omega' \quad (12.1-13)$$

if $D(\omega')$ is considered to approach zero as ω' approaches $\pm \frac{\omega_s}{2}$. Obviously, if the averaged estimate is to be equal to the actual spectrum then the property (12.1-12) must hold for large M .

If the actual spectrum is considered next as a single tone, or frequency of ω_0 , and an amplitude of A then $P_s(\omega)$ may be represented by

$$P_s(\omega) = \frac{A^2 \omega_s}{4} \left[\delta(\omega - \omega_0) + \delta(\omega + \omega_0) \right] \quad (12.1-14)$$

where this theoretical power spectrum is computed formally from the definition (11.2.3-3). The total average power is found to be

$$\begin{aligned} \bar{P} &= \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s(\omega') d\omega' \\ &= \frac{A^2}{2} \end{aligned} \quad (12.1-15)$$

as expected. If zero frequency (DC) is considered $P_s(\omega)$ is found to approach a delta function of amplitude $A^2 \omega_s$ and the average power is A^2 , also as expected.

From (12.1-8) the ensemble averaged spectrum estimate with the single tone at ω_0 , not equal to zero, is

$$\langle \hat{P}_s(\omega) \rangle = \frac{A^2}{4} \left[D(\omega - \omega_0) + D(\omega + \omega_0) \right] \quad (12.1-16)$$

Therefore a good spectrum estimator for a single tone should have the property, previously mentioned that

$$D(\omega) \simeq \omega_s \delta(\omega) \quad (12.1-17)$$

in the limit of large M . Fortunately, the two properties (12.1-12) and (12.1-16) are quite consistent mathematically. However, it will be seen that extreme peakedness of $D(\omega)$ will produce statistically unstable estimates. The wider $D(\omega)$ becomes the better the statistical stability with a concomitant loss in resolution. Good resolution conversely requires a loss in statistical stability. The two attributes must be balanced against one another in practice. Subsequent relations will be seen to quantitatively express this balance in terms of the resolution bandwidth, the statistical accuracy, and the data quantity.

To use the discrete Fourier transform and hence the fast transform algorithm (12.1-5) is written as

$$\hat{P}_s(\omega_n) = \sum_{k=-(M-1)}^M d_k \hat{\rho}_k e^{-i\omega_n kT} \quad (12.1-18)$$

with

$$\begin{aligned} \omega_n &= \frac{n\omega_s}{2M} \\ &= \frac{2\pi}{2M} \frac{n}{T}, \quad n = -(M-1), \dots, 0, \dots, M. \end{aligned}$$

The bottom index on (12.1-5) has been removed in writing (12.1-18) as use of the discrete transform causes embarrassment in defining equally spaced frequencies from zero to half the sampling frequency (Section 2.6) with an odd number of terms. In some older books, it was not an unpopular practice to introduce a weighting of $1/2$ on the end terms ($\pm M$) to maintain the symmetry of an odd number of terms about zero. However, this is quite artificial and mathematically inappropriate as these two end terms are not "linearly independent"; one of them is redundant.

Simplifying:

$$\hat{P}_s(\omega_n) = \sum_{k=-(M-1)}^M d_k \hat{\rho}_k e^{-i \frac{2\pi}{2M} nk} \quad (12.1-19)$$

which is a $2M$ point discrete transform, DFT2 of Section 2.6.1. As the number of lags is usually relatively small compared to the number of data points this part of the spectrum computation is not as desperately in need of the fast transform algorithm as is the computation of lags.

It will be seen that the proper choice of the weights d_k , or the frequency aperture $D(\omega)$, will make immediately adjacent estimates slightly correlated statistically, but estimates spaced by two discrete frequencies will be independent, at least to moments of the second order.

Before recasting the estimation formula into another form where frequency weighting, rather than time or lag weighting, is used mention should be made of certain useful compen-

sations that can be made to the lags. Often a zero frequency or "D. C. offset is contained within the data due to inherent data properties, or the method of taking data. A linear trend also may be present due to possible measurement equipment drifting or experimental trends. If these affects are not removed prior to the lag weight and transformation they will quite often obscure the actual data spectrum. For example, a large zero frequency component may completely obscure the other spectral components about zero frequency through side-lobe leakage of $D(\omega)$. Therefore it is quite often useful to remove these effects beforehand.

An excellent estimator for the mean to compensate for D. C. or zero frequency offset is (Section 13.1.8).

$$\langle \hat{x} \rangle = \frac{1}{N} \sum_{r=0}^{N-1} x_r \quad (12.1-20)$$

By using this estimator the lags may be reduced to those of a spectrum with zero actual mean. Some slight residual zero frequency component may remain from statistical effects in actual estimation. Using this mean estimator is quite simple; it can be subtracted from the data x_k , or more simply new lags $\hat{\rho}'_k$ are defined by

$$\hat{\rho}'_k = \rho_k - \langle \hat{x} \rangle^2 \quad (12.1-21)$$

Then $\hat{\rho}'_k$ is used in the previous estimator formulas instead of $\hat{\rho}_k$ to produce the spectrum estimates $\hat{P}_s(\omega)$.

Validity of this simple procedure may be seen by taking ensemble averages and examining the definitions of Section 11.1.3.

When cross-spectrums are being computed, the modification for constant D. C. offset is simply

$$\hat{\rho}_k^{xy'} = \hat{\rho}_k^{xy} - \langle \hat{x} \rangle \langle \hat{y} \rangle \quad (12.1-22)$$

where notation is obvious.

For a linear trend compensation directly on the data is probably easiest with present day computing equipment and techniques. Using the methodology of Sections 13.1.2 and 13.1.7 it is easy to generate a useful correction scheme. The raw data sequence is modified to

$$x'_k = x_k - (\hat{a}_0 + \hat{a}_1 kT) \quad (12.1-23)$$

where x_k has been assumed to contain a linear trend. By using the assumption that the data is uncorrelated from sample to sample (which is usually untrue) a conservative estimation of the coefficients a_0 and a_1 can be made in an "unbiased" manner (Sections 13.1.1 and 13.1.2). These are sometimes called "least-square" estimates and are given by

$$\hat{a}_0 = \frac{1}{AC - B^2} \sum_{r=0}^{N-1} [C - BrT] x_r$$

and

(12.1-24)

$$\hat{a}_1 = \frac{1}{AC - B^2} \sum_{r=0}^{N-1} [-B + ArT] x_r$$

where

$$A = N$$

$$B = N \frac{(N-1)}{2} T$$

$$C = \frac{(N-1) N (2N-1)}{6} T^2$$

Both the constant and linear trend corrections have some effect on the frequency windows which will not be examined here.

While weighting of the lagged products (compensated) is not a very onerous chore, another completely equivalent method of modifying the raw spectral estimates is available which sometimes reduces computation a bit.

Define the un-weighted, or "raw", spectral estimates by

$$\hat{V}(a_m) = \sum_{j=-(M-1)}^M \hat{\rho}_j e^{-i\omega_m jT}, \quad \omega_m = \frac{2\pi m}{2MT} \quad (12.1-25)$$

which is simply (12.1-18) with all $d_k = 1$. Next write the frequency weighting function $D(\omega)$ at the discrete frequencies ω_n

$$D(\omega_n) = \sum_{k=-(M-1)}^M d_k e^{-i\omega_n kT} \quad (12.1-26)$$

Therefore, from the table of Section 2.8, the discrete frequency convolution relation

$$\begin{aligned}\hat{P}_s(\omega_n) &= \sum_{k=-(M-1)}^M d_k \hat{\rho}_k e^{-i \frac{2\pi}{2M} nk} \\ &= \frac{1}{2M} \sum_{m=-(M-1)}^M D(\omega_n - \omega_m) \hat{V}(\omega_m)\end{aligned}\tag{12.1-27}$$

may be written.

Fortunately, $D(\omega)$ may be chosen in a simple form for computation and also have desirable properties for estimation. Namely let

$$\begin{aligned}D(0) &= 2Ma_0 \\ D(\omega_1) &= 2Ma_1 \\ D(-\omega_1) &= 2Ma_1\end{aligned}\tag{12.1-28}$$

and

$$D(\omega_n) = 0, \quad n = 2, 3, \dots, \omega_m.$$

Then, the estimates may be written as

$$\hat{P}_s(\omega_n) = a_0 \hat{V}(\omega_{n-1}) + a_1 \hat{V}(\omega_n) + a_0 \hat{V}(\omega_{n+1})\tag{12.1-29}$$

for $1 \leq n \leq M-1$. The frequencies $\omega_0 = 0$ and $\omega_M = \pi/T = \omega_s/2$ must be treated specially as careful examination of (12.1-27) shows that overlap from positive and negative frequencies can occur. A simple remedy is to use the estimators

$$\hat{P}_s(0) = \frac{1}{2} \hat{V}(0) + \frac{1}{2} \hat{V}(\omega_1)\tag{12.1-30}$$

and

$$\hat{P}_s(\omega_M) = \frac{1}{2} \hat{V}(\omega_{M-1}) + \frac{1}{2} \hat{V}(\omega_M)$$

which is simple arithmetic averaging.

A well known set of values for the a_i is

$$\begin{aligned} a_0 &= 1/2 \\ a_1 &= 1/4 \end{aligned} \quad (12.1-31)$$

and are often called "hanning" weights after a German meteorologist, Julius Von Hahn.* These weights allow both good statistical stability and low sidelobes and are probably as good as any other unless special properties are demanded of the estimator frequency window.

By taking the inverse discrete Fourier transform (IDFT2) of (12.1-26) the weights d_k can be found in terms of the frequency window sampled values

$$d_k = \frac{1}{2M} \sum_{n=-(M-1)}^M D(\omega_n) e^{i\omega_n kT} \quad (12.1-31)$$

and for the weights of (12.1-28)

$$\begin{aligned} d_k &= a_0 + 2a_1 \cos \omega_1 kT \\ &= a_0 + 2a_1 \cos \left(\frac{\pi}{M} k \right) \end{aligned} \quad (12.1-32)$$

For "hanning" weights

$$\begin{aligned} d_k &= \frac{1}{2} (1 + \cos (\frac{\pi}{M} k)) \\ &= \cos^2 \left(\frac{2\pi}{M} k \right) \end{aligned} \quad (12.1-33)$$

which produces another name for the weighting scheme, cosine-squared weighting. This weighting should be familiar from Section 5.7 where it was used for smoothing of transfer functions and to supply an integrating kernel that behaved as a delta function. Here it behaves as a nice low pass filter with low side-lobes.

Design of $D(\omega)$ in general is closely akin to phased array antenna pattern design. In fact the mathematical statements of the problems are identical. With a finite number of terms ($2M$) it is impossible to have arbitrary control of bandwidth (beamwidth) and side-lobes. It is desirable to make the "width" of $D(\omega)$ as narrow as possible concurrent with low leakage side-lobes. The matter is discussed more fully when statistical stability aspects are presented in Section 12.3.

* R. B. Blackman & J. W. Tukey, The Measurement of Power Spectra, Dover, 1959

It is obvious that this finite estimation procedure cannot produce infinite delta functions for a constant frequency input. However, if a tone

$$x_k = A \cos \omega_n kT \quad (12.1-34)$$

is used it can be shown that value of the estimate at the frequency ω_n is approximately

$$\hat{P}_s(\omega_n) = \frac{A^2}{4} M \quad (12.1-35)$$

which does increase with the number of lags used. It has been assumed that the weights d_k are all equal to unity. Other weighting usually reduces this value slightly. By evaluating $D(\omega)$ at zero frequency it may be seen that the d_k satisfy

$$\sum_{k=-(M-1)}^M d_k = D(0) \quad (12.1-36)$$

$D(0)$ must grow with M if $D(\omega)$ is to approach a delta function.

Mention should also be made that these estimates are for positive frequencies. Negative frequencies have been assumed in the basic spectrum definition. Therefore, if only positive frequency spectra are desired the estimated values should be doubled. Secondly, a factor of T , the data sampling interval, should be multiplied with the estimates if an analog spectrum is being computed through the use of sampled data (Section 11.2.4).

Recapitulating, the lagged estimation procedure consists of:

- (1) Forming the lagged products, most probably by fast Fourier transformation (Section 12.1.1).
- (2) Compensate the lags for D.C., or zero frequency constant offset and/or linear trends.

For a constant offset the compensation takes the form of using new lagged products.

$$\hat{\rho}'_k = \hat{\rho}_k - \left[\frac{1}{N} \sum_{r=0}^{N-1} x_r \right]^2 \quad (12.1-37)$$

For a linear trend the data should be adjust via (12.1-23).

- (3) a. Weight the lags via time weights d_k and discrete transform them to obtain the spectrum estimates

$$\hat{P}_s(\omega_n) = \sum_{k=-(M-1)}^M d_k \hat{\rho}_k e^{-i \frac{2\pi}{2M} nk}, \quad \omega_n = \frac{2\pi}{2MT} n \quad (12.1-38)$$

with

$$n = 0, 1, \dots, M$$

or rather

b. From the raw estimates

$$V(\omega_n) = \sum_{k=-(M-1)}^M \hat{\rho}_k e^{-i \frac{2\pi}{2M} nk} \quad (12.1-39)$$

and frequency weight them by the simple formulae

$$\hat{P}_s(\omega_n) = a_1 \hat{V}(\omega_{n-1}) + a_0 \hat{V}(\omega_n) + a_1 \hat{V}(\omega_{n+1}) \quad (12.1-40)$$

for $1 \leq n \leq M-1$, and

$$\hat{P}_s(0) = \frac{1}{2} \hat{V}(0) + \frac{1}{2} \hat{V}(\omega_1)$$

$$\hat{P}_s(\omega_M) = \frac{1}{2} \hat{V}(\omega_{M-1}) + \frac{1}{2} \hat{V}(\omega_M) \quad (12.1-41)$$

Calculations may be made which indicate how finely the statistical spectrum can be resolved, namely the number of lags needed, and what kind of statistical stability can be expected for a given data record length. As might be expected both quantities can not be simultaneously be improved. One must be sacrificed for improvement in the other. Increasing the data record length, of course, allows improvement in both stability and resolution. Also, a calculation will be made of the correlation of adjacent estimates in frequency. Both the lagged product estimation and digital filter methods have similar developments in formulae.

Towards these ends the variance of the spectrum estimate about the mean at a given frequency is calculated and also the covariance of estimates at different frequencies as it produces information on how closely spaced estimates must be made. Of course, the variance is simply the covariance at identical frequencies so that computation of the covariance satisfies both needs. Unfortunately, the covariance computation is quite difficult to perform in general so that the assumption of a stationary, Gaussian process is used. In any case the calculation is quite tedious, but worthwhile. It will be given in some detail as other spectrum schemes that may be thought of usually involve similar calculations.

The covariance at a pair of estimates at frequencies ω_1 and ω_2 is given by the ensemble average

$$\begin{aligned} \text{Cov} [\hat{P}_s(\omega_1), \hat{P}_s^*(\omega_2)] &= \langle \hat{P}_s(\omega_1) \hat{P}_s^*(\omega_2) \rangle - \langle \hat{P}_s(\omega_1) \rangle \langle \hat{P}_s^*(\omega_2) \rangle \\ &= \sum_{k=-M}^M \sum_{j=-M}^M d_k d_j^* e^{-i\omega_1 kT} e^{i\omega_2 jT} \end{aligned} \quad (12.1-42)$$

$$\begin{aligned} &\times \left\{ \frac{1}{N-|k|} \frac{1}{N-|j|} \sum_{r=0}^{N-1-|k|} \sum_{s=0}^{N-1-|j|} [\langle x_r x_{r+|k|} x_s x_{s+|j|} \rangle - \langle x_r x_{r+|k|} \rangle \right. \\ &\quad \left. \cdot \langle x_s x_{s+|j|} \rangle] \right\} \end{aligned}$$

Now if a process X_p ; $p = 1, 2, 3, 4$ is assumed stationary and Gaussian it can be shown from the distribution function and its Fourier transform, the characteristic function, that

$$\langle X_1 X_2 X_3 X_4 \rangle = m_{12} m_{34} + m_{13} m_{24} + m_{14} m_{23} \quad (12.1-43)$$

where

$$m_{pq} = \langle X_p X_q \rangle$$

which is the correlation of X_p and X_q . Equating X_p to the sequence x_r for various values of r it may be seen that

$$m_{pq} = \varphi((p-q)T) \quad (12.1-44)$$

with $\varphi(rT)$ the autocorrelation function of the sequence. Therefore

$$\begin{aligned} & \langle x_r x_{r+|k|} x_s x_{s+|j|} \rangle - \langle x_r x_{r+|k|} \rangle \langle x_s x_{s+|j|} \rangle \\ &= \varphi(r-s)T \varphi((r-s+|k|-|j|)T) + \varphi((r-s-|j|)T) \varphi(r-s+|k|)T \end{aligned} \quad (12.1-45)$$

Writing $\varphi(rT)$ in terms of the sampled spectrum, using the property $\varphi(-rT) = \varphi(rT)$, and defining

$$G_N(\omega', \omega'') = \frac{1}{N-|k|} \frac{1}{N-|j|} \sum_{r=0}^{N-1-|k|} \sum_{s=0}^{N-1-|j|} e^{i(\omega' - \omega'')(r-s)T} \quad (12.1-46)$$

then the power estimate covariance may be written

$$\begin{aligned} & \text{Cov} \left[\hat{P}_s(\omega_1), \hat{P}_s^*(\omega_2) \right] \\ &= \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega' \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega'' \left[P_s(\omega') P_s^*(\omega'') G_N(\omega', \omega'') \right] \end{aligned} \quad (12.1-47)$$

$$x e^{-i\omega_1 k T} e^{i\omega_2 j T} \sum_{k=-M}^M \sum_{j=-M}^M d_k d_j \left(e^{-i\omega''(|k|-|j|)T} + e^{-i(\omega''|k|+\omega'|j|)T} \right)$$

While this seems to be an almost completely unmanageable expression a relatively simple approximation eases matters considerably. Examination of $G_N(\omega', \omega'')$ shows that, for $|k|, |j|$ much less than N , it will approach a sharply peaked sampling function. If the rest

of the integrand of (12.1-47) is "smooth" over the "peak" of $G_N(\omega', \omega'')$ then one of the integrations may be performed with ω'' simply equated to ω' . $G_N(\omega', \omega'')$ acts like a Dirac delta function. It must be true, in fact, that the sequence spectrum be relatively "smooth" with respect to the width of $G_N(\omega', \omega'')$ if decent spectrum estimates are to be made. The requisite property of $G_N(\omega', \omega'')$ follows from the closed form of the summation

$$\begin{aligned} \frac{1}{N} \sum_{r=0}^{N-1} e^{-i\omega rT} &= \frac{1}{N} \frac{1 - e^{-i\omega NT}}{1 - e^{-i\omega T}} \\ &= e^{-i\omega \left(\frac{N-1}{2}\right) T} \frac{\sin\left(\frac{\omega NT}{2}\right)}{N \sin\left(\frac{\omega T}{2}\right)} \end{aligned} \quad (12.1-48)$$

Examination of this expression shows that it is a sharply peaked function for large N with a width of the order of ω_s/N . In the limit as N becomes large

$$G_N(\omega', \omega'') \simeq \frac{\omega_s}{N} \delta(\omega' - \omega''), \quad \omega_s = \frac{2\pi}{T}. \quad (12.1-49)$$

Then (12.1-48) may be written

$$\begin{aligned} \text{Cov} \left[\hat{P}_s(\omega_1), \hat{P}_s^*(\omega_2) \right] &= \frac{1}{\omega_s N} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega' \left[P_s^2(\omega') e^{-i\omega_1 k T} e^{i\omega_2 j T} \right. \\ &\quad \left. \times \sum_{k=-M}^M \sum_{j=-M}^M d_k d_j (e^{-i\omega'(|k| - |j|)T} + e^{-i\omega'(|k| + |j|)T}) \right] \end{aligned} \quad (12.1-50a)$$

$$= \frac{1}{\omega_s N} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega' \left[P_s^2(\omega') e^{-i\omega_1 k T} e^{i\omega_2 j T} \right. \\ \left. \times 2 \sum_{k=-M}^M \sum_{j=-M}^M d_k d_j e^{-i\omega' |k| T} \cos(\omega' j T) \right]$$

A few more manipulations are necessary before interpretation may be made. By changing the variable of integration $\omega' \leftarrow -\omega'$ and using the evenness of the power spectrum

$$\text{Cov} [\hat{P}_s(\omega_1), \hat{P}_s^*(\omega_2)] \\ = \frac{1}{\omega_s N} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega' \left[P_s^2(\omega') e^{-i\omega_1 k T} e^{i\omega_2 j T} \right. \\ \left. \times 2 \sum_{k=-M}^M \sum_{j=-M}^M d_k d_j e^{i\omega' |k| T} \cos(\omega' j T) \right] \quad (12.1-50b)$$

By adding these two versions ((12.1-50a) and (12.1-50b)) of the covariance

$$\text{Cov} [\hat{P}_s(\omega_1), \hat{P}_s^*(\omega_2)] \\ = \frac{1}{\omega_s N} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega' \left[P_s^2(\omega') e^{-i\omega_1 k T} e^{i\omega_2 j T} \right. \\ \left. \times 2 \sum_{k=-M}^M \sum_{j=-M}^M d_k d_j \cos(\omega' j T) \cos(\omega' k T) \right] \quad (12.1-51)$$

Using the identity

$$2 \cos(\omega' jT) \cos(\omega' kT) = \frac{1}{2} \left[e^{i\omega'(j+k)T} + e^{-i\omega'(j+k)T} + e^{i\omega'(j-k)T} + e^{-i\omega'(j-k)T} \right] \quad (12.1-52)$$

and the definition (12.1-9) the covariance may be written, finally, as

$$\begin{aligned} \text{Cov} \left[\hat{P}_s(\omega_1), \hat{P}_s^*(\omega_2) \right] &= \frac{1}{\omega_s N} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s^2(\omega') \left[D(\omega' - \omega_2) D^*(\omega' - \omega_1) + D(\omega' - \omega_2) D(\omega' + \omega_1) \right] d\omega' \end{aligned} \quad (12.1-53)$$

where the identity $D^*(\omega) = D(-\omega)$ and a change of variables have also been used.

Now $D(\omega)$ is a sharply peaked function if the mean of the estimate is to approximate the actual spectrum (12.1-12). Then the second term in (12.1-53) is essentially zero except in the vicinity of $\omega_1 \simeq \omega_2 \simeq 0$ and $\omega_1 \simeq \omega_2 \simeq \omega_s/2$, zero and half the sampling frequency, respectively. As actual functions $D(\omega)$ are not zero outside the main peak at $\omega = 0$, some leakage occurs at other frequencies depending on the size of $P_s(\omega)$. The first term in (12.1-53) measures the interaction between estimates at different frequencies. When the separation between ω_1 and ω_2 increases the only interaction is due to leakage through the side-lobes. If $P_s(\omega_2)$, the power spectrum, is large outside of a given estimation frequency ω_1 then the covariance will also be large and the estimates at the two frequencies quite intertwined. However, when $P_s(\omega)$ is relatively smooth estimates at frequencies ω_1 and ω_2 are relatively independent when they are separated by more than the "effective width" of the frequency window $D(\omega)$. See (12.1-78).

From the previous discussion (12.1-53) may be reduced approximately, when the estimates about zero and half the sampling frequency are excluded for a moment, to

$$\text{Cov} \left[\hat{P}_s(\omega_1), \hat{P}_s^*(\omega_2) \right] = \frac{1}{\omega_s N} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s^2(\omega') D(\omega' - \omega_2) D^*(\omega' - \omega_1) d\omega' \quad (12.1-54)$$

The spectrum estimate variance at an arbitrary frequency ω is given, of course, by

$$\sigma^2 = \text{Cov} [P_s(\omega), P_s^*(\omega)] \quad (12.1-55)$$

$$= \frac{1}{\omega_s N} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s^2(\omega') |D(\omega' - \omega)|^2 d\omega'$$

or by a change of variables, also

$$\sigma^2 = \frac{1}{\omega_s N} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s^2(\omega + \omega') |D(\omega')|^2 d\omega' \quad (12.1-56a)$$

and

$$\sigma^2 = \frac{1}{\omega_s N} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s^2(\omega' - \omega) |D(\omega')|^2 d\omega' \quad (12.1-56b)$$

At zero and half the sampling frequency examination of (12.1-53) indicates that σ^2 is approximately twice as large.

The relative variation of the spectrum estimate may be defined

$$\eta = \frac{\sigma}{\langle \hat{P}_s(\omega) \rangle} \quad (12.1-57)$$

and is simply the ratio of the square root of the spectrum variance, the root-mean-square spectrum error, to the mean. If the r.m.s. error is small with respect to the mean spectrum value the relative variation is also small. Graphically it can be seen from figure 12.1-1 illustrating the distribution of estimates that as the r.m.s. error decreases the mean of the estimates is more clearly defined assuming a localized unimodal distribution function for the spectrum estimate. Section 12.5 discusses the actual distribution of the spectrum estimate with a more detailed interpretation of the relative variation η .

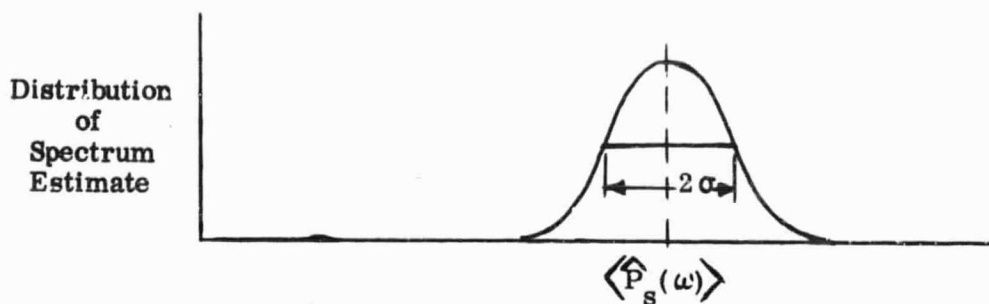


Figure 12.1-2. Relative Variation of Spectrum Estimates

From (12.1-8) and (12.1-55) the relative variation (squared) may be written as

$$\eta^2 = \frac{\omega_s}{N} \frac{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s^2(\omega') |D(\omega - \omega')|^2 d\omega'}{\left[\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s(\omega') D(\omega - \omega') d\omega' \right]^2} \quad (12.1-58)$$

This seemingly complex expression under appropriate approximation gives further information about $D(\omega')$ which in turn implies certain constraints on the number of lags M with respect to the number of data points N . It will be found that as less lags are used the relative variation becomes smaller and hence the statistical stability becomes greater; the spectrum estimate is more stable for less lags. However, as usual, a price must be paid for this stability. Resolution between adjacent spectrum estimates decreases with less lags as the function $D(\omega)$ broadens allowing contributions to an estimate from spectral components at larger spacings from each other.

Assume that $P(\omega)$ is reasonably constant in the vicinity of ω where an estimate is to be made, then as $D(\omega')$ becomes small outside of a region about $\omega = 0$, whose width will be discussed subsequently, then

$$\eta^2 = \frac{\omega_s}{N} \frac{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |D(\omega')|^2 d\omega'}{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} D(\omega') d\omega'} \quad (12.1-59)$$

from the periodicity of $D(\omega')$. Apparently the ratio

$$W_e = \frac{\left[\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} D(\omega') d\omega' \right]^2}{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |D(\omega')|^2 d\omega'} \quad (12.1-60)$$

has a lot to do with the efficacy of the estimation procedure. In terms of W_e

$$\begin{aligned} \eta^2 &= \frac{\omega_s}{N W_e} \\ &= \frac{2\pi}{N T W_e} \end{aligned} \quad (12.1-61)$$

Consider a perfectly rectangular, and unrealizable, shape for $D(\omega)$ for a moment:

$$D(\omega) = \begin{cases} G, & |\omega| \leq \frac{\Delta}{2} \\ 0, & \text{otherwise} \end{cases} \quad (12.1-62)$$

where Δ is the "ideal" bandwidth in radians per second of $D(\omega)$, then

$$W_e = \frac{A^2 \Delta^2}{A^2 \Delta} \quad (12.1-63)$$

Therefore, the nomenclature "equivalent width" for W_e is not undeserved. In this case

$$\begin{aligned} \eta^2 &= \frac{2\pi}{NT\Delta} \\ &= \frac{1}{NTf\Delta} \end{aligned} \quad (12.1-64)$$

where $f\Delta$ is the ideal $D(\omega)$ bandwidth in hertz. It can be seen that the product

$$\begin{aligned} f\Delta \eta^2 &= \frac{1}{NT} \\ &= \frac{1}{T_L} ; \quad T_L = \text{total time length of record in seconds} \end{aligned} \quad (12.1-65)$$

is determined by the total record time. Also statistical stability only increases with the square root of the record length. To increase stability by a factor of 2, the record must be 4 times as long. Faster sampling to obtain more samples does not buy better resolution or statistical stability as might be naively thought at first as the total record length in time is the basic determinant of stability and resolution. $f\Delta$ is essentially the frequency resolution. As T_L becomes large the relative accuracy and resolution both can become smaller and, hence, better. Of course, resolution may be traded for statistical stability.

Relation (12.1-65) may be thought of as a "conservation" law. Resolution and statistical stability are inextricably bound together. For a given record length increased resolution can be obtained at the expense of a loss in statistical stability and conversely.

It is quite useful to be able to express the equivalent width for a general $D(\omega)$ in terms of the lag weights d_k . This provides a means for expressing the equivalent width in terms of the number of lags for some particular $D(\omega)$'s.

Using $D(\omega)$ in terms of the d_k 's (12.1-9) the equivalent width may be expressed as

$$W_e = \frac{\omega_s}{2} \frac{2d_o^2}{\sum_{k=-(M-1)} d_k^2}$$

The fraction

$$\frac{2 d_0^2}{\sum_{k=-(M-1)}^M d_k^2} \quad (12.1-67)$$

expresses the portion of the total possible region, $\omega_s/2$, over which the spectrum may be measured.

When the uniformly weighted estimate is used all the $d_k = 1$ and

$$W_e = \frac{1}{M} \frac{\omega_s}{2} \quad (12.1-68)$$

the equivalent width is $1/M$ of the total frequency interval. For this case the relative stability and data quantity is related by

$$\eta^2 = \frac{2M}{N} \quad (12.1-69)$$

which indicates that less lags produce better statistical stability as expected.

"Hanning" has weights given by (12.1-33),

$$d_k = \frac{1}{2} \left(1 + \cos \left(\frac{\pi k}{M} \right) \right) \quad (12.1-70)$$

which produce an equivalent width

$$W_e \approx \frac{8}{3M} \frac{\omega_s}{2} \quad (12.1-71)$$

which is more than twice as wide as for uniform weighting. However, the large negative side-lobes either side of the main peak of the uniform weighting window tend to artificially reduce the effective width. Even though the resolution is diminished the "hanning" does produce significantly lower side-lobes than the uniform weighting. Side-lobe leakage with uniform weighting is quite severe. Furthermore the rate of side-lobe fall-off with "hanning" is proportional to the inverse cube of frequency whereas uniform weighting only falls off inversely with frequency. Finally, as will be seen subsequently, the correlation between adjacent estimates is also insignificant with "hanning".

Therefore, even though "hanning" does effect the balance between data quantity and stability somewhat adversely it is preferable as leakage from other frequencies is minimized allowing a more correct estimate of the actual spectrum. In fact stability is only affected by a factor of $\sqrt{8/3} = 1.64$.

For "hanning" the relative stability is related to data quantity by

$$\eta^2 = \frac{3}{4} \frac{M}{N} \quad (12.1-72)$$

As the previous discussion may indicate, the choice of windows is somewhat arbitrary. The literature of the present and past discusses window choices by many criteria however "hanning" seems to be a good compromise and extremely easy to use. A first choice of the antenna designer might be Tchebychev and/or Taylor weights but for a given resolution they have a relatively large average side-lobe leakage and are also more complex to use.

Another point of interest is the statistical dependence of adjacent estimates which affects the computational volume of estimates needed. Independence of adjacent estimates, especially when they are approximately Gaussianly distributed is best estimated by the correlation coefficient (Section 11.1.3):

$$\rho = \frac{\text{Cov} [\hat{P}_s(\omega_1), \hat{P}_s^*(\omega_2)]}{\sqrt{\text{Cov} [\hat{P}_s(\omega_1), \hat{P}_s^*(\omega_1)] \text{Cov} [\hat{P}_s(\omega_2), \hat{P}_s^*(\omega_2)]}} \quad (12.1-73)$$

Assuming, as usual, $P_s(\omega)$ "smooth" over a region of frequency encompassing adjacent estimates then from (12.1-54)

$$\rho = \frac{\frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} D(\omega' - \omega_2) D^*(\omega' - \omega_1) d\omega'}{\frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |D(\omega')|^2 d\omega'} \quad (12.1-74)$$

neglecting estimates at zero and half the sampling frequency. Using (12.1-9)

$$\rho = \frac{\sum_{k=-(M-1)}^M d_k^2 e^{-i(\omega_1 - \omega_2) kT}}{\sum_{k=-(M-1)}^M d_k^2} \quad (12.1-75)$$

If uniform weighting is used and estimates at frequencies (12.1-18)

$$\begin{aligned}\omega_1 &= \omega_m \\ &= \frac{m}{2M} \frac{2\pi}{T}\end{aligned}\quad (12.1-76)$$

and

$$\begin{aligned}\omega_2 &= \omega_n \\ &= \frac{n}{2M} \frac{2\pi}{T}\end{aligned}$$

are desired then

$$\rho = \begin{cases} 1, & n = m \\ 0, & n \neq m \end{cases} \quad (12.1-77)$$

as expected. Estimates are uncorrelated at adjacent frequencies with uniform weighting, at least to the order of second moments.

For "hanning" weights

$$\rho = \delta_{n,m} + \frac{1}{3} (\delta_{n,m+1} + \delta_{n,m-1}) + \frac{1}{6} (\delta_{n,m+2} + \delta_{n,m-2}) \quad (12.1-78)$$

where $\delta_{m,n}$ is the Kronecker delta. With these weights it can be seen that the correlation coefficient is virtually insignificant at adjacent estimates, $\rho = 1/3$, and insignificant at estimates displaced by 2 frequencies ($\rho = 1/6$).

Correlation, then, can essentially be omitted from consideration for estimates displaced $\Delta \omega = \frac{1}{M} \frac{\omega_s}{2}$ apart. Closer estimates become highly correlated making evaluation of the spectrum more frequently a redundant, and computationally wasteful process. The discrete Fourier transform is, then, certainly adequate

12.1.1 Use of the Fast Fourier Transform Algorithm to Reduce Computation Time

By use of the fast transform algorithm (Section 14) auto-correlation (lags) and cross-correlation estimates may be computed in a faster manner for a significant number of points x_r (y_r) than by direct formation of the products. The procedure for gaining this increased computational efficiency is quite similar to that for computing convolutions (Section 3.7.5) as the convolution and cross-correlation relations are quite similar (Section 2.8). Auto-correlation is simply a special case of cross-correlation whereby both sequences are identical.

The discrete cross-correlation relation (Section 2.8) is

$$z_k = \sum_{r=0}^{L-1} x_{r+k} \bar{y}_r^* \quad (12.1.1-1)$$

$$= \frac{1}{N} \sum_{n=0}^{L-1} \bar{X}_n \bar{Y}_n^* w^{nk}, \quad w = e^{i \frac{2\pi}{L}}$$

and $\bar{X}_n$ and $\bar{Y}_n$ are the L point transforms of the $\bar{x}_r$ and $\bar{y}_r$, respectively. Unfortunately, the indices in (12.1.1-1) are interpreted cyclically or modulo L which means that z_k is not, in general, equal to the desired cross-correlation (12.1-2) where indices are not interpreted modulo N . Surmounting of this problem is accomplished by appending zeroes to the N point ($N \leq L$) sequences x_r and y_r to form the L point sequences $\bar{x}_r$ and $\bar{y}_r$.

Once the sequences $\bar{x}_r$ and $\bar{y}_r$ are formed their discrete transforms (DFT1) are taken and multiplied at corresponding frequencies (n), and the inverse transform (IDFT1) taken. Depending upon L a number of lags are found.

To determine the appropriate value of L so that the desired $M + 1$ lags for use in the spectrum estimation formula (12.1-5) may be computed write (12.1.1-1) in the form

$$z_k = \sum_{r=0}^{N-1-k} \bar{y}_r^* \bar{x}_{r+k} + \sum_{r=N-k}^{L-1} \bar{y}_r^* \bar{x}_{r+k} \quad (12.1.1-2)$$

If $k = 0$ and the first lag (within a factor of $1/N$) is desired then L may be set equal to N and no zeroes appended. Only the first lag $k = 0$ is valid in this case as the cyclic term, the second in (12.1.1-2), produces incorrect results. For lags of $k = 0$, and $k = 1$, it can be seen that L should be set equal to $N + 1$ and one zero should be appended; $\bar{y}_N = \bar{x}_N = 0$ to remove the second term.

In general, then $\bar{x}_r$ and $\bar{y}_r$ are set exactly equal to the sequences x_r and y_r for the first values ($r = 0, 1, \dots, N - 1$). If $M + 1$ lags are desired then M zeroes should be appended.

Only the first $M + 1$ products should be accepted as correct. Of course, they must be modified by the normalizing factor $1/(N - |k|)$ after they are obtained.

The procedure for obtaining the lags is now clear. An $L = M + N$ point discrete Fourier transform must be taken of each of the real N point sequences x_r and y_r appended with M zeroes. These two transforms are then multiplied frequency by frequency, after taking the conjugate of one ($\bar{Y}_n$). This new sequence is then inversely discrete transformed. The first $M + 1$ members of this transform are the desired correlation products.

Recognition of certain relations for real sequence transforms allows only one application of the fast Fourier transform algorithm to compute the transforms of both x_r and y_r .

From Section 2.6.4 it is seen that

$$\bar{X}_n = \bar{X}_{L-n}^* \quad (12.1.1-3)$$

and

$$\bar{Y}_n = \bar{Y}_{L-n}^* \quad (12.1.1-3)$$

for real sequences $\bar{x}_r$ and $\bar{y}_r$.

Therefore, if the transform of the complex sequence is taken:

$$U_n = \sum_{r=0}^{L-1} (\bar{x}_r - i \bar{y}_r) W^{-rn} \quad (12.1.1-4)$$

Use of (12.1.1-3) produces

$$\begin{aligned} U_{L-n}^* &= \bar{X}_{L-n}^* + i \bar{Y}_{L-n}^* \\ &= \bar{X}_n - i \bar{Y}_n \end{aligned} \quad (12.1.1-5)$$

which results in

$$\begin{aligned} \bar{X}_n &= \frac{1}{2} [U_n + U_{L-n}^*] \\ \bar{Y}_n &= \frac{1}{2i} [U_n - U_{L-n}^*] \end{aligned} \quad (12.1.1-6)$$

It is quite easy to compute the two transforms from the one complex transform.

A comparison of minimum computing times for the direct method versus the transform method assuming special purpose hardware, is certainly of interest. General purpose computer programs are generally quite variable in computing times depending upon the machine instruction repertoire and the individual programmer.

For the direct method the computation time, TC_D , is approximately

$$TC_D = \frac{(2N-M)(M+1)}{2} (\alpha + \beta) \quad (12.1.1-7)$$

where

α = time per real addition in seconds.

β = time per real multiplication in seconds.

The base 2 fast transform method requires $2 L \log_2 L$ real multiplications and $3 L \log_2 L$ real additions with a total transform time of $(3 \alpha + 2 \beta) L \log_2 L$ seconds.

Application of the transform twice is necessary plus about $2L$ more additions are required when computing two transforms simultaneously for cross-correlations. The total time by the transform method, TC_T , is approximately, then

$$\begin{aligned} TC_T &= \left[2 (3 \alpha + 2 \beta) \log_2 L + 2 \alpha \right] L \\ &= \left[2 (3 \alpha + 2 \beta) \log_2 (M+N) + 2 \alpha \right] (M+N) \end{aligned} \quad (12.1.1-8)$$

The efficiency ratio is

$$\begin{aligned} E &= \frac{TC_D}{TC_T} \\ &= \frac{(2N-M)(M+1)(\alpha+\beta)}{2 \left[2 (3 \alpha + 2 \beta) \log_2 (M+N) + 2 \alpha \right] (M+N)} \end{aligned} \quad (12.1.1-9)$$

When the addition and multiplication times are equal, the ratio becomes

$$E = \frac{(2N-M)(M+1)}{[10 \log_2 (M+N) + 2] (M+N)} \quad (12.1.1-10)$$

For a small lag to data point ratio $(M/N) \ll 1$:

$$E \approx \frac{M+1}{5 \log_2 (M+N) + 1} \quad (12.1.1-11)$$

Examination of this formula quickly produces the observation, that when large quantities of data are available, the transform method should be used. For example, assume that 924 data points are available and 100 lags are desired, then E is approximately 2. It would take twice as long by the direct method. Maintaining a similar lag to data ratio but increasing the number of data points shows the real power of the transform method. Assuming 800 lags and 7392 data points then E is approximately 12; the transform method is faster by a factor of about 12.

12.2 ESTIMATION THROUGH THE USE OF DIGITAL FILTERS

Analog filters that possess narrow pass-bands have been used for many years to estimate the spectrum of stationary continuous processes with a resolution approximately equal to the bandwidth of the filter. The scheme is quite simple in concept though certain fundamental questions arise, as in any method, as to accuracy and statistical stability of the estimates.

The basic procedure used frequently by electrical engineers, utilizing band-pass filters for spectrum measurement, is pictorialized below in Figure 12.2-1.

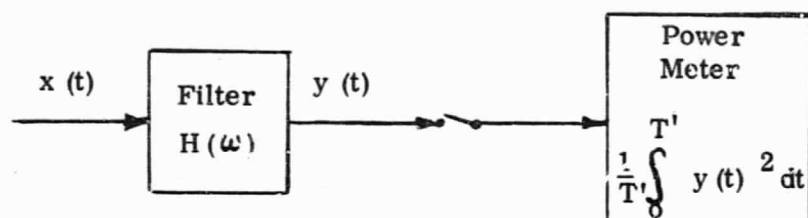


Figure 12.2-1. Spectrum Measurement by Analog Filtering

The time function $x(t)$ is passed through a sharply tuned bandpass filter, $H(\omega)$, whose center frequency, ω_o , is the frequency at which the spectrum value is desired. The bandwidth of the filter is the approximate resolution with which the spectrum can be determined. This can be seen from (11.1.11-4) where the output power P_y of the filter output $y(t)$ is related to the input power spectrum $P_x(\omega)$ by

$$P_y = \int_{-\infty}^{\infty} |H(\omega)|^2 P_x(\omega) d\omega \quad (11.1.11-4)$$

If $H(\omega)$ is localized:

$$H(\omega) = \begin{cases} \frac{1}{\sqrt{2\Delta}}, & \omega_o - \frac{\Delta}{2} \leq \omega \leq \omega_o + \frac{\Delta}{2}, \\ 0, & \text{otherwise,} \end{cases} \quad (12.2-1)$$

where Δ is the filter bandwidth in radians per second, then

$$P_y = \frac{1}{\Delta} \int_{\omega_o - \frac{\Delta}{2}}^{\omega_o + \frac{\Delta}{2}} P_x(\omega) d\omega \quad (12.2-2)$$

where an infinite integration time in the power meter has been assumed. Formula (12.2-2) demonstrates that the filter output is an average about ω_o of $P_x(\omega)$ which is the actual power spectrum. If $P_x(\omega)$ is "smooth" and doesn't vary much over the region

$\omega_o - \frac{\Delta}{2} \leq \omega \leq \omega_o + \frac{\Delta}{2}$ then P_y is a good approximation to $P_x(\omega_o)$.

The switch in the pictorialization was included, as the filter output should not be measured until the initial filter transient has decayed.

Ensuring integration of the squared output for a long enough period of time T' , is of importance as in practice a finite integrating time is actually used. Section 11.1.10 requires that $T' > \frac{2\pi}{R}$ where R in radians per second is the "smearing" of the transfer function. R should obviously be made smaller than the filter bandwidth Δ . Integration time should also be long enough to reduce statistical fluctuations in the estimate of $P_x(\omega_0)$.

Digital bandpass filters, of course, can also be used to estimate stationary sampled data spectrum. Obviously estimation of a continuous spectrum through a sampled data spectrum can be obtained only if the sampling frequency is high enough to avoid "aliasing".

The matters of frequency windows and statistical stability of spectrum estimates by digital filters follow quite closely those for the lagged product estimation method. Certain computational aspects, however, should always be taken into consideration when using digital filters. Recursive and non-recursive filters while usually supplying an easy method for implementation of spectrum analyses often require a good deal more arithmetic operations to obtain the same accuracy of estimation as by the lagged method with the fast Fourier transform. Recursive filters, at first glance, might seem to be faster as they have relatively few weights for the two or four pole filters needed. Unfortunately, this computation must be performed for every data point. Non-recursive filters, on the other hand, do not use past outputs and as the bandwidth is usually small for spectrum estimation the output may be decimated in time by a fraction of the reciprocal of the bandwidth without causing significant aliasing. As variations in procedures are quite different depending on applications no comparison of these filter methods will be made with the lagged product method as far as computation time is concerned.

Consider, now, as the spectrum estimate the squared output of a digital filter $H_z(\omega) = H_z(z)$ in abbreviated notation;

$$P_s(\omega) = \frac{1}{N} \sum_{r=0}^{N-1} y_r^2 \quad (12.2-3)$$

where y_k is the output of the digital filter and N is the number of input samples used. The filtered output is not used until the initial transient has decayed sufficiently (Section 3.10).

By taking certain averages, the ensemble average of $P_s(\omega)$ and its variance, the desirable properties of the filter $H_z(\omega)$ may be inferred. The filter characteristic will be seen to have much in common with the lagged product window $D(\omega)$.

Ensemble averaging

$$\langle P_s(\omega) \rangle = \langle y_r^2 \rangle \quad (12.2-4)$$

assuming the sequence y_r to be stationary. If the sequence is also ergodic to at least second moments then $\langle y_r^2 \rangle$ can be equated to the average over all y_r squared and the average sequence power obtained (Section 11.2.6)

$$\langle P_s(\omega) \rangle = \frac{1}{2\pi i} \oint_{\text{unit circle}} H_z(z) H\left(\frac{1}{z}\right) P_s(z) \frac{dz}{z} \quad (12.2-5)$$

in z transform notation, or

$$\langle P_s(\omega) \rangle = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |H_z(\omega')|^2 P_s(\omega') d\omega' \quad (12.2-6)$$

in frequency notation.

As in the case of "lagged product" estimation consider the case where $P_s(\omega')$ varies slowly about the frequency of estimation, ω , and the filter $H_z(\omega')$ to be bandpass filter whose center is ω . Then

$$\langle P_s(\omega) \rangle = P_s(\omega) \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |H_z(\omega')|^2 d\omega' \quad (12.2-7)$$

which is remarkably similar to the expression (12.1-3) except that $H_z(\omega')$ must be peaked at both plus and minus ω and $H_z(-\omega) = H_z^*(\omega)$. Obviously, then the digital filter transfer function must possess the property

$$\frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |H_z(\omega')|^2 d\omega' = 1 \quad (12.2-8)$$

In the limit of long data records where statistical stability is easy to come by a desirable attribute is seen to be

$$|H_z(\omega')|^2 = \frac{\omega_s}{2} [\delta(\omega - \omega') + \delta(\omega + \omega')] \quad (12.2-9)$$

As $P_s(\omega)$ is an even function the delta function at $-\omega$ causes no problems. It should be noted, however, that the gain of $|H_z(\omega')|^2$ is only half that of the lagged product window $D(\omega)$ at the peak.

If a single tone at frequency ω_0 and amplitude A (12.1-14) is applied then

$$\langle P_s(\omega) \rangle = \frac{A^2}{2} |H_z(\omega)|^2 \quad (12.2-10)$$

For this case too (12.2-9) must hold in the limit of narrow bandwidths.

As an example adjust $H_z(\omega')$ so that it has a narrow bandwidth and approximates

$$H_z(\omega') = \begin{cases} \sqrt{\frac{1}{2\Delta}} \frac{1}{\omega_s}, & \omega - \frac{\Delta}{2} \leq |\omega'| \leq \omega + \frac{\Delta}{2} \\ 0, & \text{otherwise.} \end{cases} \quad (12.2-11)$$

Then,

$$\langle P_s(\omega) \rangle = \frac{1}{\Delta} \int_{\omega - \frac{\Delta}{2}}^{\omega + \frac{\Delta}{2}} P_s(\omega') d\omega' \quad (12.2-12)$$

as expected. Again, if $P_s(\omega)$ is "smooth" around ω then, the ensemble average of the estimate will effectively be the actual spectrum of $P_s(\omega)$. As may be seen the filter amplitude develops a larger peak as the bandwidth narrows. A reasonable value for the actual filter bandwidth is the 3 decibel value and the peak filter gain may be set to that given in (12.2-11).

As with the analog method, allowance must be made for the transient to decay before averaging the square of the filter output. Further, enough samples must be allowed:

$NT > \frac{2\pi}{R}$, with T the sampling interval, such that R , the frequency spectrum "smearing" in radians per second due to a finite number of samples is significantly less than the filter bandwidth.

To perform spectrum analyses all of the techniques discussed previously for synthesizing bandpass filters are available.

For the digital filter spectrum estimation method the development of the covariance and relative stability are somewhat simpler than that of the lagged product method but results and interpretations are quite similar.

The covariance of a pair of estimates at frequencies ω_1 and ω_2 is given by

$$\begin{aligned} \text{Cov} [\hat{P}_s(\omega_1), \hat{P}_s^*(\omega_2)] &= \langle \hat{P}_s(\omega_1) \hat{P}_s^*(\omega_2) \rangle - \langle \hat{P}_s(\omega_1) \rangle \langle \hat{P}_s^*(\omega_2) \rangle \\ &= \frac{1}{N^2} \sum_{r=0}^{N-1} \sum_{s=0}^{N-1} \left[\langle y_{1,r}^2 y_{2,s}^2 \rangle - \langle y_{1,r}^2 \rangle \langle y_{2,s}^2 \rangle \right] \end{aligned} \quad (12.2-13)$$

where the front subscripts on the y's indicate the two filter outputs.

Using (12.1-43) the brackets may be written, assuming stationary Gaussian processes

$$\langle y_{1,r}^2 y_{2,s}^2 \rangle - \langle y_{1,r}^2 \rangle \langle y_{2,s}^2 \rangle = 2 \langle y_{1,r} y_{2,s} \rangle^2 \quad (12.2-14)$$

Using the convolution form of the digital filter input-output relation (Section 3.4)

$$\langle y_{1,r} y_{2,s} \rangle = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} w_k^{(1)} w_j^{(2)} \langle x_{r-k} x_{s-j} \rangle \quad (12.2-15)$$

where $w_k^{(1)}$ and $w_j^{(2)}$ are the impulse responses of the two estimating filters $H_z^{(1)}(\omega)$ and $H_z^{(2)}(\omega)$, respectively. Assuming x_k stationary

$$\varphi((r-s-k+j)T) = \langle x_{r-k} x_{s-j} \rangle, \quad (12.2-16)$$

the auto-correlation function. Writing the auto-correlation function in terms of the spectrum

$$\langle y_{1,r} y_{2,s} \rangle = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega' P_s(\omega') H_z^{(1)}(\omega') H_z^{(2)}(\omega') e^{i\omega'(r-s)T} \quad (12.2-17)$$

where the definition (Section 3.4) of the transfer function has been used. Therefore

$$\begin{aligned} \text{Cov} [\hat{P}_s(\omega_1), \hat{P}_s^*(\omega_2)] \\ = \frac{2}{\omega_s N} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega' \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega'' P_s(\omega') P_s^*(\omega'') H_z^{(1)}(\omega') H_z^{(1)*}(\omega'') H_z^{(2)}(\omega') H_z^{(2)*}(\omega'') G(\omega', \omega'') \end{aligned} \quad (12.2-18)$$

with

$$G(\omega', \omega'') = \frac{1}{N\omega_s} \left| \sum_{r=0}^{N-1} e^{i(\omega' - \omega'')r} \right|^2 \quad (12.2-19)$$

which approaches a delta function for large N

$$G(\omega', \omega'') \Rightarrow \delta(\omega' - \omega'')$$

With this approximation

$$\text{Cov} [\hat{P}_s(\omega_1), \hat{P}_s^*(\omega_2)] = \frac{2}{N\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega' P_s^2(\omega') |H_z^{(1)}(\omega')|^2 |H_z^{(2)}(\omega')|^2 \quad (12.2-20)$$

which is remarkably similar to the lagged product formula (12.1-54).

The variance at a given frequency ω is then

$$\sigma^2 = \frac{2}{N\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega' P_s^2(\omega') |H_z(\omega')|^4 \quad (12.2-21)$$

where the superscript on $H_z(\omega')$ has been removed as a single frequency is being considered. The relative variation squared is then

$$\eta^2 = \frac{2\omega_s}{N} \frac{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s^{(2)}(\omega') |H_z(\omega')|^4 d\omega'}{\left[\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s(\omega') |H_z(\omega')|^2 d\omega' \right]^2} \quad (12.2-22)$$

Again as for the lagged product method assume that $P_s(\omega')$ is relatively smooth with respect to the peakedness of $H_z(\omega')$ at ω , then approximately

$$\eta^2 = \frac{\omega_s}{NW_e'} \quad (12.2-23)$$

with W_e' the effective width for the digital filter method

$$W_e' = \frac{1}{2} \frac{\left[\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |H_z(\omega')|^2 d\omega' \right]^2}{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |H_z(\omega')|^4 d\omega} \quad (12.2-24)$$

This definition of W_e' allows the relative variation of the digital filter method to be written in a form identical to that for the lagged product method. The factor of 1/2 may appear disturbing but the fact that $H_z(\omega')$ has peaks at both plus and minus ω , the estimation frequency, compensates for it.

For example, consider $H_z(\omega')$ as an ideal filter of bandwidth Δ in radians per second and gain A

$$H_z(\omega') = \begin{cases} A, & \omega - \frac{\Delta}{2} \leq |\omega'| \leq \omega + \frac{\Delta}{2} \\ 0, & \text{otherwise} \end{cases} \quad (12.2-25)$$

then

$$\begin{aligned} W_e' &= \frac{1}{2} \frac{A^4}{A^4} \frac{(2\Delta)^2}{2\Delta} \\ &= \Delta \end{aligned} \quad (12.2-26)$$

The effective width W_e' is actually equal to the ideal filter bandwidth. The relative variation, can be interpreted in a manner identical to that for the lagged product method as in Section 12.5.

Obviously, narrow effective bandwidths produce less stable estimates than wide ones. Instead of considering lags to produce a desired resolution bandwidth, the digital filter and its effective bandwidth may be considered directly. As with the frequency window $D(\omega)$, a digital filter is desired with a fall off outside of the effective width that will not allow leakage of other frequencies into the estimate. Fortunately, the rate of fall-off of $|H_z(\omega')|^2$ outside of the center frequency ω is the concern rather than the fall-off of $|H_z(\omega')|$. This allows filters, recursive and non-recursive, with fewer coefficients. Recursive filters with at most 2 to 4 poles in their transfer functions will be found to be quite adequate.

Due to the fourth power of $|H_z(\omega')|$ involved in (12.2-24) the expression of W_e' in terms of the impulse response is not as useful as the expression in terms of z transforms

$$2W_e' = \frac{\omega \oint H_z(z) H_z\left(\frac{1}{z}\right) \frac{dz}{z}}{\oint H_z^2(z) H_z^2\left(\frac{1}{z}\right) \frac{dz}{z}} \quad (12.2-27)$$

where the contour integral is taken to be the unit circle. For recursive filters the use of residue theory of complex variable theory is helpful. However, even for the simplest filters (12.2-27) is quite difficult to evaluate. A useful and adequate procedure is to use two, and preferably four, pole bandpass filters of, say, the Butterworth type generated by the bilinear z transform (Section 6.6) with the 3 decibel bandwidth set equal to $W_e'/2$.

It may be noted that the estimation at different frequencies involves different filters which will not have identical shapes, though the effective widths can be made identical. A means of by-passing this problem, though it is hardly a severe one, is the use of complex demodulation (Sections 8.2 and 15) which is useful, in itself, as it requires only one filter design. The next section is devoted to this method.

12.3 COMPLEX DEMODULATE SPECTRUM ESTIMATION

Sections 8.2 and 15.1 discuss decomposition of a sequence into complex demodulates and practical means of obtaining them. From the developments of these sections it would appear that "mixing" with a frequency ω and low pass filtering might give a spectrum estimate if the magnitude squared of the complex output is used appropriately. Therefore, take as a spectrum estimate

$$\hat{P}_s(\omega) = \frac{1}{N} \sum_{r=0}^{N-1} |v_r|^2 \quad (12.3-1)$$

where v_r is the complex output of a pair of identical filters operating on the "mixed" signals in figure 12.3-1. Of course the output v_r is not accepted until the initial transient decays.

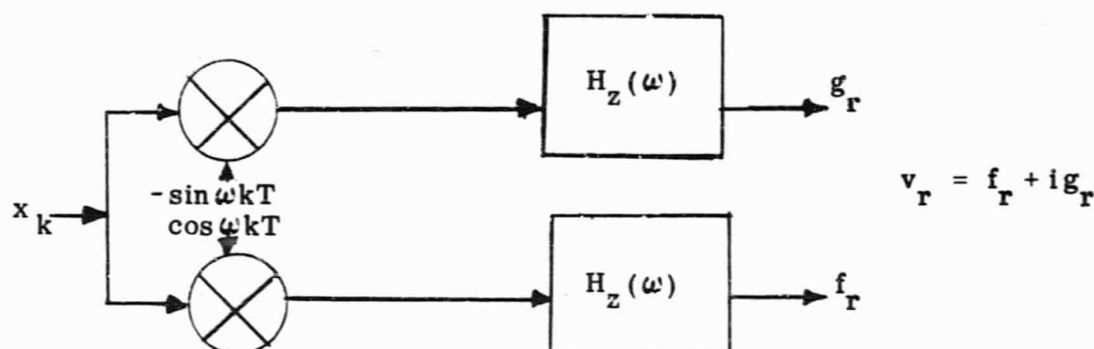


Figure 12.3-1 Complex Demodulation Spectrum Estimation

In terms of the impulse response of the "real" low pass filter $H_z(\omega)$, w_k :

$$\hat{P}_s(\omega) = \frac{1}{N} \sum_{r=0}^{N-1} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} w_k w_j x_{r-k} x_{r-j} e^{-i\omega(k-j)T} \quad (12.3-2)$$

By a methodology almost identical to that for the bandpass filter estimation technique the ensemble averaged spectrum is

$$\langle \hat{P}_s(\omega) \rangle = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s(\omega') |H_z(\omega - \omega')|^2 d\omega' \quad (12.3-3)$$

if the process x_r is assumed stationary. The low pass filtering after the mixing provides a "window" at the frequency ω through which estimates are made. The properties of

$|H_z(\omega')|^2$ apparently should be identical to those of $D(\omega)$ of lagged produce estimation. It should be noted that both positive and negative frequencies are not involved as with simple bandpass filtering.

The variance is also of interest for this technique. It is

$$\sigma^2 = \frac{1}{N} \omega_s \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s(\omega')^2 \left[|H_z(\omega - \omega')|^4 + |H_z(\omega - \omega') H_z(\omega + \omega')|^2 \right] d\omega' \quad (12.3-4)$$

for Gaussian processes and N large. It is also virtually identical to the lagged product method with $D(\omega)$ equal to $|H_z(\omega)|^2$. With a narrow bandwidth for $H_z(\omega')$ the second term in the brackets (12.3-4) contributes a negligible amount of frequencies other than around zero and half the sampling frequency and therefore approximately:

$$\sigma^2 = \frac{1}{N} \omega_s \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s(\omega')^2 |H_z(\omega - \omega')|^4 d\omega' \quad (12.3-5)$$

The relative variation is

$$\eta^2 = \frac{\omega_s}{N} \frac{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s(\omega')^2 |H_z(\omega - \omega')|^4 d\omega'}{\left[\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} P_s(\omega') |H_z(\omega - \omega')|^2 d\omega' \right]^2} \quad (12.3-6)$$

or assuming $P_s(\omega')$ "smooth" as done previously

$$\eta^2 = \frac{\omega_s}{N W_e''}$$

with the effective width of $H_z(\omega)$ for the complex demodulate method given by

$$W_e'' = \frac{\left[\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |H_z(\omega')|^2 d\omega' \right]^2}{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |H_z(\omega')|^4 d\omega'} \quad (12.3-8)$$

Certain practical aspects on the complex demodulate filtering method may be noted. The fast discrete Fourier transform realization of the filter $H_z(\omega)$ may be used if $H_z(\omega)$ is non-recursive (Section 3.7.5). Secondly, the sines and cosines that are mixed with the input sequence x_r may be computed recursively with but a few operations (Section 15.2) though occasionally for long sequences the amplitude of the wave should be renormalized to unity due to round-off.

As the transfer function of the digital filter is involved in the ensemble averaged spectrum quadratically a simply one pole low pass filter will probably be adequate in many applications. The bi-linear z transform of a simple low pass Butterworth filter given in Section 6.6, is a reasonable recursive digital filter to use. If longer computation times are not objectionable a two pole filter may be used. For the Butterworth filter of Section 6.6

$$H_z(z) = \frac{\alpha(1+z^{-1})}{1-\beta z^{-1}} \quad 12.3-9$$

with

$$\alpha = \frac{\omega_0}{1+\omega_0}$$

$$\beta = \frac{1-\omega_0}{1+\omega_0}$$

and

$$\omega_0 = \tan\left(\frac{\Omega_0 T}{2}\right)$$

Ω_0 is the digital cut-off decibel frequency.

Now using residues

$$\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |H_z(\omega')|^2 d\omega' = \omega_s \alpha \quad (12.3-10a)$$

and

$$\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |H_z(\omega')|^4 d\omega' = \omega_s \frac{\alpha}{4} \beta^2 (3 - \beta) \quad (12.3-10b)$$

Then the effective width becomes

$$W_e'' = \frac{4\omega_0 (1 + \omega_0)^2 \omega_s}{(2 + \omega_0)(1 - \omega_0)^2} \text{ radians/second,} \quad (12.3-11)$$

and from (12.3-9)

$$\omega_0 \approx \frac{\pi \Omega_0}{\omega_s} \quad (12.3-12)$$

for small digital bandwidth to sampling frequency ratios, which is usually the practical situation..

Therefore W_e'' can be approximated by

$$W_e'' \approx 2\pi \Omega_0 \quad (12.3-13)$$

Now Ω_0 is the low frequency filter 3 decibel (db.) bandwidth in radians per second from zero frequency. When "translated" by the mixing operation (See Section 15.2) the new 3 db bandwidth, Δ_e , is twice the low frequency bandwidth

$$\Delta_e = 2 \Omega_0 \text{ radians/second.} \quad (12.3-14)$$

In these terms the effective bandwidth is given by

$$W_e'' = \pi \Delta_e \quad (12.3-15)$$

The effective bandwidth is a factor of π larger than the 3 db bandwidth of the translated low frequency filter. Of course, the reason for this increase is the finite ratio of fall off of the filter, the 3 decibel width is an arbitrary measure of bandwidth. It may be noticed that π is not significantly different than the 8/3 factor of "hanning" (12.1-34).

In terms of this particular low-pass filter effective width the relative stability is given by

$$\eta^2 = \frac{2}{NT\Delta_e} \quad (12.3-16a)$$

or in terms of the low frequency bandwidth

$$\eta^2 = \frac{1}{NT\Omega_0} \quad (12.3-16b)$$

When Δ_e is set equal to $\omega_s/(2M)$, namely the same number of spectrum estimates are made.

12.4 LIMITATIONS OF A PERIODOGRAM ESTIMATE

As mentioned previously, the periodogram would be a highly unstable method of estimation. It is quite easy to find the relative stability of a periodogram with the methods used previous to this point.

The periodogram estimate using all the data points is

$$\hat{P}_s(\omega) = \frac{1}{N} \left| \sum_{k=0}^{N-1} e^{-i\omega kT} \right|^2 \quad (12.4-1)$$

The ensemble average is

$$\langle \hat{P}_s(\omega) \rangle = \frac{1}{N} \sum_{k=0}^{N-1} \sum_{j=0}^{N-1} \langle x_j x_k \rangle e^{-i\omega(k-j)T} \quad (12.4-2)$$

Assuming x_k stationary then

$$\langle \hat{P}_s(\omega) \rangle = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega' P_s(\omega') \frac{1}{N} \left| \sum_{k=0}^{N-1} e^{i(\omega' - \omega)kT} \right|^2 d\omega \quad (12.4-3)$$

which for large N becomes

$$\langle \hat{P}_s(\omega) \rangle \approx P_s(\omega). \quad (12.4-4)$$

With a large number of data samples, N, the estimate mean approaches the actual spectrum.

However, when the estimate variance is computed along with the relative stability disaster will set in.

The variance of the periodogram is

$$\begin{aligned} \sigma^2 &= \langle |\hat{P}_s(\omega)|^2 \rangle - \langle \hat{P}_s(\omega) \rangle^2 \\ &= \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{j=0}^{N-1} \sum_{r=0}^{N-1} \sum_{s=0}^{N-1} \left[\langle x_k x_j x_r x_s \rangle - \langle x_k x_j \rangle \langle x_r x_s \rangle \right] \\ &\quad \times e^{-i\omega(k-j-r+s)T} \end{aligned} \quad (12.4-5)$$

Assuming, x_k also Gaussian then relation (12.1-43) may be used to express the bracketed terms in term of covariances, which can be expressed, in turn, in terms of the actual spectrum as performed frequently in the preceding pages:

$$\sigma^2 = \frac{1}{N^2 \omega_s^2} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega' \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} d\omega'' P_s(\omega') P_s(\omega'') \quad (12.4-6)$$

$$\times \left[\sum_{k=0}^{N-1} \sum_{j=0}^{N-1} \sum_{r=0}^{N-1} \sum_{s=0}^{N-1} \left[e^{i(\omega' - \omega)(k-r)T} e^{i(\omega' - \omega)(j-s)T} \right. \right. \\ \left. \left. + e^{i(\omega' - \omega)kT} e^{-i(\omega' + \omega)sT} e^{-i(\omega'' - \omega)jT} e^{i(\omega'' + \omega)jT} \right] \right]$$

The second term in the brackets approaches zero except near zero and half the sampling frequency where it will contribute as much as the first term. From (12.1-49) the first terms may be seen to be equal to

$$G_N(\omega, \omega') G_N(\omega, \omega'') \quad (12.4-7)$$

which for large N approaches

$$N^2 \omega_s^2 \delta(\omega - \omega') \delta(\omega - \omega'') \quad (12.4-8)$$

Then the variance becomes approximately

$$\sigma^2 = P_s^2(\omega) \quad (12.4-9)$$

and the relative variation is approximately

$$\eta = \frac{P_s(\omega)}{P_s(\omega)} \quad (12.4-10)$$

$$= 1$$

which may be seen to produce a totally inaccurate estimate. Periodograms as statistical spectrum estimators should be avoided!

12.6. INTERPRETATION OF THE RELATIVE STABILITY

The relative stability was defined as the ratio

$$\eta = \frac{\sigma}{\langle P_s(\omega) \rangle}$$

This ratio is an effective measure of the stability of an estimate; a η of .1 or .05 would seem to imply a highly peaked distribution of $P_s(\omega)$ (figure 12.3-1) with "most" values lying within ± 10 or ± 5 percent of the average value of $P_s(\omega)$. However, it is customary in statistical analyses to be a bit more precise, and usefully so.

An assumption (hypothesis) is made that an actual value of η is a parameter of the distribution of $P_s(\omega)$; then under this assumption the question is raised as to what will be the fractional deviation about the true mean for a given probability of finding the estimate between the limits of deviation.

The given probability is usually called the "confidence level"; the confidence that the experimenter has that the estimate "in hand" is within the prescribed limits. Judgement must still be exercised in the choice of "confidence level" that is acceptable, but this is another level above simple subjectivity and beyond mere acceptance of η , the relative stability, as the sole indicator of spectrum estimate reliability. Confidence levels of 60%, 70%, 80%, and 90% are not uncommon to use though the application usually determines the most desirable value.

For spectrum estimation a confidence level of 80% is not unreasonable as this would mean that approximately 4 out of 5 estimates at differing frequencies would, on the average, lie between prescribed limits. The approximation comes about because of the lack of complete statistical independence of adjacent estimates.

To make the previous ideas more mathematically precise, some remarks concerning the actual distribution functions of the lagged product and digital filter estimation methods should be made. As the data values are random variables the spectrum estimate will also be a random variable with a distribution function depending upon the data distribution. It would be a fortunate circumstance if the distribution of the spectrum estimate asymptotically approached some well known distribution for a large number of data points, regardless of the data distribution function or with mild restrictions on the data distribution. In mind, of course is the wish that the asymptotic distribution would be Gaussian so that the asymptotic mean and variance of the estimates would characterize the spectrum estimate distribution completely. Motivation for this desire comes from the "central limit" theorem for independent random variables. However, for stationary data that is not independently distributed (independent data samples imply a flat ("white") spectrum) and not Gaussian it is not true, in general, that the asymptotic spectrum estimate distribution is Gaussian. Grenander and Rosenblatt (Chapter 6) discuss the matter at some length and come to the conclusion that the assumption that the distribution of the spectrum estimate being Gaussian is not unwarranted. This assumption will be made subsequently.

Only a Gaussian distribution will be discussed as it leads to simple results and allows use of tables found in almost any handbook. *

The essence of the preceding paragraphs is contained with the following definition of fractional deviation. Assume that the individual estimate is desired to be bounded within the limits:

$$\langle \hat{P}_s(\omega) \rangle - \epsilon \leq \hat{P}_s(\omega) \leq \langle \hat{P}_s(\omega) \rangle + \epsilon \quad (12.5-1)$$

where ϵ is the absolute (symmetrical) deviation limit. Suppose now that the "fractional deviation", λ is defined by

$$\epsilon = \lambda \langle \hat{P}_s(\omega) \rangle \quad (12.5-2)$$

then (12.5-1) may be written

$$1 - \lambda \leq \frac{\hat{P}_s(\omega)}{\langle \hat{P}_s(\omega) \rangle} \leq 1 + \lambda \quad (12.5-3)$$

Obviously λ controls the desired limits of the ratio of an individual estimate to the mean. From (12.5-2) the fractional deviation may be written

$$\lambda = \frac{\epsilon}{\langle \hat{P}_s(\omega) \rangle} \quad (12.5-4)$$

and if ϵ is considered as some multiple, μ , of the r.m.s. spectrum error, σ , then

$$\begin{aligned} &= \frac{\mu \sigma}{\langle \hat{P}_s(\omega) \rangle} \\ &= \mu \eta \end{aligned} \quad (12.5-5)$$

from (12.3-16). μ is also the same multiple of the relative stability.

Now a uni-variate Gaussian distribution is defined by

* Handbook of Chemistry and Physics

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\xi} e^{-\frac{(x-m)^2}{2\xi^2}} \quad (12.5-6)$$

where

m = mean

ξ^2 = variance

With the inequality (12.5-3) written as

$$-\lambda \leq \frac{\hat{P}_s(\omega)}{\langle \hat{P}_s(\omega) \rangle} - 1 \leq \lambda \quad (12.5-7)$$

then it can easily be shown that

$$x = \frac{\hat{P}_s(\omega)}{\langle \hat{P}_s(\omega) \rangle} - 1 \quad (12.5-8)$$

is a Gaussianly distributed variable, if $P_s(\omega)$ is assumed Gaussianly distributed, that has a mean zero and variance η^2 . Hence

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\eta} e^{-\frac{x^2}{2\eta^2}} \quad (12.5-9)$$

The remaining problem is to find the values μ that produce various probabilities, the "confidence levels", of x lying within the limits $\pm \lambda$. Now this probability is given by

$$p = \frac{1}{\sqrt{2\pi}\eta} \int_{-\mu\eta}^{\mu\eta} e^{-\frac{x^2}{2\eta^2}} dx \quad (12.5-10)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\mu}^{\mu} e^{-\frac{x^2}{2}} dx.$$

Useful values of the probability (confidence level) p (in percentage) and the multiple at the relative stability are given in Table 12.5-1.

p (%)	μ
60	.84
65	.935
70	1.04
75	1.15
80	1.28
85	1.44
90	1.64
95	1.96
100	∞

Table 12.5-1 Confidence Levels versus Normalized Relative Stability

After the proliferation of formulae and monologue an example is eminently worthwhile.

Suppose an estimate of a spectrum value is desired within $\pm 10\%$ ($\lambda = .1$) at a confidence level of 90%. Then examination of Table 12.5-1 indicates that

$$\mu = 1.64$$

$$= \frac{.1}{\eta}$$

(12.5-11)

from (12.3.1-5) or

$$\eta = \frac{.1}{1.64}$$

(12.5-12)

$$.061$$

In turn this accuracy requirement places a constraint on the resolution and/or record length needed by the estimation methods.

When "hanning" is used with the lagged product method

$$\eta^2 = \frac{3}{4} \frac{M}{N} \quad (12.5-12)$$

Therefore for an accuracy of $\pm 10\%$ with a confidence level per estimate of 90% (9 out of 10 estimates will be within $\pm 10\%$ of the correct value)

$$\frac{M}{N} = 5.1 \times 10^{-3} \quad (12.5-13)$$

which means that $N = 1000$ data points will allow only $M + 1 = 6$ frequency estimates over the interval zero to half the sampling frequency.

If the confidence level is relaxed to 80% for the same accuracy ($\pm 10\%$) then

$$\begin{aligned} \eta &= \frac{.1}{1.28} \\ &= .078 \end{aligned} \quad (12.5-14)$$

and for "hanning"

$$\frac{M}{N} = 8.25 \times 10^{-3} \quad (12.5-15)$$

The spectrum can be resolved to approximately 9 frequencies; an improvement of about 50% over that for the 90% confidence level. A small change in confidence level produces a much larger number of frequencies ($M + 1$) into which the spectrum may be resolved.

13. LINEAR STATISTICAL FILTERS

Quite often the problem of estimating the value of a sampled function ("signal") that has been corrupted with noise or some function of it, presents itself. Examples are target tracking with a radar and reconstruction of telemetered data. Usually, the value of the unperturbed function is desired at some particular but fixed time with respect to the sampling times. This time may, or may not, be at one of the sampling times of the data and within, or without, the sampling span.

According to whether the time of estimation is within the sampling span, in the future, or the past; the estimation problem is called smoothing, prediction, or extrapolation into the past; respectively.

Work in this section will be devoted to the exposition of two methods:

1. Reconstruction or estimation of the uncorrupted sampled function, or some function of it that can be created by a linear operator, such as differentiation, assuming that the uncorrupted sampled function can be represented as a known linear combination of linearly independent functions of time.
2. Estimation of a function which is known only statistically through the sampled auto-correlation function and the sampled cross-correlation function of the function with the "noise".

Neither of these estimation methods require that the signal or noise be stationary (Section 11.1.4). Therefore, the linear filter weights may be time varying.

This linear estimation and reconstruction will be optimized (produce "best" estimates) in the sense that the mean-square error (the variance) of the estimate will be minimized. The minimization of the first method will be made subject to the condition that the estimate will be unbiased; that is, the mean value of the estimate is the value of the unperturbed function. This can be done with this method as the form of the "signal" is assumed known.

It should be pointed out that the fact that optimum linear estimation is being used does not mean that the variance error is a minimum with respect to all possible methods of estimation. Linear estimation is optimum with respect to minimum mean-square error for data samples whose noise error is Gaussianly distributed and a few other classes of noise distributions. Optimization is only for the class of all possible linear estimators.

The great utility of linear estimators lies in three main areas:

1. Simplicity of implementation
2. Ease of analysis
3. Gaussian processes are frequently found in nature, at least approximately.

A discussion of non-linear estimation with its complexities would convince the reader that optimum linear estimation is a valuable procedure.

Before embarking upon the mathematical details of developing these filters a few words about terminology is in order.

An "estimate" is exactly that. Noisy data is combined by means of a mathematical expression which, when evaluated, results in an approximation to a desired quantity. In the absence of "noise" it is not unusual to demand of this expression, which is called an "estimator," that it not be just an estimate, but exactly the value of the quantity desired. In statistical terminology an estimator of this kind is called an "unbiased" estimator. The first method meets this requirement and is discussed in Section 13.1. Section 13.1 discusses "Weiner" filters which do not have this constraint imposed and, as such, often perform better for small "signal-to-noise" ratios.

A "representation" of a function can either be an exact expression for, or an approximation to, the function. Depending on the physical situation mathematical complexity and the intended purpose of the estimation, different representations for the same function can be used. Hopefully these representations would be equal for all values of the independent variable, or, at least equivalent in, say, a least squares approximation, equality at sampling points, or some other useful manner.

An example of distinctions between "equivalent" representations might be the use of trigonometric functions versus polynomials to approximate a function over a range. The trigonometric functions would be periodic and repeat the function outside the range of approximation. A predictor using these functions for signal representation would most likely be useless unless the signal process is also periodic with the same period. On the other hand, polynomials tend to have values which continue the actual function outside of the range of approximation.

Filters discussed in this section will all be non-recursive; further z transform notation will not be used, and realizability conditions will not be imposed (though the filters are all realizable). Notation will be quite similar to previous sections, but putting the filters in a "standard" form tends to obscure the fact that they predict, interpolate, and extrapolate into the past. Also, the frequency response is not the point of interest but the operation in the minimum r. m. s. error sense is the criteria by which the filters are judged. Finally, if the noise (or signal) statistics are not stationary then a frequency response is difficult to define as the weights of the filter vary with time.

13.1 ESTIMATION OF A FUNCTION OF KNOWN FORM OR REPRESENTATION

Define the signal representation, in the absence of noise, to be:

$$s(t) = \sum_{r=0}^M \alpha_r \phi_r(t) \quad (13.1-1)$$

where the $\phi_r(t)$ are some set of linearly independent functions. If $s(t)$ were a polynomial, for example, then

$$\phi_r(t) = t^r, \quad r = 0, 1, \dots, M \quad (13.1-2)$$

and

$$\alpha_0 = s_0 = \text{initial position,}$$

$$\alpha_1 = v_0 = \text{initial velocity,}$$

$$\alpha_2 = a_0/2 = \text{initial acceleration,}$$

etc.

Estimation then involves determination of the parameters α_r which completely specify the function $s(t)$. Two methods are available.

1. Explicit estimation of $s(t)$ with implicit evaluation of the coefficients α_r .
2. Explicit estimation of the parameters α_r and then evaluating the function $s(t)$ at a particular time "t".

The first method involves direct minimization of the root-mean-square error at the time of evaluation (which is arbitrary).

Alternatively, the second method determines the coefficients α_r in a minimum r.m.s. error manner, but begs the question of the explicit optimization of the value of $s(t)$ at time "t". This method will be discussed in Section 13.1.2. It will be seen in subsequent work that the two methods are completely equivalent (Section 13.1.3).

While the derivations in this section optimize, in a statistical sense, the estimation of the function $s(t)$; it is also possible to arrive at an identical set of results using conventional least squares curve fitting for the particular case of uncorrelated sampling.

13.1.1 Estimation of the Function Directly

The procedure in the following endeavors will be definition of the estimator; equating of the ensemble averaged estimate to the value of the unperturbed function $s(t)$ which develops constraints on the estimator coefficients or "weights"; and utilization of these constraints to compute the optimum weights from the expression for the r.m.s. error in the function $s(t)$.

Define a linear estimator $\hat{s}(t)$ for the function $s(t)$ as:

$$\hat{s}(t) = \sum_{n=0}^{\bar{N}} w_{\bar{N}-n}(t) x_{n+k} = \sum_{n=0}^{\bar{N}} w_n(t) x_{\bar{N}-n+k} \quad (13.1.1-1)$$

where the $(\bar{N} + 1)$ $w_n(t)$ are constants (or weights) which must be determined to minimize the root-mean-square noise error at some evaluation time "t". The quantities x_k are assumed to be given by the $\bar{N} + 1$ sampled values of the uncorrupted signal $s(t)$ plus noise $n(t)$,

$$x_k = s(t_k) + n(t_k). \quad (13.1.1-2)$$

The sampling times t_k are arbitrary (obviously, t_k can not be equal to t_j unless $k = j$). The free index k in (13.1.1-1) serves to indicate that weights are moved along the sequence x_k as time progresses.

Taking the ensemble average (Section 7) of equation (13.1.1-1) for setting it equal to the actual function $s(t)$, for an "unbiased" estimate, gives

$$\langle \hat{s}(t) \rangle = \sum_{r=0}^M \alpha_r \phi_r(t) = \sum_{n=0}^N w_n(t) \langle x_{N-n+k} \rangle \quad (13.1.1-3)$$

using the uncorrupted signal representation 11.1.1.

For convenience take the initial sample at $t_0(k=0)$ and let $w_n = w_n(t)$, the dependence of the estimator on the time of estimation being understood.

Inserting (13.1.1-2) into (13.1.1-3) it is found that:

$$\begin{aligned} \sum_{r=0}^M \alpha_r \phi_r(t) &= \sum_{n=0}^N w_n \sum_{r=0}^M \alpha_r \phi_r(t_{N-n}) \\ &+ \sum_{n=0}^N w_n \langle n(t_{N-n}) \rangle \end{aligned} \quad (13.1.1-4)$$

Assuming, without loss of generality, that the noise has zero mean:

$$\langle n(t_{N-n}) \rangle = 0 \quad (13.1.1-5)$$

as these values if different from zero can be included in (13.1-1); then equation (13.1.1-4) implies, as it must hold for an arbitrary function (arbitrary α_r) within the representation 13.1-1, that

$$\phi_r(t) = \sum_{n=0}^N w_n \phi_r(t_{N-n}); \quad r = 0, 1, 2, \dots, M. \quad (13.1.1-6)$$

Mathematically, equation (13.1.1-6) produces $M + 1$ constraints on the $N + 1$ unknowns w_n .

Next, calculate the variance or root-mean-square error of the estimate $s(t)$. The square error is defined as the ensemble average (Section 11.1-3).

$$\sigma_s^2 = \langle (\hat{s}(t) - \langle \hat{s}(t) \rangle)^2 \rangle \quad (13.1.1-7)$$

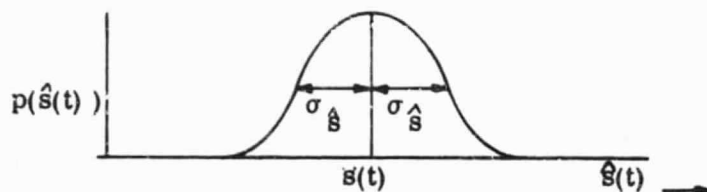


Figure 13.1. 1-1. Peaked Distribution

For probability distributions $p(\hat{s}(t))$ of $\hat{s}(t)$ which are peaked (Figure 13.1. 1-1) as is usually the case in practice, $\pm \sigma_s$ measures the "spread" or "average r. m. s. error" of the random variable $\hat{s}(t)$ about the mean which is $s(t)$ for mean zero noise. Non-zero mean noise can be accounted for by including this mean as part of the unperturbed function $s(t)$ as mentioned previously.

Upon insertion of equations (13.1. 1-2) and (13.1. 1-3) into the definition 13.1. 1-7 it is found that

$$\sigma_s^2 = \left\langle \sum_{n=0}^{\bar{N}} (w_n n(t_{\bar{N}-n}))^2 \right\rangle \quad (13.1. 1-8)$$

or

$$\sigma_s^2 = \sum_{n=0}^{\bar{N}} \sum_{j=0}^{\bar{N}} w_n w_j \langle n(t_{\bar{N}-n}) n(t_{\bar{N}-j}) \rangle$$

Defining the quantity

$$c_{nj} = c_{jn} = \langle n(t_{\bar{N}-n}) n(t_{\bar{N}-j}) \rangle \quad (13.1. 1-9)$$

as the auto-correlation function of the noise function $n(t)$ (Section 11)

$$\sigma_s^2 = \sum_{n=0}^{\bar{N}} \sum_{j=0}^{\bar{N}} w_n w_j c_{nj} \quad (13.1. 1-10a)$$

is obtained.

The problem of determining the optimum "filter" has been reduced to one of formal mathematics; namely, minimize the quadratic function of $N + 1$ variables w_n

$$\sigma_s^2(w_0, w_1, w_2, \dots, w_N) = \sum_{n=0}^{\bar{N}} \sum_{j=0}^{\bar{N}} w_n w_j c_{nj} \quad (13.1. 1-10b)$$

subject to the $M + 1$ constraints

$$\phi_r(t) = \sum_{n=0}^{\bar{N}} w_n \phi_r(t_{\bar{M}n}), \quad r = 0, 1, 2, \dots, M. \quad (13.1.1-6)$$

A well known mathematical technique, that of Lagrange multipliers, is available to assist in solving this problem. Form the function

$$\Theta = \sum_{n,j=0}^{\bar{N}} w_n w_j c_{nj} - \sum_{r=0}^M \lambda_r \left(\sum_{n=0}^{\bar{N}} w_n \phi_r(t_{\bar{N}-n}) - \phi_r(t) \right); \quad (13.1.1-11)$$

where the λ_r are constants to be determined. Finally, minimize the function Θ as if it were an unconstrained function of $N + 1$ independent variables w_s , by taking partial derivatives with respect to the weights, w_s :

$$\frac{\partial \Theta}{\partial w_s} = 0 = \sum_{n,j=0}^{\bar{N}} c_{nj} [w_n \delta_{js} + w_j \delta_{ns}] - \sum_{r=0}^M \lambda_r \sum_{n=0}^{\bar{N}} \phi_r(t_{\bar{N}-n}) \delta_{ns} \quad (13.1.1-12)$$

where

$$\frac{\partial w_n}{\partial w_s} = \delta_{ns} = \begin{cases} 1, & n=s \\ 0, & n \neq s. \end{cases}$$

$$\text{As } c_{ns} = c_{sn},$$

$$2 \sum_{n=0}^{\bar{N}} c_{sn} w_n = \sum_{r=0}^M \lambda_r \phi_r(t_{\bar{N}-s}); \quad s = 0, 1, \dots, \bar{N}. \quad (13.1.1-13)$$

Equations (13.1.1-6) and (13.1.1-13) express that fact that these are $M + 2 + \bar{N}$ linear equations in $M + 2 + \bar{N}$ unknowns w_n and λ_r . To simplify the solution of these equations write them in matrix form.

Replace equation (13.1.1-6) by the matrix equation

$$\bar{\Phi} W = \phi(t) \quad (13.1.1-14)$$

where

$$[\phi(t)]_r = \phi_r(t)$$

$$[\bar{\Phi}]_{rn} = \phi_r(t_{\bar{N}-n}),$$

and

$$[W]_n = w_n, \quad r = 0, 1, \dots, M \\ n = 0, 1, \dots, \bar{N}.$$

Similarly, equation (13.1.1-13) becomes

$$CW = \frac{1}{2} \tilde{\Phi} \Lambda; \quad (13.1.1-15)$$

where

$$[\Lambda]_r = \lambda_r, \quad r = 0, 1, \dots, M;$$

$$[C]_{ns} = c_{ns}, \quad n, s = 0, 1, \dots, N;$$

and $\tilde{\Phi}$ is the transpose of Φ :

$$[\tilde{\Phi}]_{pq} = \Phi_{qp}.$$

Therefore, the matrix equations to be solved are

$$\tilde{\Phi} W = \phi(t) \quad (13.1.1-14)$$

and

$$CW = \frac{1}{2} \tilde{\Phi} \Lambda \quad (13.1.1-15)$$

Eliminate Λ by premultiplying (13.1.1-15) by $\tilde{\Phi} C^{-1}$ and produce, using (13.1.1-14)

$$\begin{aligned} \tilde{\Phi} W &= \frac{1}{2} \tilde{\Phi} C^{-1} \tilde{\Phi} \Lambda \\ &= \phi(t) \end{aligned} \quad (13.1.1-16)$$

Using the fact that $\tilde{\Phi} C^{-1} \tilde{\Phi}$ is a square $M+1$ by $M+1$ matrix the Lagrange multipliers are:

$$\Lambda = 2 [\tilde{\Phi} C^{-1} \tilde{\Phi}]^{-1} \phi(t) \quad (13.1.1-17)$$

Combining equations (13.1.1-17) and (13.1.1-15) gives the result:

$$W = C^{-1} \tilde{\Phi} [\tilde{\Phi} C^{-1} \tilde{\Phi}]^{-1} \phi(t). \quad (13.1.1-18)$$

Another result of interest is the expression of the r. m. s. error (squared) in terms of W as given by (13.1.1-18). In matrix terms equation (13.1.1-10) for the r. m. s. error is

$$\sigma_s^2 = \tilde{W} C W, \quad (13.1.1-19)$$

and using the previous result (13.1.1-18) it is found

$$\sigma_s^2 = \tilde{\phi}(t) [\tilde{\Phi} C^{-1} \tilde{\Phi}]^{-1} \phi(t); \quad (13.1.1-20)$$

where the relations

$$\tilde{C}^{-1} = C^{-1} \quad \text{as } C \text{ is symmetric,}$$

and

$$\tilde{A}^{-1} = A^{-1}$$

for an arbitrary square matrix A that possesses an inverse, have been used.

Recapitulating, an explicit expression for the linear estimator which produces the minimum unbiased root-mean-square error at the point of estimation of the unknown function within the representation $s(t)$ has been found. An explicit expression for the r. m. s error has also been developed.

A question as to the existence of the inverse of the correlation matrix must be considered before calculations can be performed freely. For stationary processes the inverse does indeed exist. Consider the origin of the correlation matrix as zero for convenience; namely,

$$C_{rk} = \rho(r-k)T \quad ; r, k = 0, 1, \dots, \bar{N}; \quad (13.1.1-21)$$

where $\rho(\tau)$ is the sampled auto-correlation function normalized such that $\rho(0) = 1$

Write out the matrix C to inspect it:

$$C = \begin{pmatrix} \rho(0) & \rho(-1) & \rho(-2) & \dots & \rho(-\bar{N}) \\ \rho(1) & \rho(0) & & & \rho(-(\bar{N}-1)) \\ \rho(2) & \rho(1) & & & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \rho(\bar{N}) & \rho(\bar{N}-1) & & & \rho(0) \end{pmatrix} \quad (13.1.1-22)$$

For the inverse to exist, the determinant of C , $|C|$, cannot be zero. Alternatively this means that the rows (or columns) of C must be linearly independent. That is, a set of constants γ_r ; $r = 0, 1, \dots, \bar{N}$; cannot exist such that

$$\sum_{r=0}^{\bar{N}} \gamma_r C_{rk} = 0 \quad (13.1.1-23)$$

or

$$\sum_{r=0}^{\bar{N}} \gamma_r \rho((r-k)T) = 0$$

This is a well studied equation, a linear difference equation (See Section 3). Letting

$$\rho((r-k)T) = \beta^{r-k} \quad (13.1.1-24)$$

β must satisfy

$$\sum_{r=0}^{\bar{N}} \gamma_r \beta^r = 0; \quad (13.1.1-25)$$

a polynomial of the $\bar{N}$ th order which has $\bar{N}$ roots β_j . Therefore, for $\bar{N}$ distinct roots β_j , the solution to (13.1.1-25) has the form

$$\rho(kT) = \sum_{j=1}^{\bar{N}} A_j \beta_j^k ; \quad (13.1.1-26)$$

where the A_j are constants.

Now $\rho(\tau)$ has the properties (Section 11.1.6)

$$\begin{aligned} \rho(0) &= 1, \\ \rho(-\tau) &= \rho(\tau), \\ |\rho(\tau)| &\leq \rho(0). \end{aligned} \quad (13.1.1-27)$$

These properties cannot be satisfied by equation (13.1.1-26), except for rare periodic functions. Therefore, distinct roots will not do. Similarly, it can be found that multiple roots will not produce solutions with the necessary properties.

Therefore, any correlation function with the requisite properties cannot produce a non-zero determinant as the form of all functions that do produce zero determinants have been found and they do not have the properties (13.1.1-27).

In further calculations, therefore, the existence of inverses for stationary processes is of no concern.

13.1.2 Estimation of the Coefficients of the Function Representation

In a manner identical to the previous pages a linear estimator for the coefficients of $s(t)$ (13.1-1) can be defined:

$$\hat{\alpha}_r = \sum_{n=0}^{\bar{N}} w_{rn} x_{N-n} \quad r = 0, 1, \dots, M; \quad (13.1.2-1)$$

where the weights w_{rn} are determined to produce an unbiased minimum variance estimate of the actual α_r 's. The front subscript on the w_{rn} refers to the particular coefficient being estimated.

As the calculation is identical in almost every detail to that for $\hat{s}(t)$ only the results are given:

$$W_r = C^{-1} \tilde{\Phi} [\tilde{\Phi} C^{-1} \tilde{\Phi}]^{-1} I_r \quad (13.1.2-2)$$

and

$$\sigma_{\hat{\alpha}_r}^2 = \tilde{I}_r [\tilde{\Phi} C^{-1} \tilde{\Phi}]^{-1} I_r \quad (13.1.2-3)$$

where both W_r and I_r are column matrices:

$$\begin{aligned} [W_r]_n &= w_{rn}; \\ [I_r]_n &= \delta_{rn} = \begin{cases} 1, & r = n, \\ 0, & \text{otherwise,} \end{cases} \end{aligned}$$

and

$$\sigma_{\hat{\alpha}_r} = \text{r. m. s. error in estimate of } \alpha_r.$$

13.1.3 Equivalence of Function and Coefficient Estimation

To demonstrate the equivalence of function estimation at a point and coefficient estimation, write the estimator for $s(t)$ in matrix form also

$$\hat{s}(t) = \tilde{X}W; [X]_n = x_{\bar{N}-n}, \quad (13.1.3-1)$$

or

$$\hat{s}(t) = \tilde{X} C^{-1} \tilde{\Phi} [\tilde{\Phi} C^{-1} \tilde{\Phi}]^{-1} \phi(t)$$

Using the estimates for the coefficients in matrix form

$$\hat{\alpha}_r = \tilde{X}W_r = \tilde{X}C^{-1} \tilde{\Phi} [\tilde{\Phi} C^{-1} \tilde{\Phi}]^{-1} I_r, \quad (13.1.3-2)$$

and

$$\hat{s}(t) = \sum_{r=0}^M \hat{\alpha}_r \phi_r(t), \quad (13.1.3-3)$$

Where the prime on $\hat{s}'(t)$ indicates function estimation using estimated coefficients

or

$$\hat{s}'(t) = \sum_{r=0}^M \tilde{X} C^{-1} \tilde{\Phi} [\tilde{\Phi} C^{-1} \tilde{\Phi}]^{-1} I_r \phi_r(t) \quad (13.1.3-4)$$

Using the associativity of matrix multiplication

$$\hat{s}'(t) = \tilde{X} C^{-1} \tilde{\Phi} [\tilde{\Phi} C^{-1} \tilde{\Phi}]^{-1} \sum_{r=0}^M I_r \phi_r(t) \quad (13.1.3-5)$$

which becomes upon comparing (13.1.3-1) and (13.1.3-5)

$$\hat{s}(t) = \hat{s}'(t) = \tilde{X} C^{-1} \tilde{\Phi} [\tilde{\Phi} C^{-1} \tilde{\Phi}]^{-1} \phi(t) \quad (13.1.3-6)$$

and it is seen that the two methods are equivalent.

13.1.4 Generalization to Estimators of the Effect of an Arbitrary Linear Operator

Often interest centers upon performing some linear operation, L , on the data such as differentiation. Assuming noisy data samples, as before, the problem is to find a linear estimator of this operator (filter) so that its output has the least r. m. s. noise error and is exact in the absence of noise.

Derivation of this filter is identical to the previous sections. The operator is defined by its effect upon the signal representation. By the postulated linearity of the operator L , the effect upon the unperturbed function $s(t)$ is given by

$$v(t) = L s(t) = \sum_{r=0}^M \alpha_r L \phi_r(t) \quad (13.1.4-1)$$

If for example L is the differentiation operator and the $\phi_r(t)$ are trigonometric exponentials $e^{i\omega_r t}$, then

$$L s(t) = \frac{d}{dt} s(t) = \sum_{r=0}^{M=2P} \alpha_r i\omega_{P-r} e^{i\omega_{P-r} t} \quad (13.1.4-2)$$

where $L \phi_r(t) = i\omega_r e^{i\omega_r t}$

Again, the derivation in this section parallels that for the estimator of the function $s(t)$ ($L = 1$), Section 13.1.1. This time the constraint becomes, mathematically,

$$\begin{aligned} \langle \hat{v}(t) \rangle &= \left\langle \sum_{k=0}^{\bar{N}} w_k^L x_{\bar{N}-k} \right\rangle = \sum_{k=0}^{\bar{N}} w_k^L s(t_{\bar{N}-k}) \quad (13.1.4-3) \\ &= \sum_{k=0}^{\bar{N}} \sum_{r=0}^M w_k^L \alpha_r \phi_r(t_{\bar{N}-k}) \end{aligned}$$

and as the relation must hold for arbitrary α_r , it is found using the equation for an unbiased estimator:

$$v(t) = \langle \hat{v}(t) \rangle \quad (13.1.4-4)$$

that

$$\sum_{k=0}^{\bar{N}} \sum_{r=0}^M w_k^L \alpha_r \phi_r(t_{N-k}) = \sum_{r=0}^M \alpha_r L \phi_r(t) \quad (13.1.4-5)$$

or

$$\sum_{k=0}^N w_k^L \phi_r(t_{N-k}) = L \phi_r(t) \quad (13.1.4-6)$$

Using these constraints along with Lagrange multipliers in a manner identical to that for the $s(t)$ estimator (Section 13.1.1) it is found that the weights are given by

$$W^L = C^{-1} \tilde{\Phi} [\tilde{\Phi} C^{-1} \tilde{\Phi}]^{-1} \phi^L(t) \quad (13.1.4-7)$$

where C is the covariance matrix (equation 13.1.1-9), $\tilde{\Phi}$ the sampling matrix (equation 13.1.1-14) and

$$W^L = \begin{bmatrix} w_0^L \\ w_1^L \\ \vdots \\ w_{\bar{N}}^L \end{bmatrix} \quad (13.1.4-8)$$

and

$$\phi^L(t) = \begin{bmatrix} L \phi_0(t) \\ L \phi_1(t) \\ \vdots \\ L \phi_{\bar{N}}(t) \end{bmatrix}$$

The variance of the estimate is given as

$$\begin{aligned} \sigma_{\hat{V}}^2 &= W^L C \tilde{W}^L \\ &= \tilde{\phi}^L(t) [\tilde{\Phi} C^{-1} \tilde{\Phi}] \phi^L(t) \end{aligned} \quad (13.1.4-9)$$

13.1.5 Derivative Estimators

Results of the preceding section allow the expressions for derivative estimators to be written. They are

$$W^{(q)} = C^{-1} \tilde{\Phi} \{ \tilde{\Phi} C^{-1} \tilde{\Phi} \}^{-1} \phi^{(q)}(t), \quad (13.1.5-1)$$

where $W^{(q)}$ is the column matrix of the estimator weights for the q th derivative of $s(t)$ at time " t " and

$$[\phi^{(q)}(t)]_r = \frac{d^q}{dt^q} (\phi_r(t)) \quad (13.1.5-2)$$

The variance of the estimate is given as

$$\sigma_{s_q}^2 = \tilde{\Phi}^q(t) \{ \tilde{\Phi} C^{-1} \tilde{\Phi} \}^{-1} \phi^q(t). \quad (13.1.5-3)$$

13.1.6 Explicit Expression of Certain Estimators and Their Square Errors

As an aid to computation of the estimator weights and the performance measures (the root-mean-square errors) the matrix equations are presented in indicial form. (A graphical presentation of the performance of various estimators using these equations is given in Section 13.1.11).

The weights for the estimator (including the derivatives of $s(t)$)

$$\hat{s}^{(q)}(t) = \sum_{n=0}^{\bar{N}} w_n^{(q)} x_{N-n} \quad (13.1.6-1)$$

are, from matrix equation (13.1.5-1)

$$w_r^{(q)} = \sum_{n=0}^{\bar{N}} \sum_{m=0}^M \sum_{s=0}^M C_{rn}^{-1} \tilde{\Phi}_{mn} u_{ms} \phi_s^{(q)}(t); r=0, 1, \dots, \bar{N}. \quad (13.1.6-2)$$

where the definition

$$u_{kj} = u_{jk} = \{ \tilde{\Phi} C^{-1} \tilde{\Phi} \}^{-1}_{jk} \quad (13.1.6-3)$$

$$= \left[\sum_{n=0}^{\bar{N}} \sum_{m=0}^M \tilde{\Phi}_{jn} C_{nm}^{-1} \tilde{\Phi}_{km} \right]^{-1}_{jk}; j, k = 0, 1, \dots, M.$$

is made,

Similarly, the square error from equation (13.1.5-8) can be expressed as

$$\sigma_{\hat{\alpha}_q}^2 = \sum_{j=0}^M \sum_{r=0}^M \phi_r(t) \phi_j(t) u_{rj} \quad (13.1.6-4)$$

The coefficients α_n have estimator coefficients:

$$w_{nr} = \sum_{t=0}^N \sum_{m=0}^M \sum_{s=0}^M C_{rt}^{-1} \Phi_{mt} u_{mn}; \quad \begin{matrix} n = 0, 1, \dots, \tilde{N} \\ r = 0, 1, \dots, M. \end{matrix} \quad (13.1.6-5)$$

and mean square errors:

$$\sigma_{\hat{\alpha}_n}^2 = u_{nn}$$

The matrix u_{jk} can be interpreted (and is) as the cross-correlation matrix of the estimated coefficients $\hat{\alpha}_j$.

13.1.7 Specialization of the Estimators to an Important Case

An important special case met often in practice, at least approximately, is that of measurements of the corrupted function $s(t)$ which are statistically independent, mean zero, and of equal square error. That is, the correlation between successive measurements is zero ($\langle n(t_k) n(t_j) \rangle = 0$, j not equal to k) and the noise square error of each measurement has a common value $\langle n^2(t_k) \rangle = \sigma^2$ (the noise process is stationary).

Under these assumptions the matrix C is given by

$$C = \sigma^2 I; [I]_{ij} = \delta_{ij}, \text{ the identity matrix.} \quad (13.1.7-1)$$

The estimator weight and error formulas simplify to

$$W^{(q)} = \tilde{\Phi} [\tilde{\Phi} \tilde{\Phi}]^{-1} \phi^{(q)}(t) \quad (13.1.7-2)$$

or

$$w_r^{(q)} = \sum_{m=0}^M \sum_{s=0}^M \tilde{\Phi}_{mr} u'_{ms} \phi^{(q)}(t), \quad r = 0, 1, \dots, \bar{N};$$

and

$$\sigma_{\hat{s}_q}^2 = \tilde{W}^{(q)} W^{(q)} \sigma^2 \quad (13.1.7-3)$$

or

$$\frac{\sigma_{\hat{s}_q}^2}{\sigma^2} = \sum_{j=0}^M \sum_{r=0}^M \phi_j^{(q)}(t) \phi_r^{(q)}(t) u'_{jr},$$

$$u'_{jk} = \left[\sum_{n=0}^{\bar{N}} \tilde{\Phi}_{jn} \tilde{\Phi}_{kn} \right]_{jk}^{-1}; \quad j, k = 0, 1, \dots, M.$$

Similarly, for the coefficients of the representation, α_n ;

$$w_{nr} = \sum_{m=0}^M \tilde{\Phi}_{mr} u'_{mn}, \quad \begin{matrix} n = 0, 1, \dots, M \\ r = 0, 1, \dots, \bar{N} \end{matrix} \quad (13.1.7-4)$$

and

$$\frac{\sigma_{\hat{\alpha}_n}^2}{\sigma^2} = u'_{nn}. \quad (13.1.7-5)$$

13.1.8 Estimator Examples

Illustrative useful examples that can be carried out relatively easily with pencil and paper, rather than computer computation, are the polynomial representations with equally spaced samples. These calculations are carried out here for $M = 0$ and 1; which are constant and linear estimation, respectively. Also assume the special noise case of independent equal variance samples.

a. $M = 0$: constant estimation

The matrix $\tilde{\Phi}$ becomes

$$\tilde{\Phi} = \frac{[1, 1, \dots, 1]}{N+1} \quad \text{as } \phi_0(t) = 1 \text{ for all } t. \quad (13.1.8-1)$$

Therefore:

$$\tilde{\Phi} \tilde{\Phi}^T = \bar{N} + 1, \text{ (a } 1 \times 1 \text{ matrix, a scalar)} \quad (13.1.8-2)$$

and

$$W = \tilde{\Phi} [\tilde{\Phi} \tilde{\Phi}^T]^{-1} \phi(t) = \frac{1}{\bar{N} + 1} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}; \quad (13.1.8-3)$$

or

$$w_k = \frac{1}{\bar{N} + 1}$$

The square error is given by

$$\frac{\sigma_{\hat{s}}^2}{\sigma^2} = \frac{1}{\bar{N} + 1} \quad (13.1.8-4)$$

As expected, the best linear estimator of a constant using $\bar{N} + 1$ samples is:

$$\hat{s} = \sum_{m=0}^{\bar{N}} w_m x_{\bar{N}-m}$$

or

$$s = \left[\sum_{m=0}^{\bar{N}} x_m \right] / (\bar{N} + 1) \quad (13.1.8-5)$$

the well known "average" that is met first by most people.

b. $M = 1$: linear estimation (a more complicated procedure)

$$\phi_0(t) = 1, \phi_1(t) = t; \quad (13.1.8-6)$$

and the matrix Φ becomes

$$\Phi = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ \bar{N}T & \dots & 2T & T & 0 \end{bmatrix} \quad (13.1.8-7)$$

where T is the time between samples.

Also obtained is

$$\Phi \tilde{\Phi} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ \bar{N}T & \dots & 2T & T & 0 \end{bmatrix} \quad \begin{matrix} \leftarrow \bar{N}+1 \rightarrow \\ \text{Columns} \end{matrix} \quad \begin{bmatrix} 1 & \bar{N}T \\ 1 & \cdot \\ \vdots & \cdot \\ 1 & 2T \\ 1 & T \\ 1 & 0 \end{bmatrix} \quad \begin{matrix} \uparrow \\ \bar{N}+1 \\ \text{rows} \\ \downarrow \end{matrix} \quad (13.1.8-8)$$

or

$$\Phi \tilde{\Phi} = \begin{bmatrix} \bar{N}+1 & T \sum_{r=1}^{\bar{N}} r \\ T \sum_{r=1}^{\bar{N}} r & T^2 \sum_{r=1}^{\bar{N}} r^2 \end{bmatrix}; \quad (13.1.8-9)$$

a 2×2 matrix

Defining:

$$A = \bar{N} + 1$$

$$B = T \sum_{r=1}^{\bar{N}} r = \frac{\bar{N}(\bar{N}+1)}{2} T \quad (13.1.8-10)$$

$$C = T^2 \sum_{r=1}^{\bar{N}} r^2 = \frac{\bar{N}(\bar{N}+1)(2\bar{N}+1)}{6} T^2$$

(13.1.8-9) can be written

$$\Phi \tilde{\Phi} = \begin{bmatrix} A & B \\ B & C \end{bmatrix} \quad (13.1.8-11)$$

Its inverse is

$$[\Phi \tilde{\Phi}]^{-1} = \frac{1}{AC-B^2} \begin{bmatrix} C & -B \\ -B & A \end{bmatrix} \quad (13.1.8-12)$$

The estimator of the function $s(t) = b_0 + b_1 t$ is given by using the column vector

$$\phi(t) = \begin{bmatrix} 1 \\ t \end{bmatrix} \quad (13.1.8-13)$$

and therefore

$$V = \begin{bmatrix} \cdot & \bar{N}T \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & 2T \\ 1 & T \\ 1 & 0 \end{bmatrix} \quad \frac{1}{AC-B^2} \begin{bmatrix} C & -B \\ -B & A \end{bmatrix} \begin{bmatrix} 1 \\ t \end{bmatrix} \quad (13.1.8-14)$$

or

$$W = \frac{1}{AC-B^2} \begin{bmatrix} C - Bt + \bar{N} T & (-B + At) \\ C - Bt + (\bar{N} - 1) T & (-B + At) \\ C - Bt + T & (-B + At) \\ C - Bt & \end{bmatrix} \quad (13.1.8-15)$$

In terms of individual members of W:

$$w_k = \frac{1}{AC-B^2} (C - Bt + (\bar{N} - k) (-B + At)); k = 0, 1, \dots, \bar{N}. \quad (13.1.8-16)$$

For the case of only two data points ($\bar{N} = 1$) and prediction one sample period into the future ($t = 2T$) it is found that

$$\begin{aligned} w_0 &= 2 \\ w_1 &= -1 \end{aligned} \quad (13.1.8-17)$$

and the estimator is

$$\hat{s}(2T) = 2x_1 - x_0 \quad (13.1.8-18)$$

To demonstrate its efficacy, use it on samples of the function with

$$\begin{aligned} s(t) &= b_0 + b_1 t, \\ x_0 &= b_0 + b_1 t, \\ x_1 &= b_0 + b_1(t + T), \end{aligned} \quad (13.1.8-19)$$

and expect

$$\hat{s} = b_0 + b_1(t + 2T) \quad (13.1.8-20)$$

Inserting in the estimator expression with the weights of (13.1.1-17)

$$\hat{s} = 2b_0 + 2b_1(t + T) - b_0 - b_1 t$$

or

$$\hat{s} = b_0 + b_1(t + 2T); \quad (13.1.8-21)$$

which checks, as it must.

Computed next is the square error for the linear predictors

$$\sigma_{\hat{s}}^2 = \frac{(L, t)}{AC - B^2} \begin{bmatrix} C & -B \\ -B & A \end{bmatrix} \begin{bmatrix} 1 \\ t \end{bmatrix} \sigma^2 \quad (13.1.8-22)$$

or

$$\frac{\sigma_{\hat{s}}^2}{\sigma^2} = \frac{12\left(\frac{t}{T}\right)^2 - 12\bar{N}\left(\frac{t}{T}\right) + 2\bar{N}(2\bar{N}+1)}{\bar{N}(\bar{N}+1)(\bar{N}+2)}$$

A point of interest is the functional dependence of $\sigma_{\hat{s}}$ on the evaluation (prediction) time t and the number of data samples $\bar{N} + 1$. It is seen that $\sigma_{\hat{s}}$ varies linearly with time t and inversely as the $\sqrt{\bar{N} + 1}$. The limiting forms become

$$\frac{\sigma_{\hat{s}}}{\sigma} \approx \frac{2\sqrt{3}}{\bar{N}(\bar{N}+1)(\bar{N}+2)} \frac{t}{T} \quad \text{for } \left|\frac{t}{T}\right| \gg 1; \quad (13.1.8-23)$$

and

$$\frac{\sigma_{\hat{s}}}{\sigma} \approx \frac{2}{\sqrt{\bar{N}+1}}, \quad \text{for } \bar{N} + 1 \text{ large.} \quad (13.1.8-24)$$

For the case of unequal sampling times and estimation of a linear function, it is interesting to compute the normalized noise error. Computation of the weighting coefficients, while certainly useful when smoothing, is of no particular theoretical interest.

As in the equal sample interval case the matrix

$\Phi \tilde{\Phi}$ is given by

$$\Phi \tilde{\Phi} = \begin{bmatrix} A' & B' \\ B' & C' \end{bmatrix} \quad (13.1.8-25)$$

where in this case

$$\begin{aligned} A' &= \sum_{j=0}^{\bar{N}} 1 = \bar{N} + 1, \\ B' &= \sum_{j=0}^{\bar{N}} t_j, \\ C' &= \sum_{j=0}^{\bar{N}} t_j^2. \end{aligned} \quad (13.1.8-26)$$

Similarly the inverse is

$$[\Phi \tilde{\Phi}]^{-1} = \frac{1}{A'C' - B'^2} \begin{bmatrix} C' & -B' \\ -B' & A' \end{bmatrix} \quad (13.1.8-27)$$

and

$$\begin{aligned} \frac{\sigma_s^2}{\sigma^2} &= \frac{(1, t)}{A'C' - B'^2} \begin{bmatrix} C' & -B' \\ -B' & A' \end{bmatrix} \begin{bmatrix} 1 \\ t \end{bmatrix} \\ &= \frac{(1, t)}{A'C' - B'^2} \begin{bmatrix} C' - B't \\ -B' + A't \end{bmatrix} \end{aligned} \quad (13.1.8-28)$$

or

$$\frac{\sigma_s^2}{\sigma^2} = \frac{C' - 2B't + A't^2}{A'C' - B'^2} \quad (13.1.8-29)$$

To put this result in a more interesting form define the mean of the measurement times

$$m_t = \frac{\sum_{j=0}^{\bar{N}} t_j}{\bar{N} + 1} = \frac{B'}{A'} \quad (13.1.8-30)$$

and the spread (squared) of the measurement times

$$\sigma_t^2 = \frac{\sum_{j=0}^{\bar{N}} (t_j - m_t)^2}{\bar{N} + 1} = \frac{C'}{A'} - \left(\frac{B'}{A'} \right)^2 \quad (13.1.8-31)$$

and find that

$$\frac{\sigma_s^2}{\sigma^2} = \frac{t^2 - 2 \frac{B'}{A'} t + \frac{C'}{A'}}{A' \left(\frac{C'}{A'} - \left(\frac{B'}{A'} \right)^2 \right)} \quad (13.1.8-32)$$

Completing the square in the numerator

$$\frac{\sigma_s^2}{\sigma^2} = \frac{t^2 - \frac{2B'}{A'} t + \frac{C'}{A'} + \left(\frac{B'}{A'} \right)^2 - \left(\frac{B'}{A'} \right)^2}{a' \sigma_t^2} \quad (13.1.8-33)$$

or finally

$$\frac{\sigma_s^2}{\sigma^2} = \frac{1}{\bar{N} + 1} \left(1 + \frac{(t - m_t)^2}{\sigma_t^2} \right) \quad (13.1.8-34)$$

Obviously, σ_s is a minimum when the prediction time is equal to the mean of the measurement times. For equal sampling times this means that the smoothing is best ("optimum") when the prediction, or smoothing times, is used in the middle of the data span $t = \frac{NT}{2}$ (also a "delay" of $\frac{NT}{2}$ in time). This result does not hold for an arbitrary order predictor in general.

13.1.9 Representation Error

Another facet of these estimators is the error made when using an inadequate representation in their design. That is, M is chosen, when designing the estimator, smaller than the actual number of functions $\phi_j(t)$ that are present, say, M' in the signal representation. This will cause a "representation error" even without noise.

An example would be the use of an estimator, designed for errorless prediction of a straight line, upon a quadratic input. After development of error formulas for the general function representation $s(t)$ in equation (13.1-1), the error will be evaluated in this special case.

Define the representation error as the difference between the noiseless function $s(t)$ and the estimated output $\hat{s}(t)$

$$\epsilon_M^{M'} = \hat{s}(t) - s(t) \quad (13.1.9-1)$$

where the subscript M and superscript M' indicate the explicit dependence of the error on the predictor design, M and the signal M'. Obviously, if M = M' no error will be present (other than noise).

$$\epsilon_M^{M'} = \sum_{r=0}^{M'} \alpha_r \phi_r(t) - \sum_{m=0}^{\bar{N}} w_m \sum_{r=0}^{\bar{N}} \alpha_r \phi_r(t_{\bar{N}-m}). \quad (13.1.9-2)$$

Use the equations of constraint (13.1.1-6) for the unbiased estimation of the signal when M of the functions $\phi_r(t)$ are present. These constraints were used in calculating the w_m 's. They are

$$\phi_r(t) - \sum_{m=0}^{\bar{N}} w_m \phi_r(t_{\bar{N}-m}) = 0 \quad (13.1.1-6)$$

Upon splitting the second sum in equation 13.1.9-2 into two parts over the index r; namely, from $0 \leq r \leq M$ and $M+1 \leq r \leq M'$ it is found, upon using the constraints, that

$$\epsilon_M^{M'} = \sum_{r=M+1}^{M'} \alpha_r [\phi_r(t) - \sum_{m=0}^{\bar{N}} w_m \phi_r(t_{\bar{N}-m})]. \quad (13.1.9-3)$$

Obviously, if the predictor were designed for a representation of M' order, the term in brackets would be zero, as it is the equation of constraint for that representation. The error would then be zero as it should be.

The special cases of M = 1, M' = 2 are evaluated for the polynomial signal representation. Phrased another way, find the quadratic error for a linear predictor.

The weights were given, for the linear input sampled at equally spaced times, in equation 13.1.8-16 as:

$$w_m = \frac{1}{AC - B^2} (C - Bt + (\bar{N}-m)(-B + At)) \quad (13.1.8-16)$$

Assuming the quadratic input

$$s(t) = \sum_{r=0}^2 \alpha_r \phi_r(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$$

the representation errors are found

$$\epsilon_{(1)}^{(2)} = \alpha_2 \left[t^2 - \sum_{m=0}^{\bar{N}} w_m (\bar{N} - m)^2 T_s^2 \right]$$

or

(13.1.8-17)

$$\epsilon_{(1)}^{(2)} = \alpha_2 \left[t^2 - \frac{C^2 - B^3 + (A B^2 - CB)t}{AC - B^2} \right]$$

13.1.10 Orthogonal Representations

If the function representation is assumed to be orthogonal over the sampling span, then certain simplifications result. By "orthogonal over the sampling span" the following is meant.

Let, as before,

$$s(t) = \sum_{r=0}^M \alpha_r \phi_r(t) \quad (13.1-1)$$

then the functions $\phi_r(t)$ possess the property

$$\sum_{k=0}^{\bar{N}} \phi_r(t_k) \phi_j^*(t_k) = \delta_{rj} = \begin{cases} 1, & r = j \\ 0, & r \neq j \end{cases} \quad (13.1.10-1)$$

(where the asterisk indicates complex conjugation) for the set of points t_k .

For the special case discussed previously with $C = \sigma^2 I$, the matrix can be evaluated (the previous derivation modified slightly to allow complex functions $\phi_r(t)$)

$$[\tilde{\Phi} \tilde{\Phi}^*]_{jn} = \sum_{k=0}^{\bar{N}} \tilde{\Phi}_{jk} \tilde{\Phi}_{nk}^* = \sum_{n=0}^{\bar{N}} \phi_j(t_{\bar{N}-k}) \phi_n^*(t_{\bar{N}-k}) \quad (13.1.10-2)$$

and immediately

$$[\tilde{\Phi} \tilde{\Phi}^*] = [\tilde{\Phi} \tilde{\Phi}^*]^{-1} = I \quad (13.1.10-3)$$

Namely, $[\tilde{\Phi} \tilde{\Phi}^*]$ is the identity matrix, which is its own inverse. The corresponding weight vector for the estimator is given as

$$w(q) = \tilde{\Phi}^* \phi^{(q)}(t) \quad (13.1.10-4)$$

and the square error is

$$\sigma_{sq}^2 = \hat{\phi}^{(q)}(t) \phi^{(q)}(t) \sigma^2 \quad (12.1.10-5)$$

13.1.10 Graphical Presentation of Polynomial Representation Estimator Properties With Stationary, Uncorrelated, Noise

As an indication of the efficiency of estimators of functions of a known linear form, various graphs are presented demonstrating performance of polynomial predictors using equi-spaced data. These predictors were discussed in Section 13.1.8.

The polynomial predictors here are specialized to stationary noise that is uncorrelated between samples, this is not atypical of the aircraft tracking problem using radar. Pulses are usually spaced many times the correlation time of the receiver (the reciprocal of the bandwidth).

Examination of the estimator weights W , for the polynomial predictor of a given order in terms of the matrices Φ and ϕ , shows that they are functions of the ratio

$$\alpha = \frac{t_p}{NT} \quad \text{the normalized prediction time} \quad (13.1.11-1)$$

and of $\bar{N}$,

where

- t_p = prediction time measured from the beginning of the data span in seconds
- $N+1$ = the number of data points used in the predictor
- T = the time between equi-spaced data samples
- NT = the total span of time covered by the data in seconds.

For various ranges of α three classes of predictors may be defined:

1. $\alpha < 0$, extrapolation into the past;
2. $0 \leq \alpha \leq 1$, smoothing;
3. $\alpha > 1$, prediction into the future.

Similarly, as the noise has been chosen to be stationary and uncorrelated, the correlation matrix is given by

$$C = \sigma I; I = \text{identity matrix}$$

with σ equal to the r.m.s error of an individual data sample.

These facts allow plotting of the "normalized rms error"

$$\frac{\sigma_s}{\sigma} \quad (13.1.11-2)$$

as a function of $\alpha = t_p / NT$.

Of course, graphs for fixed α , the normalized prediction time with the number of data points $\bar{N}+1$ varied are also of interest for various predictor orders and are included.

Graphs follow with interpretations where needed.

Figure 13.1.11-1 is a plot of the normalized r. m. s estimation error versus the normalized estimation time. It can be seen from Figure 13.1.11-1, and proven mathematically, that the predictor errors are symmetrical about $\alpha = .5$, the center of the data span. Actually, only errors for $\alpha \geq .5$ need be presented, but values for $\alpha \geq 0$ are included for clarity.

The errors are obviously least for smoothing, $0 \leq \alpha \leq 1$, and increase rapidly for extrapolation and prediction times outside the data span, $\alpha < 0$, $\alpha > 1$; further prediction, $\alpha > 1$, is best for lower order predictors.

Finally, error minimums depend on the order of the predictor and are found at prediction times within the data span. For odd order (M) predictors the error minimums occur in the center of the data span. Even order predictors have minimum errors at points other than the data span center, except for $M = 0$ which is a constant estimator and has a constant error regardless of the prediction time (See Section 13.1.8).

Figure 13.1.11-2 depicts the error (normalized) for fixed α 's with the number of data points varying. Figures 13.1.11-2a, b, and c are for predictor orders $M = 1, 2$, and 3, respectively. It can be seen that increasing the number of data samples has the desired effect of reducing the estimation error.

In Figure 13.1.11-3 the effect of increasing the predictor order can be observed. Two particular normalized prediction times have been chosen; smoothing in the middle of the interval, $\alpha = .5$, and prediction one sampling interval ahead of the data span (radar tracking), $t_p = (\bar{N} + 1) T$ or $\alpha = (\bar{N} + 1)/\bar{N}$. In Figure 13.1.11-3a, midpoint smoothing the errors are seen to increase with order (M) though successive even-odd order pairs have the same value; hence the rough curves, which are obviously not valid at points of the abscissa other than the integers. Figure 13.1.11-3b illustrates the same phenomena of increasing error with order. Obviously, the curve being estimated should be fitted with the least possible order. Statistical accuracy is lost as the order of fitting is increased.

Frequency response magnitudes, for the two cases $\alpha = .5$ (midpoint smoothing) and, $\alpha = (\bar{N} + 1)/\bar{N}$ (prediction one interval in the future) are given in Figure 13.1.11-4. The sampling frequency has been normalized to 1 sample per second. Figure 13.1.11-4a ($M=1$) and b ($M=2$) for midpoint smoothing behave somewhat like low pass filters, but it can be seen from Figures 13.1.11-4c ($M=1$) and 13.1.11-4b ($M=2$) that predictor behavior changes violently when α increases to 1 unit of prediction, $\alpha = (\bar{N}+1)/\bar{N}$. Frequency response is a tenuous idea at best when describing "optimum" estimators.

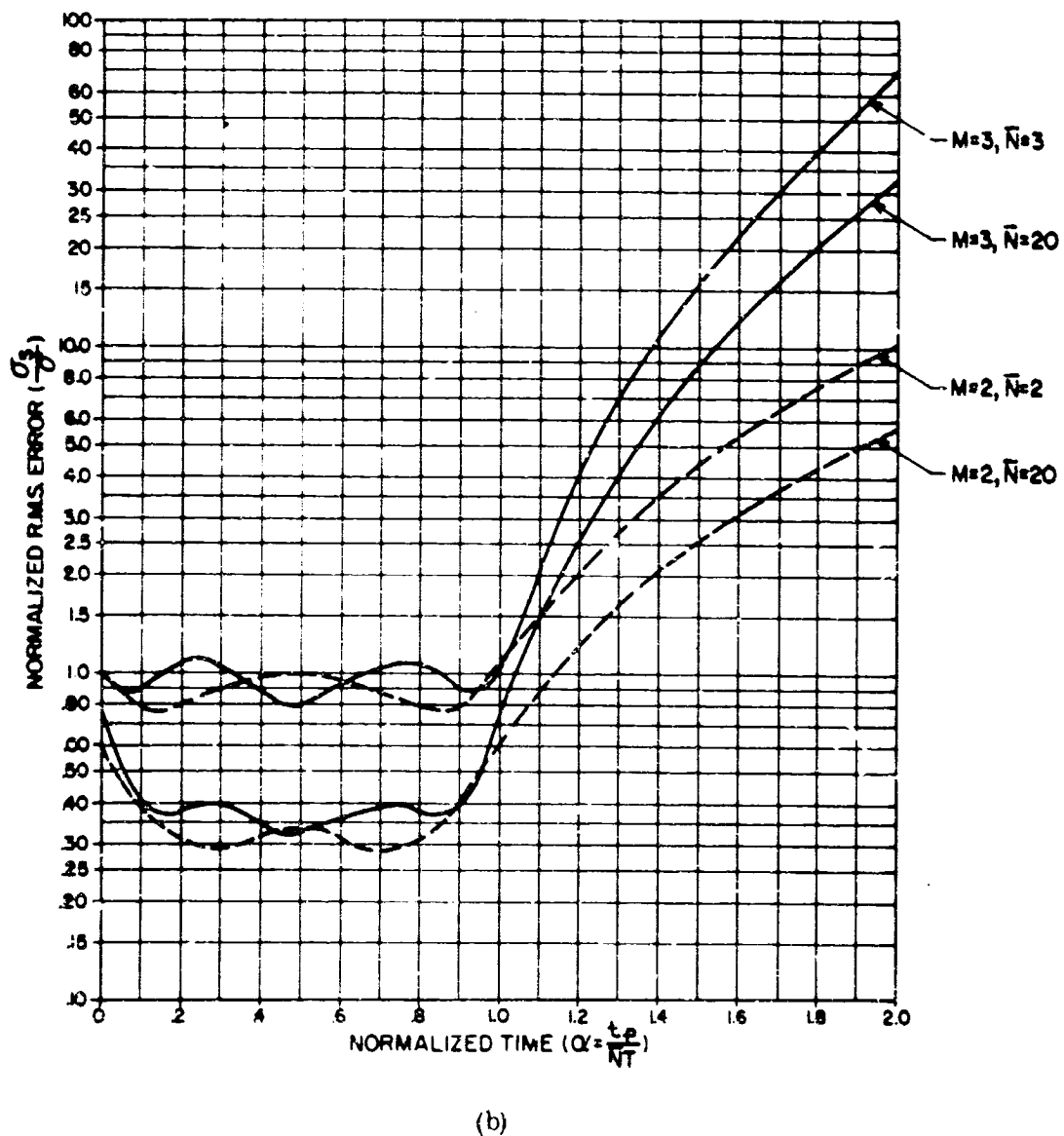
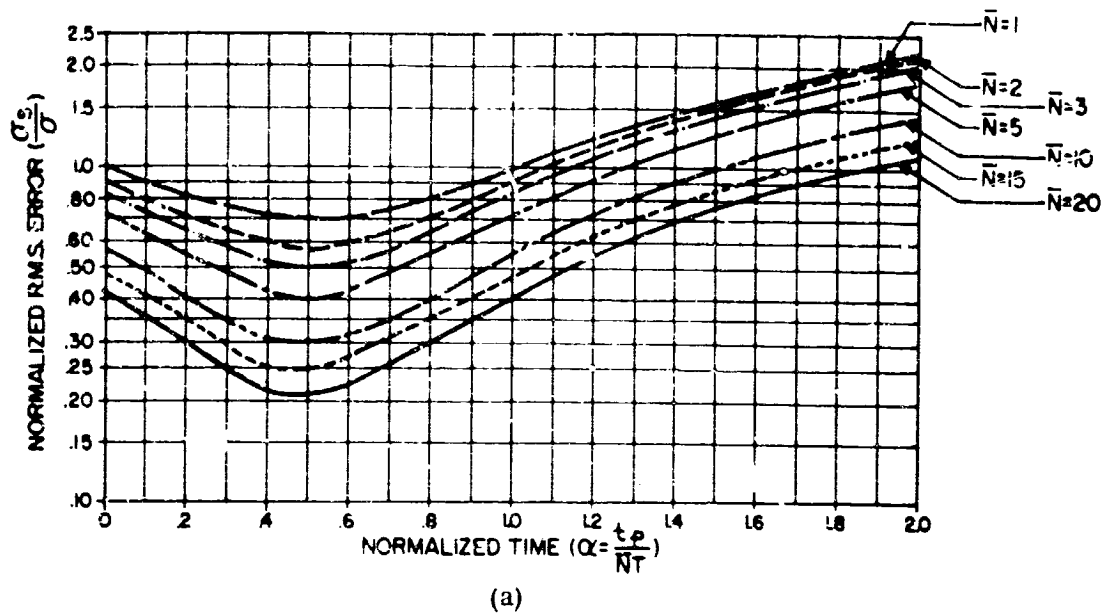


Figure 13.1. 11-1. Normalized RMS Error vs. Normalized Prediction Time

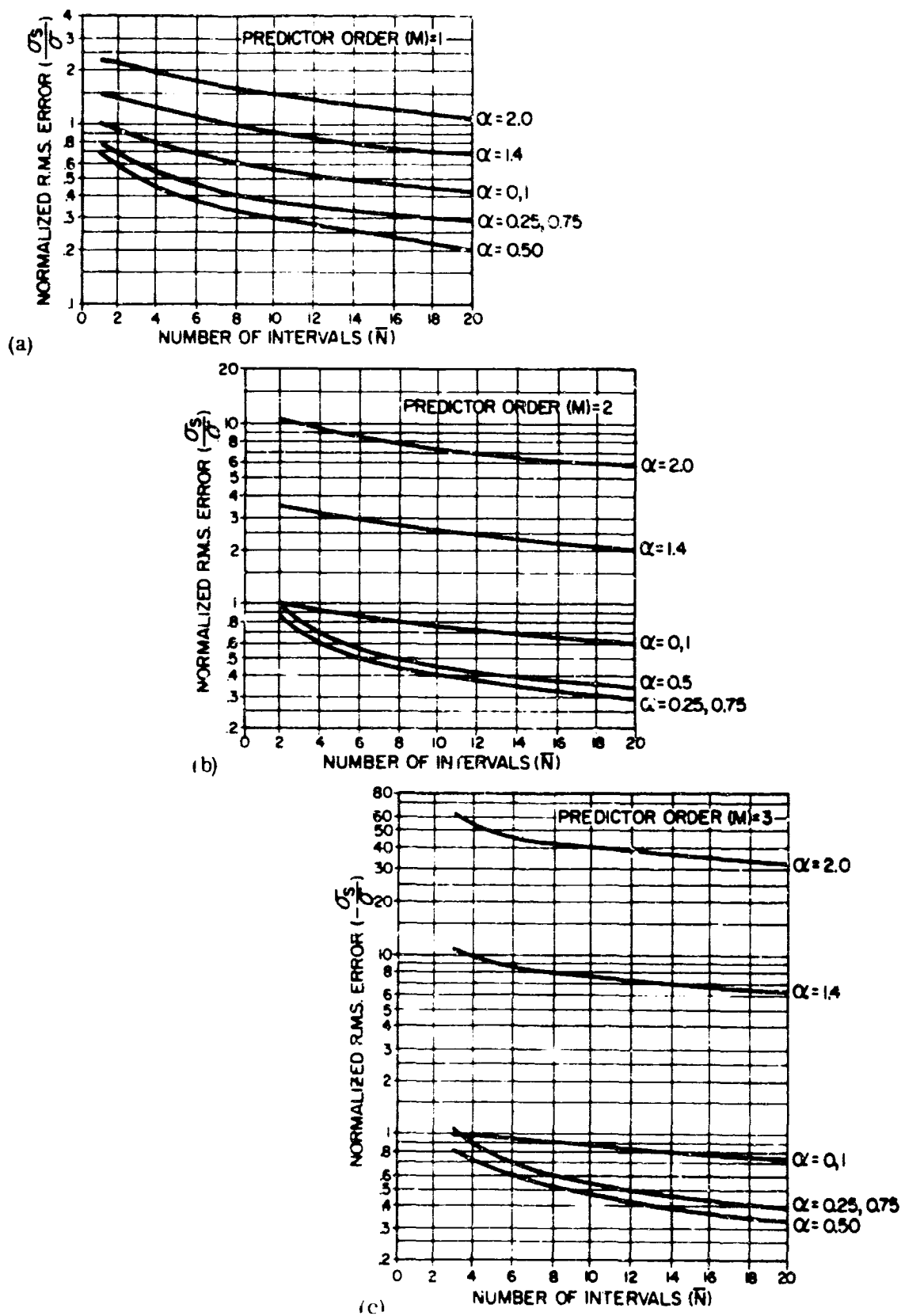
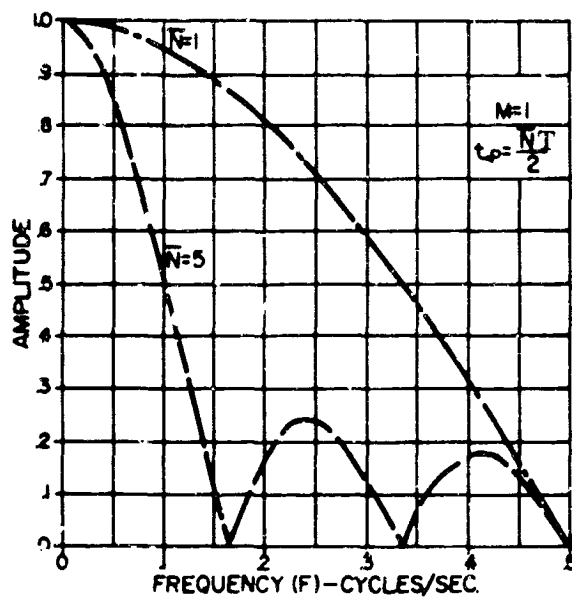
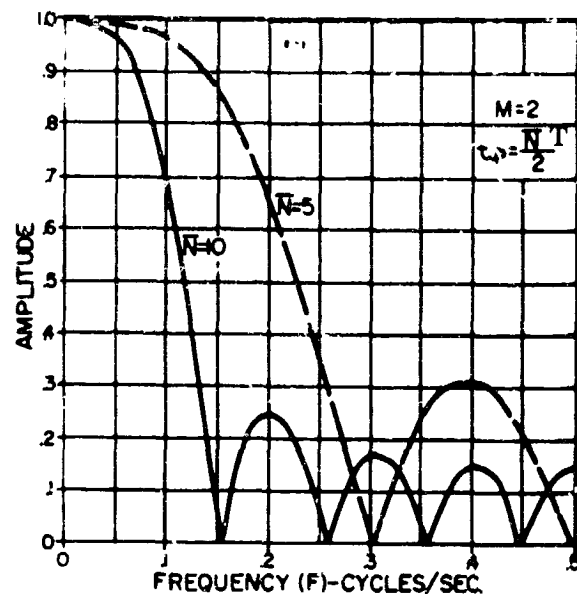


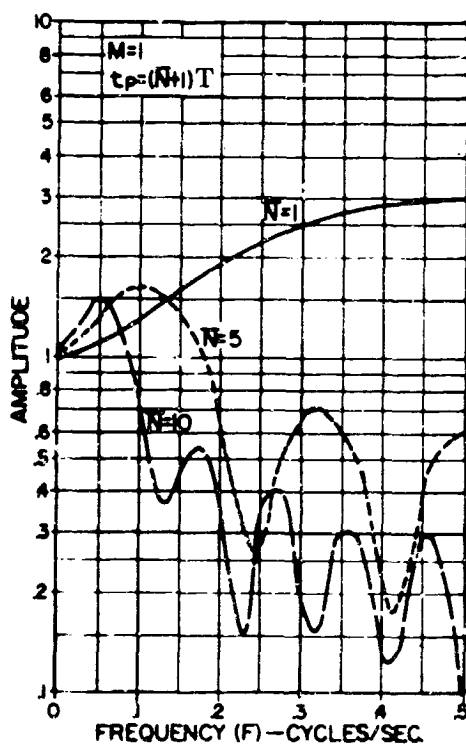
Figure 13.1 11-2 Normalized RMS Error as a Function of the Number of Date Samples $\bar{N} \geq 1$



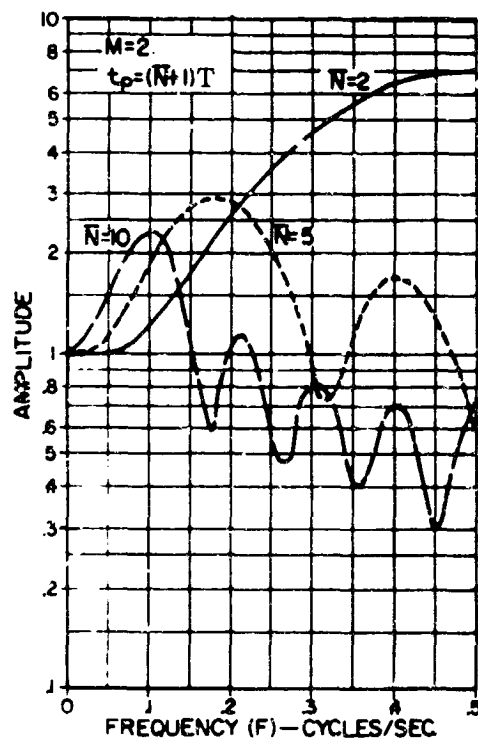
(a)



(b)



(c)



(d)

Figure 13.1. 11-4. Frequency Response of Polynomial Estimators

13.2 WEINER FILTERS

Another estimator, or filter, of considerable interest is the sampled data version of the "Weiner" filter. Many signals to be estimated in the presence of noise are themselves random, or at least appear to be so. Sometimes the lack of knowledge of the details of the signal structure, other than certain "average" properties to be discussed, force the use of the hypothesis of randomness.

Define the filter as before (Section 13.1) to be a weighting sequence w_m such the output of the filter $\hat{s}(t)$ is given by

$$\hat{s}(t) = \sum_{m=0}^{\overline{N}} w_m x_{\overline{N}-m}, \quad (13.2-1)$$

where x_k is a data sample at time t_k . This is the typical non-recursive linear filter form seen in previous sections. What will distinguish the filter of this section will be the assumptions under which the weights w_m are determined.

The assumptions under which the filter is designed are:

1. The data samples are given by a linear combination of the noiseless signal $s(t)$ plus noise $n(t)$ of zero mean. Non-zero mean noise can be compensated by subtracting the mean from the output. Mathematically,

$$x(t) = s(t) + n(t) \quad (13.2-2)$$

2. The output of the filter, $\hat{s}(t)$ is the estimate of the uncorrupted function $s(t)$ at some arbitrary, but fixed time t with respect to the data span.
3. Both noise and signal are described by their auto-correlation functions, their cross-correlation function, and their means. It is not necessary to assume that either noise or signal are stationary random processes at this point, which allows time varying filters.
4. Performance of the filter is judged by the ensemble averaged square error between $s(t)$ and $\hat{s}(t)$ (or its square root, the r.m.s error):

$$\begin{aligned} \epsilon^2 &= \text{average square error} \\ &= \left\langle (s(t) - \hat{s}(t))^2 \right\rangle \end{aligned} \quad (13.2-3)$$

Determination of the filter (the w_m 's) shall be such that the square error is minimized.

Proceeding with the derivation along these lines the square error is evaluated using equations of Section 11.1.

$$\begin{aligned}
 \epsilon^2 &= \left\langle \left(s(t) - \sum_{m=0}^{\bar{N}} w_m s(t_{\bar{N}-m}) - \sum_{m=0}^{\bar{N}} w_m n(t_{\bar{N}-m}) \right)^2 \right\rangle \quad (13.2-4) \\
 &= \left\langle s(t)^2 \right\rangle - 2 \sum_{m=0}^{\bar{N}} w_m \left\langle s(t) s(t_{\bar{N}-m}) \right\rangle \\
 &\quad - 2 \sum_{m=0}^{\bar{N}} w_m \left\langle s(t) n(t_{\bar{N}-m}) \right\rangle \\
 &\quad - 2 \sum_{m=0}^{\bar{N}} \sum_{k=0}^{\bar{N}} w_m w_k \left\langle s(t_{\bar{N}-m}) n(t_{\bar{N}-k}) \right\rangle \\
 &\quad + \sum_{m=0}^{\bar{N}} \sum_{k=0}^{\bar{N}} w_m w_k \left\langle s(t_{\bar{N}-m}) s(t_{\bar{N}-k}) \right\rangle \\
 &\quad + \sum_{m=0}^{\bar{N}} \sum_{k=0}^{\bar{N}} w_m w_k \left\langle n(t_{\bar{N}-m}) n(t_{\bar{N}-k}) \right\rangle.
 \end{aligned}$$

Minimize ϵ^2 with respect to all the w_r 's by taking the partial derivatives and setting them equal to zero. Using the relation

$$\frac{\partial (w_m w_k)}{\partial w_r} = w_m \delta_{rk} + w_k \delta_{mr}; \quad \delta_{mr} = \begin{cases} 1 & m = r \\ 0 & m \neq r \end{cases} \quad (13.2-5)$$

it is found that:

$$\begin{aligned}
 \frac{\partial \epsilon^2}{\partial w_r} = 0 &= -2 \left(\left\langle s(t) s(t_{\bar{N}-r}) \right\rangle + \left\langle s(t) n(t_{\bar{N}-r}) \right\rangle \right) \quad (13.2-6) \\
 &\quad - 2 \sum_{m=0}^{\bar{N}} w_m \left\langle s(t_{\bar{N}-m}) n(t_{\bar{N}-r}) \right\rangle \\
 &\quad - 2 \sum_{m=0}^{\bar{N}} w_m \left\langle s(t_{\bar{N}-r}) n(t_{\bar{N}-m}) \right\rangle \\
 &\quad + 2 \sum_{m=0}^{\bar{N}} w_m \left(\left\langle s(t_{\bar{N}-r}) s(t_{\bar{N}-m}) \right\rangle + \left\langle n(t_{\bar{N}-r}) n(t_{\bar{N}-m}) \right\rangle \right);
 \end{aligned}$$

$$r = 0, 1, \dots, \bar{N}$$

where the fact that

$$\langle s(t_{\bar{N}-r}) s(t_{\bar{N}-m}) \rangle = \langle s(t_{\bar{N}-m}) s(t_{\bar{N}-r}) \rangle \quad (13.2-7)$$

has been used.

To solve these equations convert them to matrix form. Define the matrices:

1. The signal sample correlation matrix

(13.2-8)

$$[\Phi_{ss}]_{mr} = [\Phi_{ss}]_{rm} = \langle s(t_{\bar{N}-r}) s(t_{\bar{N}-m}) \rangle ; r, m = 0, 1, \dots, \bar{N};$$

2. The noise sample correlation matrix

(13.2-9)

$$[\Phi_{nn}]_{mr} = [\Phi_{nn}]_{rm} = \langle n(t_{\bar{N}-r}) n(t_{\bar{N}-m}) \rangle ; r, m = 0, 1, \dots, \bar{N};$$

3. The signal-noise cross-correlation matrix

(13.2-10)

$$[\Phi_{sn}]_{mr} = \langle s(t_{\bar{N}-m}) n(t_{\bar{N}-r}) \rangle ; r, m = 0, 1, \dots, \bar{N};$$

4. The prediction time signal-sample time signal correlation matrix (a column vector)

$$[\Phi_s^t]_r = \langle s(t) s(t_{\bar{N}-r}) \rangle ; r = 0, 1, \dots, \bar{N}; \quad (13.2-11)$$

5. The prediction time signal-sample time noise cross-correlation matrix (a column vector)

$$[\Phi_n^t]_r = \langle s(t) n(t_{\bar{N}-r}) \rangle ; r = 0, 1, \dots, \bar{N}; \quad (13.2-12)$$

6. The weight vector

$$[W]_r = w_r ; r = 0, 1, \dots, \bar{N} \quad (13.2-13)$$

Equation (13.2-6) for the weight vector may now be rewritten as

$$[\Phi_{ss} + \Phi_{nn} - \Phi_{sn} - \tilde{\Phi}_{sn}] W = \Phi_s^t + \Phi_n^t \quad (13.2-14)$$

where $\tilde{\Phi}_{sn}$ is the transpose of Φ_{sn} :

$$[\tilde{\Phi}_{sn}]_{rm} = [\Phi_{sn}]_{mr} \quad (13.2-15)$$

A formal solution to equation (13.2-14) can be written; namely,

$$W = [\Phi_{ss} + \Phi_{nn} - \Phi_{sn} - \tilde{\Phi}_{sn}]^{-1} [\Phi_s^t + \Phi_n^t] \quad (13.2-16)$$

where A^{-1} = the inverse of A.

The words "formal solution" are used because it depends on the existence of the inverse matrix. An inverse of a matrix exists if its determinant is not zero. That is, the solution 13.2-16 holds if

$$|\Phi_{ss} + \Phi_{nn} - \Phi_{sn} - \tilde{\Phi}_{sn}| \neq 0. \quad (13.2-7)$$

In Section 13.1.1 (equations 13.1.1-21 etc.) the problem of the existence of the inverse the matrix C is discussed. The elements of C were the auto-correlation function of a stationary noise function $n(t)$. (13.2-7) which appears quite complicated as it is the sum of four matrices, is actually the auto-correlation function of the difference between signal and noise. If it is assumed that these processes are stationary, then, by Section 13.1.1, the matrix of equation (13.2-16) also possesses an inverse. Therefore, at least for stationary processes the "optimum" filter can be found.

13.2.1 Specialization of the Solution to an Important Case

Assume here that the signal and noise are uncorrelated as, for example, a transmitted speech waveform corrupted by receiver noise. Mathematically phrased this means

$$\Phi_{sn} = 0 \quad (13.2.1-1a)$$

and

$$\Phi_n^t = 0 \quad (13.2.1-1b)$$

Therefore, in this instance,

$$W = [\Phi_{ss} + \Phi_{nn}]^{-1} \Phi_s^t \quad (13.2.1-2)$$

14. THE FAST FOURIER TRANSFORM AND VARIOUS MECHANIZATIONS

The fast Fourier transform is a significantly more rapid method of computing the discrete Fourier transform (Section 2.6) than has been used conventionally. By taking advantage of the periodic properties of the quantity $W^j = e^{2\pi ij/N}$ and certain symmetries when N is highly composite, namely composed of many factors, the amount of computation may be reduced considerably from that using the definition directly.

Recall the definition of the discrete Fourier transform (DFT1)

$$X_n = \sum_{j=0}^{N-1} x_j W^{-jn}, \quad W = e^{i \frac{2\pi}{N}} \quad (2.6.1-)$$

with the inverse transform given by (IDFT1)

$$x_j = \frac{1}{N} \sum_{n=0}^{N-1} X_n W^{jn} \quad (2.6.1-)$$

This particular form of the transform is not the only form that may be used, but it has become conventional in the literature and will be discussed subsequently. Another convenient version of the transform and its relation to the one used in this section is also discussed in Section 2.6.1. It may be noted that the discrete Fourier transform and its inverse have essentially the same form as far as numerical calculations are concerned; only the proportionality factor and the sign of the exponent of W differ.

The fundamental version of the algorithm to be described here is that of Cooley and Tukey which essentially resolves an N point transform to many smaller transforms. A slight modification, decimation in frequency, is also discussed. The algorithm will be specialized in detail to the case where N is a power of two; $N = 2^m$.

Subsequent discussion shows that there are two distinct modes of mechanizing the fast transform, serial and parallel, as for the conventional method of discrete transformation. By serial mechanization the successive performance of each basic mathematical operation of addition, multiplication, or storing is meant. Parallel mechanization, on the other hand, allows a number of simultaneous multiplication, addition, and storing operations. More hardware, of course, is required for the parallel mechanization. The conventional method of computing the transform is also capable of being computed in either mode. Examination of (2.6.1-1) indicates that, in its most simple form, the transform for each X_n may be computed first for $n = 0$, then $n = 1$, etc., with N successive multiplications and additions (complex) for each n requiring N^2 total operations sequentially. If the hardware is available the N different X_n may be computed individually at the same time by N multipliers and adders using only N operations. When speed is mandatory it may be necessary to mechanize the parallel version.

The fast Fourier transform algorithm, when mechanized has an advantage in speed over the conventional transform by a factor proportional to $N/\log_2 N$ when parallel versions are compared and when serial versions are compared. In both modes of transformation the parallel method is N times faster than the serial method.

Storage necessary for the fast algorithm can be that for the N data points. In each phase of the algorithm, the various computations can be done "in place", that is as new quantities are formed they are stored in the places occupied by their constituents. The discrete transform, at the end of the algorithm, resides in the original data locations. Unfortunately, the transform is in "bit reversed" order, to be discussed subsequently if the data was in original time order. By "bit reversing" the data the transform can be obtained in normal order if necessary. Therefore, when the procedure is mechanized by a general purpose computer, "bit reversing" routines are useful and require only a few registers as additional storage.

Accuracy of the fast algorithm is also approximately better than the direct method by a factor of $N/\log_2 N$ as that many less operations are necessary.

14.1 BASIS OF THE FAST TRANSFORM ALGORITHM

In subsequent discussion a paper by J. Cooley presented at the SHARE XXIX, August 16-18, 1967, the contents of the IEEE Transactions on Audio and Acoustics, June 1967, and the paper of J. W. Cooley and J. W. Tukey "An Algorithm for the Machine Calculation of Complex Fourier Series", Mathematics of Computation, April, 1965, have been heavily relied upon in describing basic versions of the $N \log_2 N$ class of algorithms. A more complete set of references is given in the Bibliography.

The discrete transform is repeated:

$$X_n = \sum_{j=0}^{N-1} x_j \bar{W}^{jn} \quad (14.1-1)$$

with $\bar{W} = e^{-\frac{2\pi i}{N}}$ used to be more compatible with the available literature. Various definitions used are by no means uniform. Definition of the sign of the transform exponent has been maintained at minus for the transform from the time to frequency domain throughout this document consistent with popular scientific and engineering usage in the last quarter century with continuous functions.

Assume at first that N is composed of two arbitrary integer factors $N = r \cdot s$ and define a one-to-one mapping between the integers used in (14.1-1) and a new set of variables

$$n = n_1 r + n_0 \quad (14.1-2)$$

with

$$0 \leq n < N, \quad 0 \leq n_0 < r, \quad 0 \leq n_1 < s$$

and

$$j = j_1 s + j_0 \quad (14.1-3)$$

with

$$0 \leq j < N, \quad 0 \leq j_0 < s, \quad 0 \leq j_1 < r.$$

Effectively, this allows considering X_n and x_j as a pair of two dimensional arrays enabling transformation in two steps through formation of an intermediate transform.

Thus

$$X(n_1, n_0) \equiv X_n$$

$$= \sum_{j=0}^{N-1} x_j \bar{W}^{jn} \quad (14.1-4)$$

$$= \sum_{j_0=0}^{s-1} \sum_{j_1=0}^{r-1} x(j_1, j_0) \bar{W}^{n(j_1 s + j_0)}$$

with the definition

$$x(j_1, j_0) = x_j \quad (14.1-5)$$

As

$$\bar{W}^{nj_1 s} = \bar{W}^{n_0 j_1 s} \quad (14.1-6)$$

since $\bar{W}^{kN} = 1$, $X(n_1, n_0)$ may be written

$$X(n_1, n_0) = \sum_{j_0=0}^{s-1} x_1(n_0, j_0) \bar{W}^{(n_0 r + n_0) j_0} \quad (14.1-7)$$

with

$$X_1(n_0, j_0) = \sum_{j_1=0}^{r-1} x(n_1, n_0) \bar{W}^{n_0 j_1 s}$$

Now the basis for the speed of the fast transform algorithm lies in these last equations. There are N elements in the array x_1 each requiring r operations or a total of Nr operations (complex multiplications and additions). Computing all the X_n from the x_1 's requires, similarly Ns operations. Therefore, by merely breaking the discrete transform into two steps via (14.1-7) the total number of operations is

$$N(r + s)$$

which is substantially less than N^2 . For example with $r = 9$ and $s = 10$ a factor of about 4.7 is saved in the amount of computation.

Notice next that both sections of (14.1-7) are capable of further decomposition in an identical manner as long as r and/or s are themselves composite numbers. Therefore, by successive application of (14.1-7) an m -step algorithm for m factors of N would require

$$N(r_1 + r_2 + \dots + r_m)$$

operations where

$$N = \prod_{i=1}^m r_i$$

More factors provide the greatest gains. If all the factors r_i are assumed equal to r then

$$m = \log_r N \quad (14.1-8)$$

and the number of operations is

$$T(r) = r N \log_r N \quad (14.1-9)$$

In order to find approximately the most efficient base factor compute the ratio of $T(r)$ to that for a base of two for comparison:

$$\frac{T(r)}{T(2)} = \frac{r N \log_r N}{2 N \log_2 N}$$

(14.1-10)

$$= \frac{r}{2 \log_2 r}$$

Following Cooley and Tukey the ratio for various r follows:

r	$\frac{r}{2 \log_2 r}$
2	1.00
3	.94
4	1.00
5	1.08
6	1.16
7	1.25
8	1.34
9	1.41
10	1.51

While a base factor of 3 is most efficient mathematically, the use of 2 and 4, which have significant practical advantages, differs but by a few percent (6%). In fact sensitivity to small changes in the base is small.

14.2 THE BASE FACTOR 2 ALGORITHM

Many variants of the base factor 2 algorithm are possible. Only the methods which do not require additional storage beyond that required for the data will be discussed. The methods differ in two basic factors, how the powers of $\bar{W}$ are applied, and in the initial and final storage positions of the data and its transform. This second property of data and transform position is denoted by "bit reversal" in contradistinction to natural time order. The term refers to storage in a position whose index is the number obtained by reversing the order of the binary digits in the natural sequence position in binary representation.

The first method to be discussed has been called "time decimation" and will be given in two forms, bit reversal of the transform and bit reversal of the data. The nomenclature of time decimation comes about because of the particular method of factoring the powers of $\bar{W}$ in terms of the successive (binary) digits of "j", the time index (see Section 2.6.2). When the factoring of $\bar{W}$ is in terms of successive digits of "n", the frequency

index, the implementation of the fast transform is often denoted as "frequency decimation". These methods will be discussed finally.

In all, four methods will be discussed.

14.2.1 Time Decimation With a Bit Reversed Transform

In this algorithm form the indices are specified by

$$j = j_{m-1} 2^{m-1} + \dots + j_1 \cdot 2 + j_0 \quad (14.2.1-1)$$

$$n = n_{m-1} 2^{m-1} + \dots + n_1 \cdot 2 + n_0$$

with $m = \log_2 N$ ($N = 2^m$) and j_k and n_k are equal to either 0 or 1. Obviously, j and n are simply integers expressed in binary form with the most significant digits j_{m-1} and n_{m-1} , respectively. Development of this form and subsequent forms will use a particularly indicative notation. Arrays will be denoted by their subscripts in binary form with the m digits written from left to right in order or descending significance. New arrays defined in each development use the same convention. For example, the array

$$y(0, 0, 1, 0, 1) = y(5) \quad (14.2.1-2a)$$

in decimal notation. In bit reversed form $y(0, 0, 1, 0, 1)$ becomes

$$y(1, 0, 1, 0, 0) = y(24). \quad (14.2.1-2b)$$

Then

$$X_n \equiv X(n_{m-1}, \dots, n_0) \quad (14.2.1-3)$$

$$= \sum_{j_0} \sum_{j_1} \dots \sum_{j_{m-1}} X(j_{m-1}, \dots, j_0) \bar{W}^{n_{j_{m-1}} 2^{m-1} + \dots + n_{j_0}}$$

with the sum over $j_k = 0, 1$. As $\bar{W}^{kN} = 1$ (k an integer) then

$$\bar{W}^{nj} = \bar{W}^{n_j m-1} 2^{m-1} \bar{W}^{n_j m-2} 2^{m-2} \dots \bar{W}^{n_j 0} \quad (14.2.1-4)$$

$$= \bar{W}^{n_0 j_{m-1} 2^{m-1}} \bar{W}^{(n_1 \cdot 2 + n_0) j_{m-2} 2^{m-2}} \dots \bar{W}^{n_j 0}$$

may be written. The utility of this particular factoring lies in the separation of the various summation indices (j_k) into separate terms allowing "nesting" of successive summations into a recursive algorithm: the fast Fourier transform.

Sum over the innermost index j_{m-1} and obtain

$$X(n_{m-1}, \dots, n_0) = \sum_{j_0} \sum_{j_1} \dots \sum_{j_{m-2}} x_1(n_0, j_{m-2}, \dots, j_0) \bar{W}^{(n_1 \cdot 2 + n_0) j_{m-2} 2^{m-2}} \dots \bar{W}^{n_j 0} \quad (14.2.1-5)$$

with x_1 given by

$$x_1(n_0, j_{m-2}, \dots, j_0) = \sum_{j_{m-1}} x(j_{m-1}, \dots, j_0) \bar{W}^{n_0 j_{m-1} 2^{m-1}} \quad (14.2.1-6)$$

Forming x_1 is the first basic step of the algorithm and it should be noted that, as n_0, j_{m-2} , etc. are allowed to vary, N sums are formed. The next summation, of course, is over j_{m-2} where N quantities x_2 will be computed.

Namely,

$$x_2(n_0, n_1, j_{m-3}, \dots, j_0) = \sum_{j_{m-2}} x_1(n_0, j_{m-2}, \dots, j_0) \bar{W}^{(n_1 \cdot 2 + n_0) j_{m-2} 2^{m-2}} \quad (14.2.1-7)$$

Continuing, the general term is

$$\begin{aligned}
 & x_q (n_0, \dots, n_{q-2}, n_{q-1}, j_{m-q-1}, \dots, j_0) \\
 &= \sum_{j_{m-q}=0}^1 \left[x_{q-1} (n_0, \dots, n_{q-2}, j_{m-q}, \dots, j_0) \right. \\
 & \quad \left. \cdot \bar{W}^{(n_{q-1} \cdot 2^{q-1} + \dots + n_0) j_{m-q} 2^{m-q}} \right]
 \end{aligned} \tag{14.2.1-8}$$

with $q = 1, 2, \dots, m = \log_2 N$. Writing the two term summation in open form:

$$\begin{aligned}
 & x_q (n_0, \dots, n_{q-1}, j_{m-q-1}, \dots, j_0) \\
 &= x_{q-1} (n_0, \dots, n_{q-2}, 0, \dots, j_0) \\
 & \quad + x_{q-1} (n_0, \dots, n_{q-2}, 1, \dots, j_0) \cdot W^{(n_{q-1} 2^{q-1} + \dots + n_0) 2^{m-q}}
 \end{aligned} \tag{14.2.1-9}$$

it may be observed that only those two x_{q-1} with 0 and 1 in the q^{th} bit position are involved in the computation for a particular x_q . This may be seen more clearly if the quantities

$$\begin{aligned}
 K_R &= n_0 2^{q-2} + \dots + n_{q-2} \\
 J &= j_{m-q-1} 2^{m-q-1} + \dots + j_0
 \end{aligned} \tag{14.2.1-10}$$

and

$$I = J + K_R \cdot 2^{m-q+1}$$

are defined. Notice that the significant place 2^{m-q} has not been represented.

Then (14.2.1-9) may be written in a more cogent form

$$x_q(I) = x_{q-1}(I) + x_{q-1}(I + 2^{m-q}) \bar{W}^K \cdot 2^{m-q} \quad (14.2.1-11)$$

and

$$x_q(I + 2^{m-q}) = x_{q-1}(I) - x_{q-1}(I + 2^{m-q}) \bar{W}^K \cdot 2^{m-q}$$

where K is the natural form of the bit reversed K_R :

$$K = n_{q-2} 2^{q-2} + \dots + n_0 \quad (14.2.1-12)$$

and the identity

$$\bar{W}^{n_{q-1} j_{m-q} 2^{m-1}} = (-1)^{n_{q-1} j_{m-q}} \quad (14.2.1-13)$$

has been used.

It can be seen upon examining (14.2.1-9), or the more useful form (14.2.1-11) for programming, exactly what is required for implementing the algorithm. These statements have been summarized previously but may now be stated in detail.

Only two values computed in a preceding iteration (the $q-1^{st}$) are needed to compute two values at the next stage. Furthermore, these values are not needed in the computation of any other x_q and, therefore, the new values may be stored in place of the old ones. No new storage is needed for the transform.

The number of serial mathematical operations can also be clearly stated at this point. At each of the q stages of the algorithm ($q = 1, 2, \dots, m = \log_2 N$) $N/2$ of the two point transforms given by (14.2.1-11) must be formed. Examination of these equations indicate one complex multiplication and two complex additions per two point transform. Therefore, there are a total of $(N/2) \log_2 N$ complex multiplications and $N \log_2 N$ complex additions for a complete discrete transform using the base two algorithm. Translated into real operations, $2N \log_2 N$ multiplications and $3N \log_2 N$ additions are required. Also, as no other values are needed, the computation for $N/2$ pairs may be performed simultaneously or in parallel. Only $m = \log_2 N$ parallel operations are necessary.

Finally, examination of the final step of the algorithm ($q = m = \log_2 N$) shows that the output transform is stored in bit reversed locations

$$X(n_m, \dots, n_0) = x_m(n_0, \dots, n_m) \quad (14.2.1-14)$$

The transform must be "shuffled" if natural order is desired. Powers of $\bar{W}$ used are in natural order as may be seen from (14.2.1-11).

An illustrative program, in a bastardization of the FORTRAN language, may be written.

```

DIMENSION X ( . . . )
COMPLEX X
INTEGER Q
Do 1 Q = 1, M
Do 1 KR = 0, 2Q-1 -1
**U =  $\bar{W}$  ** (K * 2M-Q)
Do 1 J = 0, 2M-Q -1

I = J + KR * 2M-Q+1

T = X(I + 2M-Q) * U
X(I + 2M-Q) = X(I) - T
X(I) = X(I) + T
1 CONTINUE

```

** K is used though it is the bit reversed K_R that was defined. As the sines and/or cosines must be stored in a table or generated by a special purpose machine they can be stored or generated in bit reversed form and K_R used as the input; K_R need not be reversed in the loop of the procedure above.

As certain multiplications are simply by ± 1 and $\pm i$ they may, if the complexity is warranted, be treated as individual cases. $\bar{W}$ can also be extracted from a table of sines for one quadrant. For the inverse transform it is only the sign of $\bar{W}$ that is changed, otherwise the algorithm remains the same within a constant factor.

A flow graph is often useful in assisting in understanding the processing involved in both serial and parallel implementation.

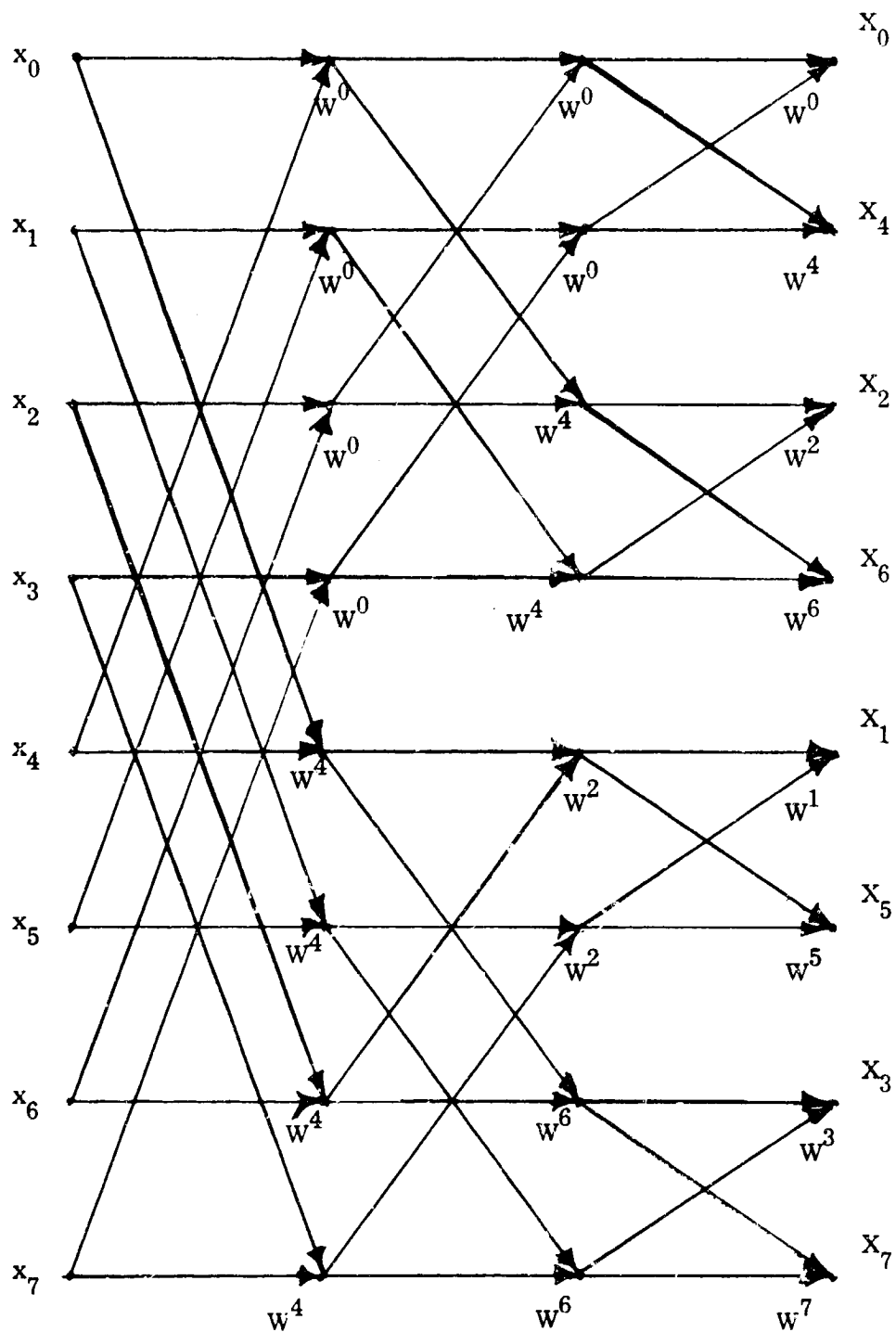


Figure 14.2.1-1 Flow Graph of Time Decimation Method
With Naturally Ordered Input Sequence

This graph is simply (14.2.1-11) illustrated. Nodes on a particular vertical line indicate operations at the q^{th} iteration. Iterations progress from left to right. Arrows indicate which nodes from a previous iteration contribute to a new computation. The results of computation may be considered to be stored (in complex form) at a node at a given stage. Quantities contributing to a node via an arrow are assumed to be summed at a node. If an arrow has a symbol adjacent to it the quantity passing over that path is weighted (multiplied) by the value of the symbol.

Figure 14.2.1-1 assists in making more apparent the assertion that "in place" computation may be used as it may be seen that any two nodes used at a given stage are the only contributors to nodes on the same horizontal lines directly to the right.

14.2.2 Time Decimation With a Bit Reversed Input Sequence

Another form of the algorithm is immediately suggested by the previous one. In this case the input sequence is scrambled first by a bit reversal operation allowing the transform to appear in natural order. Another slight advantage might be the appearance of powers of $\bar{W}$ in natural order. This form of the algorithm does not differ from the preceding version in so far as amount of computation, computation without additional storage, and allowing parallel computation are concerned.

The derivation of the basic formulae will be shortened a bit from that of the previous method as both have essentially the same structure.

Indices n and j are defined in binary notation as before. In this form however the bit reversed array

$$x_0(j_0, \dots, j_{m-1}) = x(j_{m-1}, \dots, j_0) \quad (14.2.2-1)$$

is defined and is used in the first stage of the algorithm. Equation (14.1-1) may be written this time

$$X_n = X(n_{m-1}, \dots, n_0) \quad (14.2.2-2)$$

$$= \sum_{j_0} \sum_{j_1} \dots \sum_{j_{m-1}} x_0(j_0, \dots, j_{m-1}) \bar{W}^{nj_{m-1}2^{m-1} + nj_0}$$

Sum over the innermost index and define

$$x_2(j_0, \dots, j_{m-2}, n_0) = \sum_{j_{m-1}} x_0(j_0, \dots, j_{m-1}) \bar{W}^{n_0 j_{m-1} 2^{m-1}} \quad (14.2.2-3)$$

where (14.2.1-4) has been used. Similarly successive sums may be defined

$$x_2(j_0, \dots, j_{m-3}, n_1, n_0) = \sum_{j_{m-2}} x_1(j_0, \dots, j_{m-2}, n_0) \bar{W}^{(n_1 \cdot 2 + n_0) j_{m-2} 2^{m-2}}$$

(14.2.2-4)

and generally

$$\begin{aligned} & x_q(j_0, \dots, j_{m-q-1}, n_{q-1}, n_{q-2}, \dots, n_0) \\ &= \sum_{j_{m-q}=0}^1 \left[x_{q-1}(j_0, \dots, j_{m-q}, n_{q-2}, \dots, n_0) \right. \\ & \quad \left. \cdot \bar{W}^{(n_{q-1} \cdot 2^{q-1} + \dots + n_0) j_{m-q} 2^{m-q}} \right] \end{aligned} \quad (14.2.2-5)$$

with $q = 1, 2, \dots, m = \log_2 N$. At the last stage:

$$X_n = x_m(n_{m-1}, \dots, n_0)$$

In open form

$$\begin{aligned} & X_q(j_0, \dots, j_{m-q-1}, n_{q-1}, \dots, n_0) \\ &= x_{q-1}(j_0, \dots, 0, n_{q-2}, \dots, n_0) \\ & \quad + x_{q-1}(j_0, \dots, 1, n_{q-2}, \dots, n_0) \cdot \bar{W}^{(n_{q-1} \cdot 2^{q-1} + \dots + n_0) 2^{m-q}} \end{aligned} \quad (14.2.2-6)$$

As before quantities may be defined to clarify these equations

$$K = n_{q-2} 2^{q-2} + \dots + n_0, \quad (14.2.2-7)$$

$$J = j_0 2^{m-q-1} + \dots + j_{m-q-1},$$

and

$$I = K + J \cdot 2^q.$$

With these indices the fast algorithm with shuffled input and naturally order transform may be written

$$x_q(I) = x_{q-1}(I) + x_{q-1}(I + 2^{q-1}) \overline{W}^K \cdot 2^{m-q} \quad (14.2.2-8)$$

and

$$x_q(I + 2^{q-1}) = x_{q-1}(I) - x_{q-1}(I + 2^{q-1}) \overline{W}^K \cdot 2^{m-q}$$

where (14.2.1-13) has been used. It may be seen that K is used in natural bit order throughout.

An illustrative program for this case follows.

```
DIMENSION X ( . . . )
```

```
COMPLEX X
```

```
INTEGER Q
```

```
Do 1 Q = 1, M
```

```
Do 1 K = 0, 2Q-1-1
```

```
U =  $\overline{W}$  ** (K * 2M-Q)
```

```
Do 1 J = 0, 2M-Q-1
```

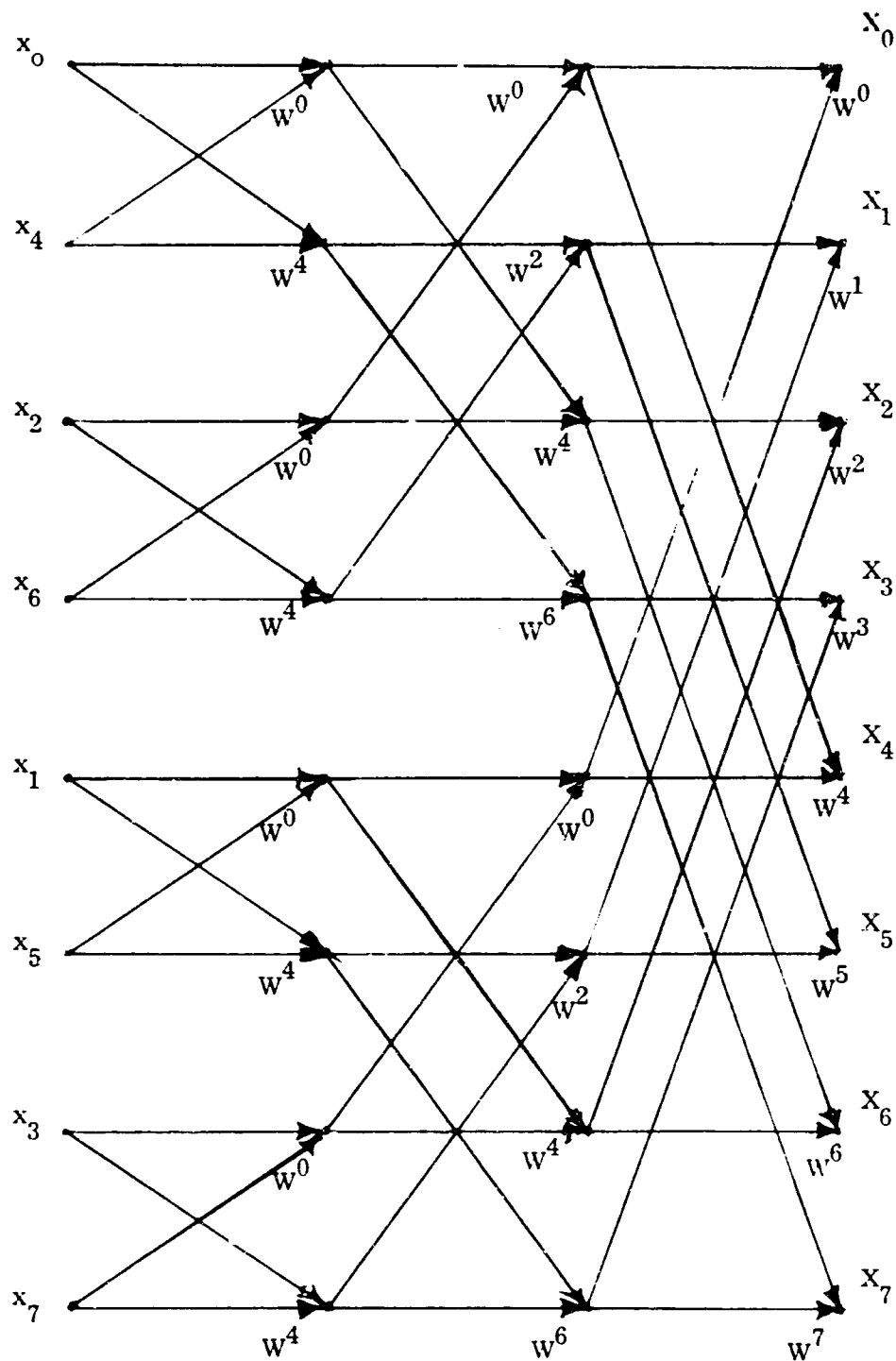


Figure 14.2.2-1 Flow Graph of Time Decimation Method
With Bit Reversed Input Sequence

$$I = K + J \cdot 2^Q$$

$$T = X(I + 2^{Q-1}) * U$$

$$X(I + 2^{Q-1}) = X(I) - T$$

$$X(I) = X(I) + T$$

1 CONTINUE

A flow chart of this form of the algorithm follows (Figure 14.2.2-1)

14.2.3 Frequency Decimation With a Bit Reversed Transform

Examination of the previous two methods indicates that the fundamental difference between the two methods is the "addressing" of the successive arrays developed at each stage of the algorithms. A certain factoring of $\bar{W}$ was assumed (14.2.1-4). Due to symmetry in the indices n and j an equivalent factoring may be used.

$$\begin{aligned} \bar{W}^{nj} = \bar{W}^{j_0 n_{m-1} 2^{m-1} + \dots + j_{m-2} 2^{m-2} + \dots + j_0} n_1 2^1 \\ \cdot \bar{W}^{(j_{m-1} 2^{m-1} + \dots + j_0) n_0} \end{aligned} \quad (14.2.3-1)$$

As further developments are quite similar to that of the previous methods only results will be stated. Define as usual, by bit position,

$$\begin{aligned} X_n &= X(n_{m-1}, n_{m-2}, \dots, n_0) \\ &= \sum_{j_0} \sum_{j_1} \dots \sum_{j_{m-1}} x(j_{m-1}, \dots, j_0) \bar{W}^{jn} \end{aligned} \quad (14.2.3-2)$$

Summing over the inner sums first, successive sums may be obtained using (14.2.3-1)

$$x_1(n_0, j_{m-2}, \dots, j_0) = \sum_{j_{m-1}} x(j_{m-1}, \dots, j_0) \bar{W}^{j n_0} \quad (14.2.3-3)$$

$$x_2(n_0, n_1, j_{m-3}, \dots, j_0) = \sum_{j_{m-2}} x_1(n_0, j_{m-2}, \dots, j_0) \cdot \bar{W}^{(j_{m-2} 2^{m-2} + \dots + j_0) n_1 2}$$

and generally

$$x_q(n_0, \dots, n_{q-2}, n_{q-1}, j_{m-q-1}, \dots, j_0)$$

$$= \sum_{j_{m-q}=0}^1 \left[x_{q-1}(n_0, \dots, n_{q-2}, j_{m-q}, j_{m-q-1}, \dots, j_0) \right. \\ \left. \cdot \overline{W}^{(j_{m-q} 2^{m-q} + \dots + j_0) n_{q-1} 2^{q-1}} \right] \quad (14.2.3-4)$$

In open form

$$x_q(n_0, \dots, n_{q-1}, j_{m-q-1}, \dots, j_0)$$

$$= \left[x_{q-1}(n_0, \dots, 0, j_{m-q-1}, \dots, j_0) + (-1)^{n_{q-1}} x_{q-1}(n_0, \dots, 1, j_{m-q-1}, \dots, j_0) \right] \\ \cdot \overline{W}^{(j_{m-q-1} 2^{m-q-1} + \dots + j_0) n_{q-1} 2^{q-1}} \quad (14.2.3-5)$$

where the fact that

$$\overline{W}^{n_q 2^{m-1}} = (-1)^{n_{q-1}} \quad (14.2.3-6)$$

has been used.

Defining

$$J = j_{m-q-1} 2^{m-q-1} + \dots + j_0 \quad (14.2.3-7)$$

$$K_R = n_0 2^{q-2} + \dots + n_{q-2}$$

and

$$I = J + K_R \cdot 2^{m-q+1}$$

the simplified forms may be written:

$$x_q(I) = x_{q-1}(I) + x_{q-1}(I + 2^{m-q}) \quad (14.2.3-8)$$

and

$$x_q(I + 2^{m-q}) = \left[x_{q-1}(I) - x_{q-1}(I + 2^{m-q}) \right] \bar{W}^J \cdot 2^{q-1}$$

As in the other forms of the algorithm, the same number of operations are necessary for computation "in place" calculation may be performed, and parallel calculation may be used. Also, the exponent of $\bar{W}$ is in its natural order.

Bit reversal of the transform is obtained as

$$\begin{aligned} X_n &= X(n_{m-1}, \dots, n_0), \\ &= x_m(n_0, \dots, n_{m-1}). \end{aligned} \quad (14.2.3-9)$$

An illustrative FORTRAN program in abortive notation follows.

```

DIMENSION X ( . . . )
COMPLEX X
INTEGER Q
Do 1 Q = 1, M
Do 1 J = 0, 2M-Q - 1
U =  $\bar{W}^{** (J * 2^{Q-1})}$ 
Do 1 KR = 0, 2Q-1 - 1
T = (X(I) - X(I + 2M-Q)) * U

```

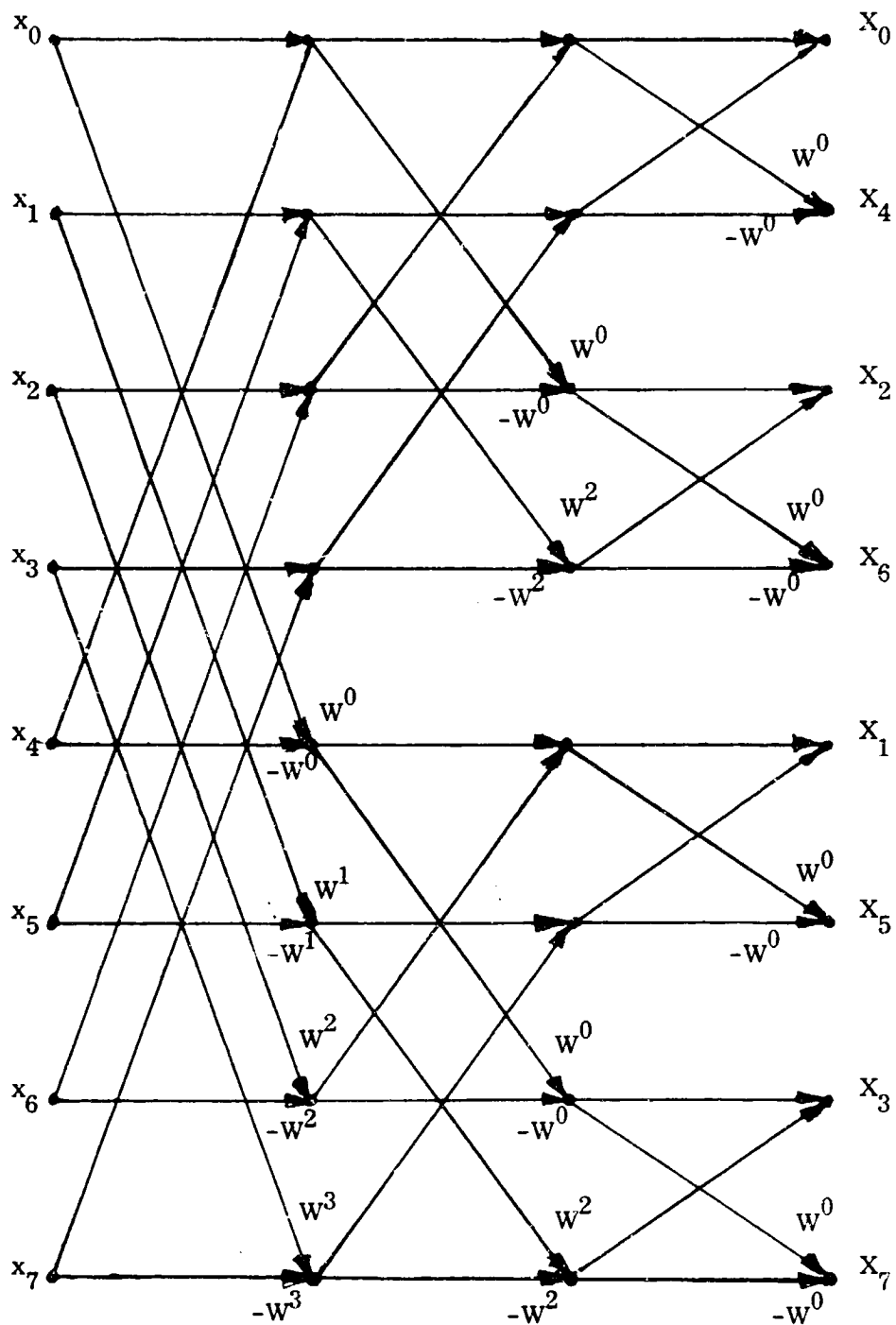


Figure 14.2.3-1 Flow Graph of Frequency Decimation Method With Naturally Ordered Input Sequence

$$X(I) = X(I) + X(I + 2^{M-Q})$$

$$X(I + 2^{M-Q}) = T$$

1 CONTINUE

A flow chart of this form of the algorithm follows (Figure 14.2.3-1).

14.2.4 Frequency Decimation With a Bit Reversed Input Sequence

A fourth version of the fast transform combines both bit reversal of the input sequence and frequency decimation. It is also the last of the basic in place methods. Combinations of the various methods can be used in any given problem, though the use of mixed methods in most situations is not necessary or desirable.

As this method follows immediately from the preceding ones it is also summarized.

Define the bit reversed input sequence

$$x_0(j_0, \dots, j_{m-1}) = x(j_{m-1}, \dots, j_0) \quad (14.2.4-1)$$

then

$$X_n = X(n_{m-1}, \dots, n_0) \quad (14.2.4-2)$$

$$= \sum_{j_0} \sum_{j_1} \dots \sum_{j_{m-1}} x_0(j_0, \dots, j_{m-1}) \bar{W}^{jn}$$

with binary index conventions as before. Using the representation (14.2.3-1) for $\bar{W}^{jn}$ successive sums may be obtained

$$x_1(j_0, \dots, j_{m-2}, n_0) = \sum_{j_{m-1}} x_0(j_0, \dots, j_{m-1}) \bar{W}^{jn_0} \quad (14.2.4-3)$$

$$x_2(j_0, \dots, j_{m-3}, n_1, n_0) = \sum_{j_{m-2}} x_1(j_0, \dots, j_{m-2}, n_0) \cdot \bar{W}^{(j_{m-2} 2^{m-2} + \dots + j_0) n_1 2}$$

and generally

$$\begin{aligned}
 & x_q(j_0, \dots, j_{m-q-1}, n_{q-1}, n_{q-2}, \dots, n_0) \\
 &= \sum_{j_{m-q}=0}^1 \left[x_{q-1}(j_0, \dots, j_{m-q}, n_{q-2}, \dots, n_0) \right. \\
 & \quad \left. \cdot \bar{W}^{(j_{m-q} 2^{m-2} + \dots + j_0) n_{q-1} 2^{q-1}} \right]
 \end{aligned} \tag{14.2.4-4}$$

In open form

$$\begin{aligned}
 & x_q(j_0, \dots, n_{q-1}, n_{q-2}, \dots, n_0) \\
 &= \left[x_{q-1}(j_0, \dots, 0, n_{q-2}, \dots, n_0) \right. \\
 & \quad \left. + (-1)^{n_{q-1}} x_{q-1}(j_0, \dots, 1, n_{q-2}, \dots, n_0) \right] \\
 & \quad \cdot \bar{W}^{(j_{m-q-1} 2^{m-q-1} + \dots + j_0) n_{q-1} 2^{q-1}}
 \end{aligned} \tag{14.2.4-5}$$

At the last stage of the algorithm

$$X(n_{m-1}, \dots, n_0) = x_m(n_{m-1}, \dots, n_0)$$

which is a naturally ordered transform.

For the simplified forms define

$$K = n_{q-2} 2^{q-2} + \dots + n_0 \tag{14.2.4-6}$$

$$J_R = j_0 2^{m-q-1} + \dots + j_{m-q-1}$$

and

$$I = K + J_R \cdot 2^q$$

then

$$x_q(I) = x_{q-1}(I) + x_{q-1}(I + 2^{q-1})$$

and

$$x_q(I + 2^{q-1}) = [x_{q-1}(I) - x_{q-1}(I + 2^{q-1})] \bar{W}^{J \cdot 2^{q-1}}$$

with J the bit reversed, natural, form of J_R .

This version of the fast transform has the same amount of computations, maintains "in place" computation, and allows parallel computation as do the other methods. It does require bit reversal of the exponent of $\bar{W}$.

An abortive FORTRAN procedure follows.

```

      DIMENSION X ( . . . )
      COMPLEX X
      INTEGER Q
      Do 1 Q = 1, M

      Do 1 JR = 0, 2M-Q - 1

      ** U =  $\bar{W}$  ** (J * 2Q-1)

      Do 1 K = 0, 2Q-1 - 1

      T = (X(I) - X(I + 2Q-1)) * U

      X(I) = X(I) + X(I + 2Q-1)

      X(I + 2Q-1) = T

      1 CONTINUE
  
```

** As for the time decimation, bit reversed transform method the sines and/or cosines may be stored or generated in bit reversed order allowing J rather than J_R to be written.

A flow chart of this form of the algorithm follows (Figure 14.2.4-1).

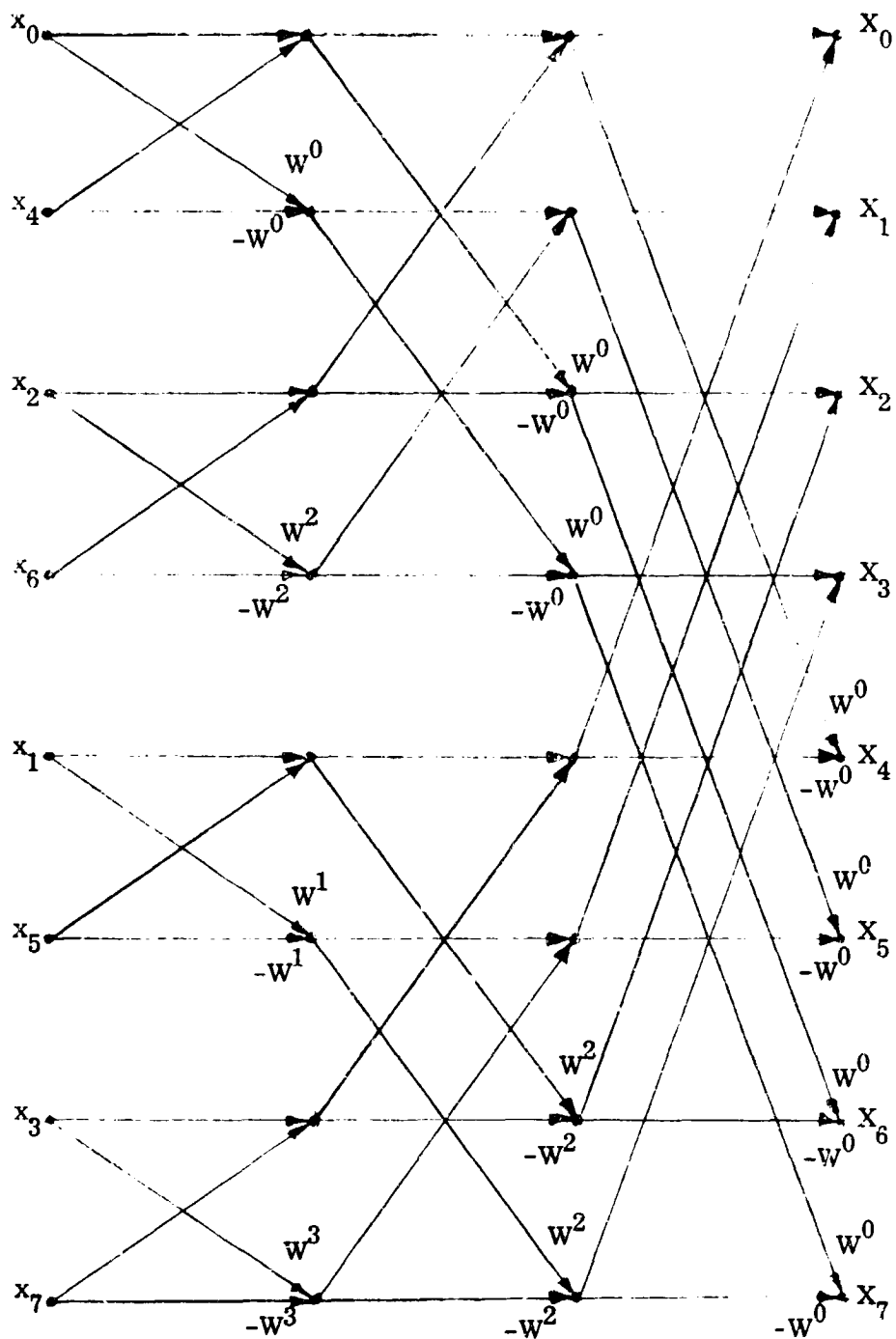


Figure 14.2.4-1 Flow Graph of Frequency Decimation Method
With a Bit Reversed Input Sequence

14.3 Multidimensional Discrete Fourier Transformation

Many physical problems require discrete Fourier transformation in more than one dimension. Two and three dimensional discrete Fourier transformations are defined similarly to one dimensional ones. By using a few simple formulae they can be reduced to one dimensional transforms, hence computation by the fast Fourier transformation is allowed.

The two dimensional transform is (in a form similar to the DFT1)

$$X_{n_1, n_2} = \sum_{j_1=0}^{N_1-1} \sum_{j_2=0}^{N_2-1} x_{j_1, j_2} w_1^{-j_1 n_1} w_2^{-j_2 n_2}$$

with

$$w_1 = e^{i \frac{2\pi}{N_1}}$$

and

$$w_2 = e^{i \frac{2\pi}{N_2}}$$

Inverse transformation is given by

$$x_{j_1, j_2} = \frac{1}{N_1 N_2} \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} X_{n_1, n_2} w_1^{j_1 n_1} w_2^{j_2 n_2} \quad (14.3-2)$$

To break the two dimensional transformation into one dimensional ones

$$A(n_1, j_2) = \sum_{j_1=0}^{N_1-1} x_{j_1, j_2} w_1^{-j_1 n_1} \quad (14.3-3)$$

is defined, which takes N_1 one dimensional transforms, and then

$$X_{n_1, n_2} = \sum_{j_2=0}^{N_2-1} A(n_1, j_2) W_2^{-j_2 n_2} \quad (14.3-4)$$

may be calculated with N_2 more one dimensional transforms. In all $N_1 + N_2$ one dimensional transforms are needed to compute a two dimensional one.

The three dimensional transform is

$$X_{n_1, n_2, n_3} = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \sum_{n_3=0}^{N_3-1} x_{j_1, j_2, j_3} W_1^{-j_1 n_1} W_2^{-j_2 n_2} W_3^{-j_3 n_3} \quad (14.3-5)$$

with an inverse

$$x_{j_1, j_2, j_3} = \frac{1}{N_1 N_2 N_3} \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \sum_{n_3=0}^{N_3-1} X_{n_1, n_2, n_3} W_1^{j_1 n_1} W_2^{j_2 n_2} W_3^{j_3 n_3} \quad (14.3-6)$$

To compute the transform define the two dimensional transform

$$A(n_1, n_2, j_3) = \sum_{j_1=0}^{N_1-1} \sum_{j_2=0}^{N_2-1} x_{j_1, j_2, j_3} W_1^{-j_1 n_1} W_2^{-j_2 n_2} \quad (14.3-7)$$

of which there are N_3 , each requiring $N_1 + N_2$ one dimensional transforms. The entire transform is given by

$$X_{n_1, n_2, n_3} = \sum_{j_3=0}^{N_3-1} A(n_1, n_2, j_3) W_3^{-j_3 n_3} \quad (14.3-8)$$

requiring $N_1 N_2$ one dimensional transforms. A total of $N_1 N_2 + N_2 N_3 + N_1 N_3$ one dimensional transforms are necessary.

15. APPLICATIONS OF DIGITAL FILTERS

Most of the text has been devoted to general properties and aspects of digital sequences and filtering. Applications of some of the material to a few standard problems and devices is useful both from a pedagogical viewpoint and for presentation of actual data processing techniques. Many of the techniques to be discussed are commonly used in analog systems but until recently have not been considered widely in the domain of digital processing.

15.1 TRANSLATION OF A SIGNAL SPECTRUM TO ANOTHER FREQUENCY: "MIXING" OR "MODULATION"

A problem of frequent interest is the translation of a given frequency spectrum to another center frequency. The result is eminently useful, as exemplified by broadcast systems, commercial spectrum analyzers, IF processing rather than RF processing, phase and frequency tracking, etc.

The basic method is rather simple. The original signal is simply multiplied with a sine wave at another frequency, say ω_0 . The resultant spectrum will appear at the sum and difference of the original spectrum center frequencies (plus possible spurious frequencies due to the round-off in digital systems, and non-linearities in analog systems). This can be seen by taking the transform of the given signal $x(t)$ (t may be continuous or discrete, $t = kT$, and Fourier or z transforms taken accordingly) multiplied by a sine wave which is the "mixer" (multiplier) output $y(t)$.

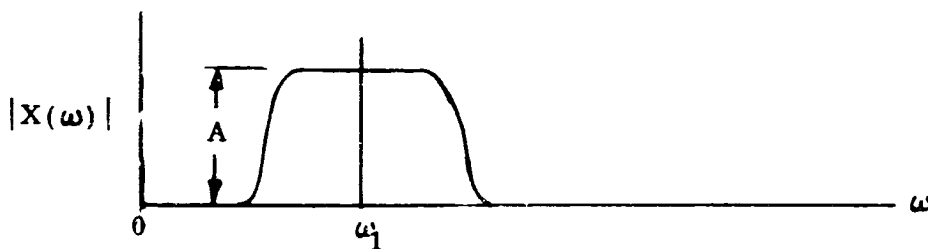
Now

$$\begin{aligned} y(t) &= x(t) \sin \omega_0 t \\ &= x(t) \frac{e^{i\omega_0 t} - e^{-i\omega_0 t}}{2i} \end{aligned} \tag{15.1-1}$$

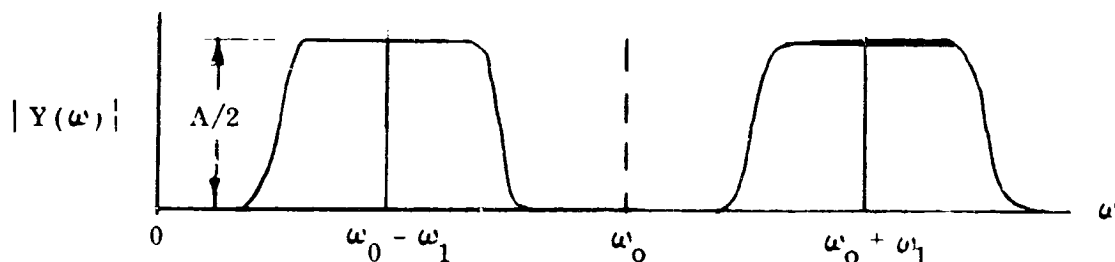
and its Fourier transform is

$$Y(\omega) = \frac{1}{2i} [X(\omega - \omega_0) - X(\omega + \omega_0)]$$

where $X(\omega)$ is the transform of $x(t)$. If $X(\omega)$ is localized and has its center frequency at ω_1 as in figure 15.1-1a then $Y(\omega)$ will have two localized components of center frequencies $\omega_1 + \omega_0$ and $\omega_1 - \omega_0$ as in figure 15.1-1b with half amplitude.



(a) Spectrum Prior to Mixing



(b) Spectrum After Mixing

Figure 15.1-1

A particular case of interest occurs when $X(\omega)$ is centered at $\omega = 0$, or D.C., then one term appears at ω_0 and the other at $-\omega_0$. If ω_0 is greater than the total bandwidth of $X(\omega)$, including negative frequencies, then the spectrum of $X(\omega)$ has been translated to ω_0 with no other terms present (barring round-off and non-linearities) for positive (physical) ω . Figure 15.1-2. This process is commonly known as "amplitude modulation" of a carrier of frequency, ω_0 , by a modulating signal $x(t)$.

If ω_0 is not greater than the total bandwidth, then the two terms overlap in the neighborhood of zero frequency.

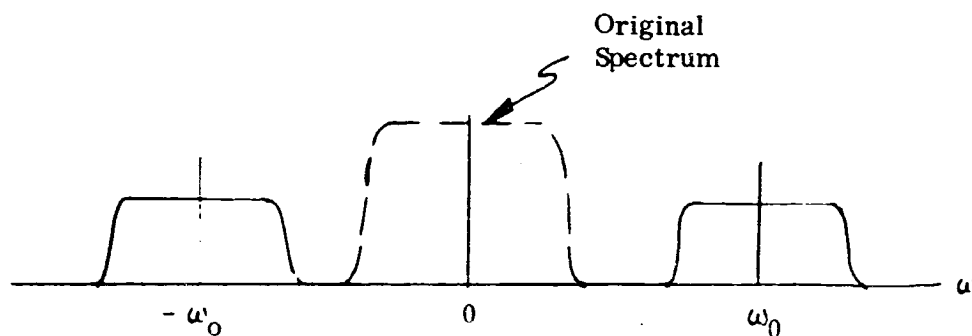


Figure 15.1-2. Translation of D.C. Spectrum to ω_0

While the previous paragraphs worked with transforms and spectrums to illustrate the mixing process for general signal spectrums, $x(t)$, a simple example demonstrates the process in the time domain. Consider the particular waveform $x(t) = \sin \omega_1 t$, which produces

$$\begin{aligned}
 y(t) &= \sin \omega_1 t \sin \omega_0 t \\
 &= \frac{A}{2} [\cos (\omega_1 - \omega_0) t - \cos (\omega_1 + \omega_0) t]
 \end{aligned}
 \tag{15.1-3}$$

As expected two terms, "sidebands", appear above and below the carrier frequency, ω_0 , one at the sum frequency, the other at the difference.

Symbolically, mixing (or multiplying) is denoted by a special symbol for used in block diagrams and flow graphs. Figure 15.1-3 illustrates the matter.

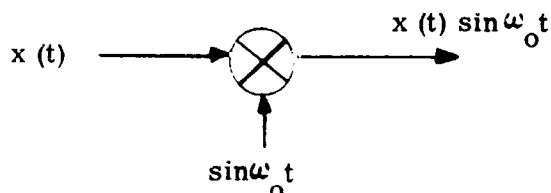


Figure 15.1-3. Mixing Block Diagram

15.1.1 Extraction of "Complex Modulation" by "Mixing"

The emminence of mixing techniques lies not only in "modulation" as discussed previously, but in "demodulation". Curiously, the same technique used to impose modulation on a carrier can be used to demodulate, or extract, the original information.

A modulated carrier is given by (Section 8.2)

$$s(t) = x(t) \cos \omega_0 t - y(t) \sin \omega_0 t \quad (15.1.1-1)$$

Obviously, the simple system of Figure 15.1.1-1 will allow demodulation.

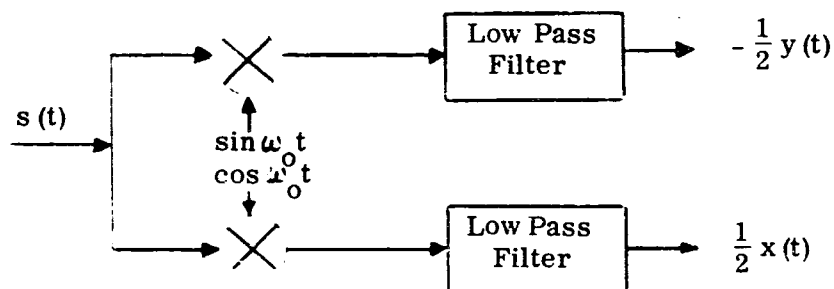


Figure 15.1.1-1 Demodulation by Mixing

The mathematics of the situation involves use of the simple identities

$$\sin^2 \omega_0 t = \frac{1}{2} - \frac{1}{2} \cos 2\omega_0 t$$

and

$$\cos^2 \omega_0 t = \frac{1}{2} + \frac{1}{2} \cos 2\omega_0 t$$

(15.1.1-2)

The two mixer outputs are, then,

$$s(t) \sin \omega_0 t = \frac{1}{2} x(t) \sin 2\omega_0 t - \frac{1}{2} y(t) + \frac{1}{2} y(t) \cos 2\omega_0 t.$$

and

(15.1.1-3)

$$s(t) \cos \omega_0 t = \frac{1}{2} x(t) + \frac{1}{2} x(t) \cos 2\omega_0 t - \frac{1}{2} y(t) \sin 2\omega_0 t.$$

If the spectra of $x(t)$ and $y(t)$ are such that they are approximately zero above $1/2 \omega_0$ then low-pass filtering at $(1/2)\omega_0$ will result in removal of the 2nd harmonic ($2\omega_0$) terms, albeit with delay depending on the sharpness of the filter cut-off needed. Neglecting the delay induced by the low-pass filtering the mixer outputs are, therefore,

$$s(t) \sin \omega_0 t = -\frac{1}{2} y(t)$$

and

(15.1.1-4)

$$s(t) \cos \omega_0 t = \frac{1}{2} x(t)$$

which are proportional to the complex modulation components. Assumed of course, in the previous analysis was the knowledge of the frequency ω_0 (and the phase) of the carrier.

In some applications the carrier frequency is not known, only the fact that it is fixed at some value. Assuming this value to be $\omega_0 = \omega_0 + \Delta$, then the mixer input is

$$\bar{s}(t) = \bar{x}(t) \cos \bar{\omega}_0 t - \bar{y}(t) \sin \bar{\omega}_0 t,$$

(15.1.1-5)

$$= x(t) \cos \omega_0 t - y(t) \sin \omega_0 t$$

where

$$x(t) = \cos \Delta t \bar{x}(t) - \sin \Delta t \bar{y}(t)$$

and

$$y(t) = \sin \Delta t \bar{x}(t) + \cos \Delta t \bar{y}(t)$$

Obviously, the effect of having an input to the mixers which differs from their reference frequency, ω_0 , by Δ is to have modulations $x(t)$ and $y(t)$ which are a linear combination of the original modulations. In fact $x(t)$ and $y(t)$ are simply complex modulated signals at a carrier of Δ

The linear transformation is an orthogonal one, preserving lengths, and having a rotation angle of Δt . Lengths are preserved in the sense that the modulation amplitude is unchanged, namely

$$\begin{aligned} A(t) &= \sqrt{\bar{x}(t)^2 + \bar{y}(t)^2} \\ &= \sqrt{x(t)^2 + y(t)^2}. \end{aligned} \quad (15.1.1-6)$$

If $\bar{x}(t)$ and/or $\bar{y}(t)$ is constant, then $x(t)$ and $y(t)$ are simply sine waves, with some phase angle, oscillating at a frequency Δ : $x(t)$ and $y(t)$ are 90° or $\frac{\pi}{2}$ radians different in phase.

15.2 DESIGN OF A "MOVING WINDOW" FILTER

In some applications a filter which moves its bandpass characteristic with time in a controllable manner is of use. That is, the center frequency of the filter is a known, specified function of time.

One approach, of course, is the building of a bank of stationary filters and switching their outputs as a function of time. Another approach is the adjusting of parameters of a single filter with time. Both of these methods can lead to extremely complex systems or systems with long computing times.

An alternate approach, commonly used in continuous systems, "mixes" the original signal with a carrier which is adjustable, then remodulates the carrier after a filtering operation. By appropriately utilizing "mixing" techniques a stationary filter can effectively be moved. Figure 15.2-1 illustrates the basic problem of a moving filter.

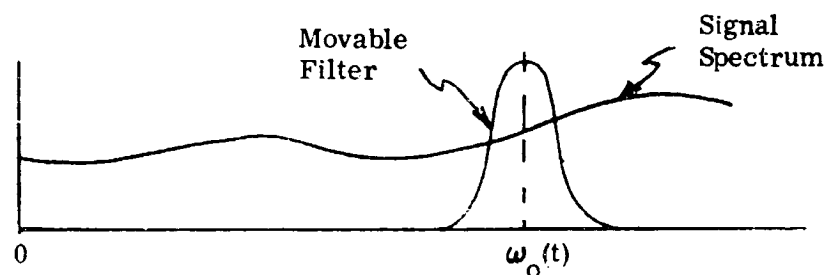


Figure 15.2-1 A "Moving" Filter

As the filter is moved with time the overall system is not really describable by a simple transfer function, but as long as the filter is moved slowly with respect to its bandwidth, then the system can be treated as such. Moving the filter "slowly" means that the distance the filter center frequency is moved, $\Delta \omega$, in time, Δt , must be such that

$$\frac{\Delta \omega}{\Delta t} < \frac{BW}{\tau} \quad (5.2-2)$$

where

BW = filter bandwidth in radians,

and

τ = filter impulse response time in seconds.

Now

$$BW \cdot \tau \geq 4,$$

or

$$\frac{1}{\tau} \leq BW/4$$

Therefore $\Delta \omega$ must be such that

$$\Delta \omega \ll (BW)^2 \Delta t / 4 \quad (15.2-3)$$

How much $\Delta \omega$ must be smaller than the right side of the inequality is dependent upon the application at hand.

A block diagram of the "moving" filter using "mixing" (Section 15.1) follows. Continuous or sampled signals may be used. Sampled signals have one slight peculiar limitation. This matter will be discussed after the basics of the 'moving' filter.

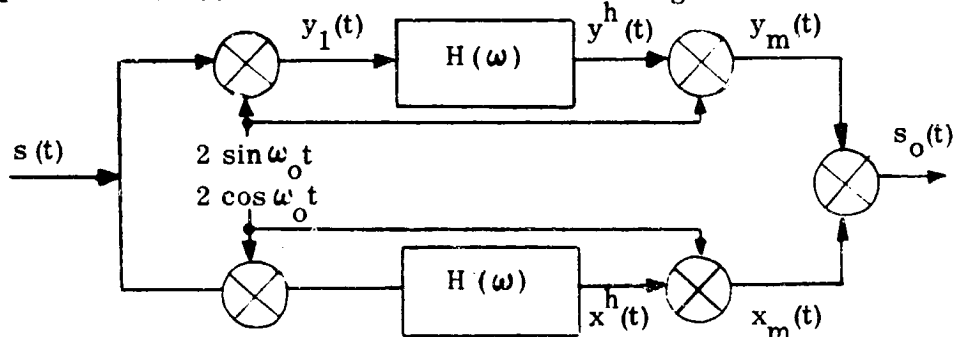


Figure 15.2-2 Moving Filter Implementation

Computation of the output $s_o(t)$ for a fixed mixing frequency, ω_0 , can be computed in either the time domain, as in the previous section, or through transform analysis. Transform analysis will be used for diversity.

Using the complex exponential representation of $\sin \omega_0 t$ and $\cos \omega_0 t$ the various system outputs in transform form, using capital letters to indicate the individual transforms, may be written as (where the subscripts and superscripts refer to Figure 15.2-1).

$$Y_1(\omega) = -\frac{i}{2} S(\omega - \omega_0) + \frac{i}{2} S(\omega + \omega_0),$$

$$X_1(\omega) = \frac{1}{2} S(\omega - \omega_0) + \frac{1}{2} S(\omega + \omega_0),$$

$$Y^h(\omega) = -\frac{i}{2} H(\omega) S(\omega - \omega_0) + \frac{i}{2} H(\omega) S(\omega + \omega_0),$$

$$X^h(\omega) = \frac{1}{2} H(\omega) S(\omega - \omega_0) + \frac{1}{2} H(\omega) S(\omega + \omega_0), \quad (15.2-4)$$

$$\begin{aligned} Y_m(\omega) &= -\frac{1}{4} H(\omega - \omega_0) S(\omega - 2\omega_0) + \frac{1}{4} H(\omega - \omega_0) S(\omega) \\ &= +\frac{1}{4} H(\omega + \omega_0) S(\omega) - \frac{1}{4} H(\omega + \omega_0) S(\omega + 2\omega_0), \end{aligned}$$

$$\begin{aligned} X_m(\omega) &= \frac{1}{4} H(\omega - \omega_0) S(\omega - 2\omega_0) + \frac{1}{4} H(\omega - \omega_0) S(\omega) \\ &+ \frac{1}{4} H(\omega + \omega_0) S(\omega) + \frac{1}{4} H(\omega + \omega_0) S(\omega + 2\omega_0), \end{aligned}$$

and finally,

$$\begin{aligned} S_0(\omega) &= Y_m(\omega) + X_m(\omega), \\ &= \frac{1}{2} \left[H(\omega - \omega_0) + H(\omega + \omega_0) \right] S(\omega) \end{aligned}$$

It can be seen that the equivalent filter, for a fixed ω_0 , is

$$H_{eq}(\omega) = \frac{1}{2} \left[H(\omega - \omega_0) + H(\omega + \omega_0) \right] \quad (15.2-5)$$

The impulse response of $H_{eq}(s)$ is real as

$$H_{eq}^*(-\omega) = H_{eq}(\omega) \quad (15.2-6)$$

In words, the sequence of operations is

1. Mix the input with the frequency ω_0 to obtain the translated spectrum.
2. Filter the desired output at (or about) zero frequency with a fixed parameter linear filter, $H(\omega)$.

3. Remodulate to the original frequency, ω_0 .
4. Combine the sine and cosine components via summation. The high frequency component is thereby cancelled.

The desired overall filter characteristic is completely determined by $H(\omega)$ translated to ω_0 . If $H(\omega)$ is a low-pass filter whose upper cut-off frequency is $\omega_B/2$, then $H_{eq}(\omega)$ is a bandpass filter of bandwidth ω_B .

15.2.1 Digital Considerations

Analogue and digital considerations are identical with one exception. Mixing and sampling cause aliasing, even if the system is completely digital, as mixing and sampling are commutative operations. Therefore as the mixing frequency approaches half the sampling frequency, the second harmonic is approaching zero frequency due to aliasing, allowing the second harmonic to appear within the bandpass of $H(\omega)$ producing distortion. If the bandwidth of $H(\omega)$ is considered to be $\omega_B/2$ then the mixing frequency should not be allowed to approach half the sampling frequency by at least the amount $(\omega_s/2) - \omega_B/2$, $\omega_s/2$ is equal to one-half the sampling frequency in radians per second. Actually the mixer frequency should not approach quite that close as practical filters do not fall-off to zero immediately outside of their bandwidth.

15.3 A DIGITAL OSCILLATOR

In a digital system it is often necessary to generate sampled sine and cosine waves of fixed or varying frequency. A pair of simple methods are available which may be implemented as linear recursive filters. Feedback techniques may be used with bandpass filters but analysis is quite complex. Only the sine and/or cosine of the basic increment between samples must be known. They can either be computed or stored. By changing these numbers the frequency of the oscillator(s) may be changed to any desired frequency without modifying the oscillator amplitude except for round-off errors

By using the trigonometric identities

$$\sin(A + B) = \sin A \cos B + \cos A \sin B$$

and

(15.3-1)

$$\cos(A + B) = \cos A \cos B - \sin A \sin B$$

it can be easily shown that

$$S_k = \cos \theta S_{k-1} + \sin \theta C_{k-1}$$

and

(15.3-2)

$$C_k = -\sin \theta S_{k-1} + \cos \theta C_{k-1}$$

where

$$S_k = \sin k\theta$$

$$C_k = \cos k\theta$$

The relations are obviously recursive and can be rewritten in the form

$$S_k - \cos \theta S_{k-1} = \sin \theta C_{k-1}$$

and

$$C_k - \cos \theta C_{k-1} = -\sin \theta S_{k-1}$$

(15.3-3)

Defining the transforms of S_k and C_k as $S(z)$ and $C(z)$, respectively; the z transform of (15.3-3) produces

$$\frac{S(z)}{C(z)} = \frac{\sin \theta z^{-1}}{1 - \cos \theta z^{-1}} \quad (15.3-4a)$$

$$= H_S(z)$$

and

$$\frac{C(z)}{S(z)} = \frac{-\sin \theta z^{-1}}{1 - \cos \theta z^{-1}} \quad (15.3-4b)$$

$$= H_C(z).$$

$H_S(z)$ and $H_C(z)$ can be implemented separately to compute S_k and C_k as only previous output values are needed.

In using these equations as oscillators the sampled values of

$$\sin \omega t$$

or

(15.3-5)

$$\cos \omega t$$

are desired at successive times $t = \kappa T$, where T is the sampling time and $\omega = 2\pi f$, the oscillator frequency is radians per second. Therefore, let

$$\theta = \omega T \quad (15.3-6)$$

in the equations.

It can be seen that $\cos \omega T$ and $\sin \omega T$ must be computed before the oscillator is initiated.

Equation (15.3-3) needs the pair of initial conditions

$$\begin{aligned} S_0 &= A \sin \theta_0 \\ C_0 &= A \cos \theta_0 \end{aligned} \quad (15.3-7)$$

Where θ_0 is any number and A is the amplitude desired. A simple choice is $\theta_0 = 0$ or

$$\begin{aligned} S_0 &= 0 \\ C_0 &= 1 \end{aligned} \quad (15.3-8)$$

To demonstrate that the sine/cosine oscillator is stable in amplitude, neglecting round-off, as the frequency is changed (15.3-3) is written in matrix form

$$\begin{pmatrix} S_k \\ C_k \end{pmatrix} = U(\omega) \begin{pmatrix} S_{k-1} \\ C_{k-1} \end{pmatrix} \quad (15.3-9)$$

where

$$U(\omega) = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}, \quad \theta = \omega T$$

Now the determinant of $U(\omega)$, $|U(\omega)|$, is equal to 1 and $U(\omega)$ is also an orthogonal matrix. Such a matrix represents a pure rotation by the angle θ without changing length. Therefore, as S_{k-1} and C_{k-1} are such that

$$A = \sqrt{S_{k-1}^2 + C_{k-1}^2} \quad (15.3-10a)$$

then it is also true that

$$A = \sqrt{S_k^2 + C_k^2}$$

Amplitude is preserved!

The matrix $U(\omega)$ can be changed at any time $t = mT$ to a new frequency and the oscillator will then oscillate at this frequency and at the same amplitude. Only round-off in the development of $\sin \theta$ and $\cos \theta$ affects the amplitude.

By taking a second difference of equation (15.3-2) two other computational procedures are available:

$$S_k = 2 \cos \theta S_{k-1} - S_{k-2}$$

and

(15.3-11)

$$C_k = 2 \cos \theta C_{k-1} - C_{k-2}$$

These equations only depend on their own past values and not on each other as the original ones do. However, these equations are much more dependent on round-off error.

The individual sine/cosine generators, equation (15.3-11), each need a pair of initial conditions

$$S_0 = A \sin \theta_0$$

$$S_1 = A \sin (\theta + \theta_0)$$

and

(15.3-12)

$$C_0 = A \cos \theta$$

$$C_1 = A \cos (\theta + \theta_0)$$

Formulae (15.3-11) for an oscillator are amplitude sensitive to frequency changes. Along with round-off noise, this indicates that they should be avoided unless absolutely necessary. Amplitude sensitivity can be avoided if an extra initial value is used based on the previous value plus the new frequency.

15.4 DIGITAL COMPLEX DEMODULATE FM DETECTOR

As frequency modulation detectors working at the carrier frequency may be at a much too high a frequency for digital processing directly it is useful to be able to extract the frequency modulation from the complex modulation by "coherent" processing.

Given the frequency modulated signal in terms of its carrier with phase modulation

$$s(t) = A \sin (\omega_0 t + \varphi(t)) \quad (15.4-1)$$

the frequency modulation is defined by

$$f(t) = \frac{1}{2\pi} \frac{d\varphi}{dt} \quad (15.4-2)$$

To find a processor expand $s(t)$

$$s(t) = A \sin \omega_0 t \cos \varphi(t) + A \cos \omega_0 t \sin \varphi(t)$$

Now by mixing and low pass filtering the two modulation components may be extracted (Section 15.1.1).

Namely

$$\begin{aligned} v &= A \cos \varphi(t) \\ &= \mathcal{L}_p [s(t) \sin \omega_0 t] \end{aligned} \quad (15.4-4a)$$

and

$$\begin{aligned} u &= \frac{A}{2} \sin \varphi(t) \\ &= \mathcal{L}_p [s(t) \cos \omega_0 t] \end{aligned} \quad (15.4-4b)$$

where $\mathcal{L}_p$ indicates low-pass filtering of the mixing product.

To compute the frequency modulation $\dot{\varphi}(t)$ consider

$$\frac{u}{v} = \tan \varphi(t) \quad (15.4-5)$$

and take the time derivative of the ratio:

$$\frac{d}{dt} \left(\frac{u}{v} \right) = \frac{d}{dt} [\tan \varphi(t)]$$

or

$$\frac{v\dot{u} - u\dot{v}}{v^2} = \sec^2(\varphi(t)) \dot{\varphi}(t)$$

and finally

$$f(t) = \frac{1}{2\pi} \dot{\varphi}(t)$$

$$= \frac{2}{\pi A^2} (vu - uv)$$

While this expression may seem abstract it is capable of simple implementation by linear digital filters and multiplication as illustrated by figure (15.5-1)

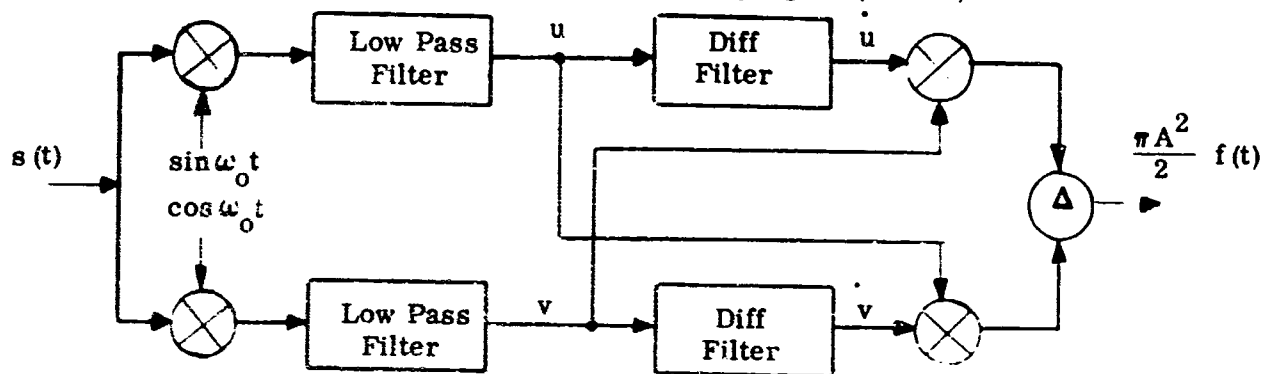


Figure 15.5-1 Complex Demodulate FM Detector

16. "PURE" DIFFERENTIATION AND INTEGRATION

16.1 DIFFERENTIATION

Discussion to this point has centered on integration of constant coefficient, integro-differential equations that describe "causal" physical systems, that is, no output can occur before an input is applied. This implies, as mentioned previously in Section 4.1, that the numerator of the transfer function $H(s)$ must have a degree less than or equal to that of the denominator.

This remark leads to difficulties in the realization of "simple" differentiation, $v(t) = dy/dt$, by a physical system of either the analogue or digital kinds without delay. A "pure" or "ideal" differentiator is not realizable due to the transfer function, $H_D(s) = s$, of numerator degree higher than that of the denominator. Higher orders of differentiation are correspondingly more difficult.

If, however, the quantity being differentiated is generated by a physical system having a transfer function $H(s)$ such that its denominator order is less than that of the numerator then the differentiation is realizable as the over-all transfer function $sH(s)$ describes a realizable system. Similarly, higher order differentiation, say n^{th} order, is realizable if the numerator of $H(s)$ is of such degree that $s^n H(s)$ is realizable.

16.1.1 Elementary Differentiation Considerations

The operation of analytic differentiation is expressed by the symbol

$$v(t) = \frac{dy(t)}{dt} \quad (16.1.1-1)$$

Taking the Laplace transform of $v(t)$, $V(s)$, and integrating by parts it is found that

$$V(s) = sY(s) - y(0+) \quad (16.1.1-2)$$

where

$Y(s)$ = the Laplace transform of $y(\cdot)$,

and

$y(0+)$ = the initial value of $y(t)$.

Now the operator s is not realizable by a physical system without delay: some future data must be known before $v(t)$ can be found. However a useful approximation consists in replacing (16.1.1-2) by the physically realizable equation

$$\frac{1}{\omega_0} \frac{dv(t)}{dt} + v(t) = \frac{dy(t)}{dt} \quad (16.1.1-3)$$

Letting ω_0 become large makes this equation approach the desired differentiation operation. Taking the Laplace transform of this modified equation:

$$V(s) = \frac{sY(s)}{1 + \frac{s}{\omega_0}} + \frac{1}{\omega_0} \frac{v(0+)}{(1 + \frac{s}{\omega_0})} - \frac{y(0+)}{1 + \frac{s}{\omega_0}} \quad (16.1.1-4)$$

Notice that this equation also approaches the Laplace transform of the unmodified equation for large ω_0 . $v(0+)$ is the initial value of the derivative which for ω_0 large contributes little to $V(s)$.

The modified transfer function is seen to be

$$H_D^M(s) = \frac{s}{1 + \frac{s}{\omega_0}} \quad (16.1.1-5)$$

The term

$$- \frac{y(0+)}{1 + \frac{s}{\omega_0}} \quad (16.1.1-6)$$

has an inverse transform

$$-y(0+) \omega_0 e^{-\omega_0 t} \quad (16.1.1-7)$$

which cancels the sharp output "spike" which occurs if the term is not present. For inputs which are "turned-on" at $t = 0$ in a discontinuous manner $y(0+)$ is equal to zero and a sharp peak occurs in the output. If the input is assumed to have existed prior to $t = 0$ and is not equal to zero then the "spike" cancelling term is necessary.

An example will clarify these remarks. Assume the following two inputs.

a) A simple linear function of time that has existed prior to $t = 0+$:

$$y_1(t) = \alpha t + \beta \quad (16.1.1-8)$$

b) The same function of time, except that it is "turned-on" at $t = 0$:

$$y_2(t) = \begin{cases} \alpha t + \beta, & t > 0, \\ 0, & t < 0 \end{cases} \quad (16.1.1-9)$$

The Laplace transform, of course, produces solutions only for $t \geq 0$, but the initial conditions appearing take account of conditions before $t = 0$.

Case a:

Taking the transform of $y_1(t)$ it is found to be

$$\begin{aligned} Y_1(s) &= \int_0^{\infty} (\alpha t + \beta) e^{-st} dt \\ &= \frac{\alpha}{s^2} + \frac{\beta}{s} \end{aligned} \quad (16.1.1-10)$$

Using the modified equation for differentiation (assuming ω_0 large so that $v(0+)/\omega_0$ approximates zero) and the initial condition $y(0+) = \beta$

$$\begin{aligned} V_1(s) &= \frac{\omega_0 s}{s + \omega_0} \left(\frac{\alpha}{s^2} + \frac{\beta}{s} \right) - \frac{\omega_0 \beta}{s + \omega_0} \\ &= \frac{\omega_0}{s + \omega_0} \frac{\alpha}{s} \end{aligned} \quad (16.1.1-11)$$

Taking the inverse transform

$$v_1(t) = \alpha(1 - e^{-\omega_0 t}), \quad t > 0; \quad (16.1.1-12)$$

which for large ω_0 gives

$$\begin{aligned} v_1(t) &= \frac{dy(t)}{dt}, \\ &= \alpha, \end{aligned} \quad (16.1.1-13)$$

as expected.

Case b:

The transform of $y_2(t)$ is also

$$Y_2(s) = \frac{\alpha}{s^2} + \frac{\beta}{s}, \quad (16.1.1-14)$$

but the initial condition is zero in this case. This input corresponds to that depicted in Figure 16.1.1-1.

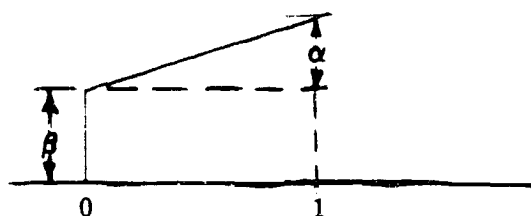


Figure 16.1.1-1. Linear Function "Turned-On"

The output transform of the modified equation is

$$V_2(s) = \frac{\omega_o}{s + \omega_o} \frac{\alpha}{s} + \frac{\omega_o \beta}{s + \omega_o} \quad (16.1.1-15)$$

with the inverse transform, the time function, given by

$$v_2(t) = \alpha(1 - e^{-\omega_o t}) + \omega_o \beta e^{-\omega_o t} \quad (16.1.1-16)$$

While the first term is identical to that in case (a) for ω_o large, the second term is a sharp "spike" that has a height of $\omega_o \beta$ and decays rapidly. This corresponds to approximately differentiating the fast rise of height β at $t = 0$. As ω_o approaches infinity the "spike" term approaches a Dirac delta function which is the derivative of the discontinuity at $t = 0$. Choice of ω_o is dictated by the range of frequencies over which the approximation is desired.

16.1.2 Numerical Considerations for Differentiation

The ideal differentiator has a frequency response given by

$$H_D(i\omega) = i\omega \quad (16.1.1-17)$$

If a non-recursive digital filter is necessary, due to the bandwidth of the sampled data approaching half the sampling frequency, then the methods of Section 5 must be used. A Fourier series expansion of $H_D(i\omega)$ is necessary, with convergence factors, and delay to the middle of the non-recursive filter data span by a "central" non-recursive filter. If $2N + 1$ weights are needed a time delay of NT seconds occurs. With larger N the approximation improves also. Of course, this method is not applicable to "real-time" filtering.

To take care of non-zero initial conditions if the discontinuity is undesirable the simple alternative is to subtract the initial conditions, from each input sample. That is, if the input sample is

$$y_j = y(jT) \quad (16.1.1-18)$$

the input to the differentiating filter should be

$$y_j' = y_j - y(0+) \quad (16.1.1-19)$$

which simply shifts the constant level, no longer allowing a discontinuity at $t = 0$.

If recursive numerical filters are allowable, the data has frequency content that does not extend to half the sampling frequency, then the data reconstruction methods of Section 7 can be used with less delay. Again the initial value of $y(0+)$ is best taken care of by subtracting it from each member of the input sequence as above

With both methods of Section 7 it is sufficient to use the unrealizable differentiator transfer function

$$H_D(s) = s \quad (16.1.1-20)$$

as both methods, filter and polynomial input reconstruction, produce effective transfer functions that are realizable.

Use of the modified transfer function (16.1.1-5)

$$H_D^M(s) = -\frac{s}{1 + \frac{s}{\omega_0}} \quad (16.1.1-5)$$

is effectively performed in the filter reconstruction method of obtaining sampled data filters where the factor $\omega_0/(s + \omega_0)$ is part of the reconstruction filter.

In the case where $y(t)$ is generated by a physical system $H(s)$, the methods of Section 7 may be applied for differentiation most accurately by considering the composite transfer function $sH(s)$ as in Figure 16.1.2-1.

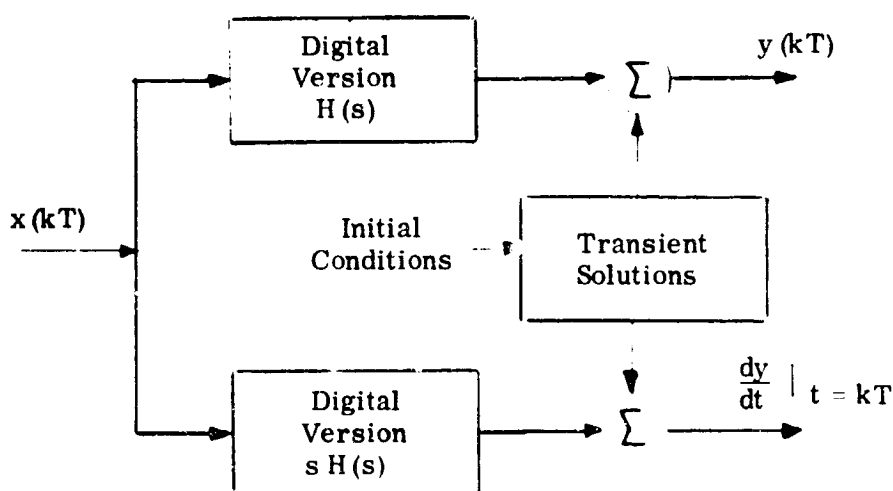


Figure 16.1.2-1. Differentiation of a Function Generated by a Filter $H(s)$ by a Parallel Method

It should be remembered that $y(kT)$ and $\left. \frac{dy}{dt} \right|_{t=kT}$ have delays specified by the digital filter synthesis method used. These delays are discussed in Sections 5 and 7 where synthesis methods are discussed. The "transient" solutions must also be delayed by this amount (see Section 6.4).

If simplicity of numerical processing, or a "black box" approach, is desired then an alternate procedure follows in Figure 16.1.2-2.

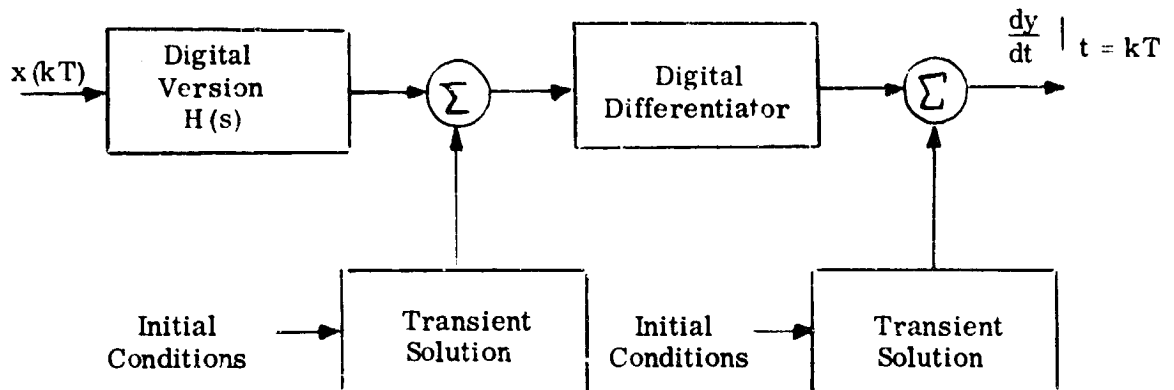


Figure 16.1.2-1. Differentiation by the "Black Box" Approach

16.2 INTEGRATION

The subject of integrating a constant coefficient linear differential equation has been discussed extensively. Here the subject is specialized to simple "definite" integration defined by

$$y(t) = \int_0^t x(\tau) d\tau \quad (16.2-1)$$

Upon taking the Laplace transfer

$$\frac{Y(s)}{X(s)} = \frac{1}{s} \quad (16.2-2)$$

Higher order integrations merely increase the inverse power of s by the number of integrations.

If $X(s)$ is generated by a physical system $H(s)$, the composite transfer function $H(s)/s$ is realizable and the techniques of Sections 5, 6, and 7 may be used directly with no further discussion.

However, if integration is desirable as a data processing function then certain other considerations are necessary. The operator $1/s$ is not bandlimited by any frequency and falls

off as slowly as possible with increasing frequency.

If the modified z transform of Section 6.1 is taken a digital transfer function corresponding to $1/s$ is

$$\begin{aligned} H_z(z) &= \frac{T}{1 - z^{-1}} - \frac{T}{2} \\ &= \frac{T}{2} \frac{1 + z^{-1}}{1 - z^{-1}} \end{aligned} \quad (16.2-3)$$

Strangely, the z transform, modified, produces the same result as the bi-linear z transform of Section 6.7. The difference equation corresponding to (16.2-3) is

$$y_k = y_{k-1} + \frac{T}{2} (x_k + x_{k-1}) \quad (16.2-4)$$

Examination of Figure 16.2-1 will show that this procedure is equivalent to approximation of the area under $x(t)$ by trapezoids, the well known "trapezoidal" rule.

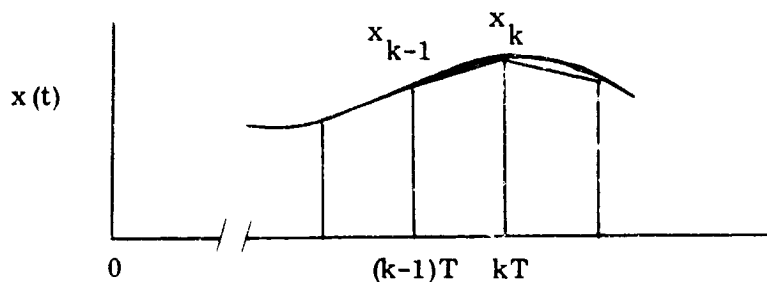


Figure 16.2-1. Trapezoidal Rule Integration

In terms of frequency

$$H_z(\omega) = \frac{1}{i} \frac{T}{2} \cot \frac{\omega T}{2} \quad (16.2-5)$$

which for low frequencies is approximately

$$H_z(\omega) \simeq \frac{1}{i\omega} \quad (16.2-6)$$

as it should be.

The most simple numerical integration scheme is rectangular approximation

$$y_k = y_{k-1} + T x_{k-1} \quad (16.2-7)$$

which has a z transform transfer function

$$H_z(z) = \frac{T}{1 - z^{-1}} \quad (16.2-8)$$

and a frequency response

$$H_z(\omega) = \frac{e^{i \frac{\omega T}{2}}}{i \frac{T}{2} \sin \frac{\omega T}{2}} \quad (16.2-9)$$

It also approximates the ideal integrator for low frequencies

$$H_z(\omega) \simeq \frac{1}{i \omega} \quad (16.2-10)$$

Another well known elementary integration procedure, Simpson's rule, uses exact fitting of a quadratic to the three most recent data points x_k , x_{k-1} , and x_{k-2} , with analytical integration of the expression from $(k-2)T \leq t \leq kT$, resulting in the recursive scheme

$$y_k = y_{k-2} + \frac{T}{3} (x_k + 4x_{k-1} + x_{k-2}) \quad (16.2-11)$$

The z transfer function is

$$H_z(z) = \frac{\frac{T}{3} (1 + 4z^{-1} + z^{-2})}{1 - z^{-2}} \quad (16.2-12)$$

with a frequency response

$$H_z(\omega) = \frac{\frac{T}{3} (2 + \cos \omega T)}{i \sin \omega T} \quad (16.2-13)$$

which also behaves as an ideal integrator for low frequencies

$$H_z(\omega) \simeq \frac{1}{i \omega} \quad (16.2-14)$$

By extending the curve fitting to a cubic and integrating over three intervals in the past Simpson's 3/8 rule is obtained:

$$y_k = y_{k-3} + \frac{3T}{8} (x_k + 3x_{k-1} + 3x_{k-2} + x_{k-3}) \quad (16.2-15)$$

The resultant z transfer function is

$$H_z(z) = \frac{\frac{3T}{8} (1 + 3z^{-1} + 3z^{-2} + z^{-3})}{1 - z^{-3}} \quad (16.2-16)$$

with the frequency response

$$H_z(\omega) = \frac{3T}{i8} \frac{\cos\left(\frac{3\omega T}{2}\right) + 3\cos\left(\frac{\omega T}{2}\right)}{\sin\left(\frac{3\omega T}{2}\right)} \quad (16.2-17)$$

which also behaves as an ideal integrator for low frequencies

$$H_z(\omega) \approx \frac{1}{i\omega}$$

In order to compare the efficacy of the different schemes take the magnitude ratio, $R(\omega)$, of various frequency responses to the ideal as a measure of the error:

$$\text{Rectangular Rule: } R(\omega) = \frac{\omega}{\frac{2}{T} \sin \frac{\omega T}{2}} \quad (16.2-18a)$$

$$\text{Trapezoidal Rule: } R(\omega) = \frac{\omega T}{2} \cot \frac{\omega T}{2} \quad (16.2-18b)$$

$$\text{Simpson's Rule: } R(\omega) = \frac{\omega T}{6} \frac{2 + \cos \omega T}{\sin \omega T} \quad (16.2-18c)$$

$$\text{Simpson's } 3/8 \text{ Rule: } R(\omega) = \frac{3\omega T}{8} \frac{\cos\left(\frac{3\omega T}{2}\right) + 3\cos\left(\frac{\omega T}{2}\right)}{\sin\left(\frac{3\omega T}{2}\right)} \quad (16.2-18d)$$

For the various schemes the ratio $D(\omega)$ is plotted in decibels in Figure 16.2-2 for clarity. Furthermore the frequency scale has been normalized to unit sampling frequency: $\omega' = \omega T$. Negative decibels indicate attenuation of the various methods with respect to the ideal. Obviously, Simpson's 1/3 rule and the trapezoidal rule are preferable, however Simpson's 1/3 rule amplifies errors at high frequencies while the trapezoid rule attenuates them. If a nominal error can be tolerated through low frequencies, then the rectangular rule has the least average error in the entire frequency range.

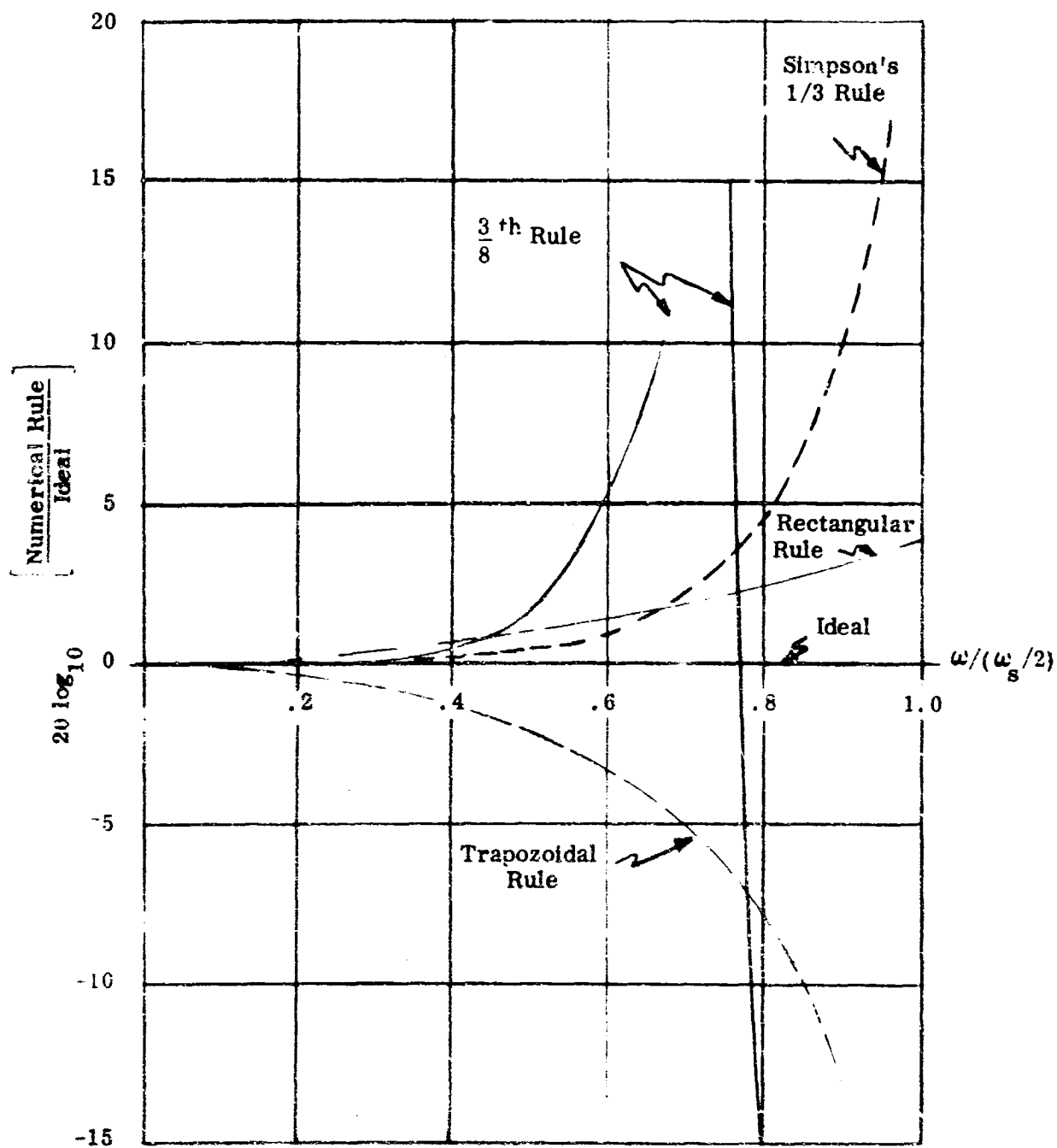


Figure 16.2-2. Plots of Integration Procedure Comparisons With Ideal

As details of the error ratios are sometimes useful at low frequency some values are given in Table 16.2-1.

$\frac{\omega}{\pi}$	Rectangular	Trapezoidal	Simpson's 1/3	Simpson's 3/8
0	0	0	0	0
.05	.0089	-.0179	.0000	.0000
.10	.0357	-.0719	.0005	.0011
.15	.0805	-.1629	.0024	.0057
.20	.1433	-.2925	.0079	.0186
.25	.2244	-.4633	.0198	.0482
.30	.3239	-.6785	.0424	.1077
.35	.4420	-.9426	.0818	.2189
.40	.5792	-1.2616	.1462	.4194
.45	.7358	-1.6433	.2471	.7765
.50	.9121	-2.0982	.4006	1.4236

Table 16.2-1. Comparison Ratio, $D(\omega)$,
for Various Integration Schemes

To this point it should be obvious that all these techniques are examples of the utility of polynomial extrapolation of Section 7.2. Only the rectangular method has a phase shift different from the ideal integrator and this phase shift is an advance in time by half a sampling time. Also all these procedures are unstable using the definition of stability given in Section 3.4.2. However, analog integration is also unstable in the same sense. Integration by its very nature accumulates the inputs so this should not be an excessive concern.

If phase shifts along with delay are allowable, the bandpass filter data reconstruction method of Section 7.4 may be used to obtain integration procedures over a wider bandwidth. Of course, the transfer function used along with this technique is $H(s) = 1/s$.

APPENDIX A CONVERSION OF A POLYNOMIAL TO AN EXPANSION IN TCHEBYCHEF POLYNOMIALS OF THE FIRST AND SECOND KINDS

Given the expression

$$\sum_{k=0}^N \alpha_k T_k(x) = \sum_{k=0}^N a_k x^k \quad (A-1)$$

The computation of the a_k 's in terms of the known α_k 's is desired. To solve, multiply both sides by $T_m(x)$.

Therefore, equation (A-1) becomes

$$\sum_{k=0}^N \alpha_k x^k T_m(x) = \sum_{k=0}^N a_k T_k(x) T_m(x) \quad (A-2)$$

For future benefit expand

$$x^k T_m(x) \quad (A-3)$$

in Tchebychev polynomials so that appropriate orthogonality relations may be used.

Using the identity

$$2x T_m(x) = T_{m+1}(x) + T_{m-1}(x) \quad (A-4)$$

it is found that

$$2^2 x^2 T_m(x) = T_{m+2}(x) + 2 T_m(x) + T_{m-2}(x) \quad (A-5)$$

and

$$2^3 x^3 T_m(x) = T_{m+3}(x) + 3 T_{m+1}(x) + 3 T_{m-1}(x) + T_{m-3}(x) \quad (A-6)$$

A general relationship can be recognized; namely

$$2^k x^k T_m(x) = \sum_{p=0}^k C(k, p) T_{m+k-2p} \quad (A-7)$$

where $C(k, p)$ is the binomial coefficient:

$$C(k, p) = \frac{k!}{p! (k-p)!} \quad (A-8)$$

This result can be checked by induction. The relation holds for fixed m and $k = 1, 2$, and 3 from equations A-4, 5, and 6. Assume it holds for k and try to deduce it for $k+1$. Then, by induction, the result holds for all k .

Multiplying (A-7) by x and using the identity (A-4) gives

$$2^{k+1} x^{k+1} T_m(x) = \sum_{p=0}^k C(k, p) T_{m+k+1-2p}(x) + T_{m+k-1-2p}(x) \quad (A-9)$$

Changing indices in the second term; namely

$$p = s-1 \text{ or } s = p+1 \quad (A-10)$$

and have:

$$\sum_{p=0}^k C(k, p) T_{m+k-1-2p}(x) = \sum_{s=1}^{k+1} C(k, s-1) T_{m+k+1-2s}(x) \quad (A-11)$$

As s is a dummy index insert this result in (A-9) and find that

$$\begin{aligned} 2^{k+1} x^{k+1} T_m(x) &= \sum_{p=1}^k [C(k, p) + C(k, p-1)] T_{m+k+1-2p}(x) \\ &\quad + C(k, 0) T_{m+k+1}(x) \\ &\quad + C(k, k) T_{m-(k+1)}(x) \end{aligned} \quad (A-12)$$

Now:

$$C(k, 0) = C(k, k) = C(k+1, 0) = C(k+1, k+1) = 1 \quad (A-13)$$

It can be shown, with a bit of algebra, that

$$C(k, p) + C(k, (p-1)) = C(k+1, p) \quad (A-14)$$

for $p = 1, 2, \dots, k$.

Therefore, the result: equation A-9 can be written as

$$2^{k+1} x^{k+1} T_m(x) = \sum_{p=0}^{k+1} C(k+1, p) T_{m+k+1-2p}(x) \quad (A-15)$$

which is identical with A-7 where k is simply replaced by $k+1$. Therefore the formula is the correct expansion in Tchebychev polynomials of $x^k T_m(x)$ for all k .

The problem then becomes a matter of using the orthogonality relation*

* Korn and Korn, Mathematical Handbook for Scientists and Engineers

$$\int_{-1}^1 \frac{T_m(x) T_n(x)}{\sqrt{1-x^2}} dx = \begin{cases} \frac{\pi}{2}, & m = n \neq 0, \\ \pi, & m = n = 0, \\ 0, & \text{otherwise} \end{cases} \quad (\text{A-16})$$

properly.

Defining

$$d_{mk} = \int_{-1}^1 \frac{x^k T_m(x)}{\sqrt{1-x^2}} dx \quad (\text{A-17})$$

then

$$a_0 = \sum_{k=0}^N \frac{d_{0k}}{\pi} \alpha_k, \quad (\text{A-18})$$

and

$$a_m = \sum_{k=0}^N \frac{2d_{mk}}{\pi} \alpha_k, \quad m > 0 \quad (\text{A-19})$$

Using the expansion (A-7) and the orthogonality relations:

$$\frac{d_{mk}}{\pi} = \frac{1}{2^k} \cdot \begin{cases} 0, & m > k \\ 0, & m+k \text{ odd} \\ 1, & m=k \\ C(k, (m+k)/2), & \text{otherwise.} \end{cases} \quad (\text{A-20})$$

Inserting this result in the expression for the a_m 's gives

$$a_0 = \sum_{k=0}^{\left\{ \frac{N+1}{2} \right\}} \frac{1}{2^{2k}} C(2k, k) \alpha_{2k}, \quad (\text{A-21})$$

m even:

$$a_m = \sum_{k=\frac{m}{2}}^{\left\{ \frac{N+1}{2} \right\}} \frac{2}{2^{2k}} C(2k, (m+2k)/2) \alpha_{2k} \quad (\text{A-22})$$

m odd:

$$a_m = \sum_{k=\frac{m+1}{2}}^{\left\{ \frac{N+1}{2} \right\}} \frac{2}{2^{2k-1}} C(2k-1, (m+2k-1)/2) a_{2k-1} \quad (A-23)$$

By $\left\{ \frac{N+1}{2} \right\}$ the truncated integer is meant.

As an example the 4th order Tchebychev polynomial is used. Namely; set

$$\begin{aligned} \sum_{k=0}^4 \alpha_k x^k &= 1 - 8x^2 + 8x^4 \\ &= T_4(x) \\ &= \sum_{k=0}^4 a_k T_k(x). \end{aligned} \quad (A-24)$$

See that

$$\alpha_1 = \alpha_3 = 0 \quad (A-25)$$

$$\alpha_0 = 1, \alpha_2 = -8, \text{ and } \alpha_4 = 8.$$

Using the preceding procedures the expected result should be $a_0 = a_1 = a_2 = a_3 = 0, a_4 = 1$.

By direct computation: find

$$\begin{aligned} a_0 &= C(0, 0) \alpha_0 + \frac{1}{2^2} C(2, 1) \alpha_2 + \frac{1}{2^4} C(4, 2) \alpha_4 \\ &= 1 + \frac{2}{2^2} x (-8) + \frac{6}{2^4} x^2 \cdot 8 \\ &= 0 \end{aligned} \quad (A-26)$$

All odd coefficients are automatically zero as equation (A-23) involves only odd α_k which are zero.

Similarly

$$a_2 = \frac{2}{2^2} C(2, 2) \alpha_2 + \frac{2}{2^4} C(4, 3) \alpha_4 \quad (\text{A-27})$$

$$= \frac{2}{2^2} \times (-8) + \frac{2}{2^4} \times 4 \times 8$$

or $a_2 = -4 + 4$

$$a_2 = 0$$

Finally, compute

$$a_4 = \frac{2}{2^4} C(4, 4) \alpha_4 \quad (\text{A-28})$$

or

$$a_4 = 1,$$

and have

$$1 - 8x^2 + 8x^4 = T_4(x) \quad (\text{A-29})$$

as expected.

Trying the relations for an odd case, namely

$$5x - 20x^3 + 16x^5 = T_5(x) \quad (\text{A-30})$$

produces

$$\alpha_0 = \alpha_2 = \alpha_4 = 0 \quad (\text{A-31})$$

$$\alpha_1 = 5, \alpha_3 = -20, \text{ and } \alpha_5 = 16.$$

The expected result is finding $a_0 = a_1 = a_2 = a_3 = a_4 = 0$ and $a_5 = 1$. Of course, from equation A-22 the even a_m 's are zero.

Compute:

$$\begin{aligned}
 a_1 &= \frac{2}{2} C(1, 1) \alpha_1 + \frac{2}{2^3} C(3, 2) \alpha_3 + \frac{2}{2^5} C(5, 3) \alpha_5 \\
 &= \alpha_1 + \frac{2 \times 3}{2^3} \alpha_3 + \frac{2}{2^5} \times 10 \alpha_5 \\
 &= 5 - \frac{6}{8} \times 20 + \frac{2}{2^5} \times 10 \times 16
 \end{aligned} \tag{A-32}$$

or

$$a_1 = 0;$$

and

$$\begin{aligned}
 a_3 &= \frac{2}{2^3} C(3, 3) \alpha_3 + \frac{2}{2^5} C(5, 4) \alpha_5 \\
 &= \frac{2}{2^3} \times (-20) + \frac{2}{2^5} \times 5 \times (16)
 \end{aligned} \tag{A-33}$$

or $a_3 = 0;$

and

$$\begin{aligned}
 a_5 &= \frac{2}{2^5} C(5, 5) \alpha_5 \\
 &= \frac{2 \times 16}{2^{-5}} = 1.
 \end{aligned} \tag{A-34}$$

Therefore

$$5x - 20x^3 + 16x^5 = T_5(x) \tag{A-35}$$

as expected.

The other important computation is finding, for the imaginary part, the b_k 's in terms of the β_k 's from the relation

$$\sum_{k=0}^{N-1} \beta_k x^k = \sum_{k=1}^N b_k U_k(x) \tag{A-36}$$

As before (equation A-2) multiply both sides by $\sqrt{1-x^2} U_m(x)$ and express $x^k U_m(x)$ in terms of the $U_n(x)$. The $U_n(x)$ satisfy the same recursion relation as did the $T_n(x)$, therefore, the relationship

$$2^k x^k U_m(x) = \sum_{p=0}^k C(k, p) U_{k+m-2p}(x) \quad (A-37)$$

can automatically be written as the derivation is identical to that for equation (A-7).

Next, using the orthogonality relation for the $U_n(x)$ *; namely,

$$\int_{-1}^1 \sqrt{1-x^2} U_m(x) U_n(x) dx = \begin{cases} \frac{\pi}{2}, & m = n \\ 0, & \text{otherwise;} \end{cases} \quad (A-38)$$

$m, n = 1, 2, \dots$;

and obtain:

$$\frac{\pi}{2} b_m = \sum_{k=0}^{N-1} \beta_k \sum_{p=0}^k \frac{1}{2^k} C(k, p) \int_{-1}^1 \sqrt{1-x^2} U_{k+m-2p}(x) dx; \quad (A-39)$$

$m = 1, 2, \dots, N$.

Defining the quantities

$$f_{mk} = \sum_{p=0}^k \frac{1}{2^k} C(k, p) \int_{-1}^1 \sqrt{1-x^2} U_{k+m-2p}(x) dx, \quad (A-40)$$

they can be evaluated using the orthogonality relation (A-38):

$$\frac{2}{\pi} f_{mk} = \frac{1}{2^k} \cdot \begin{cases} 0, & k < m - 1 \\ 0, & k+m \text{ even} \\ 1, & k = m-1 \\ \frac{2m}{k+m+1} C(k, (k+m+1)/2), & m+1 \leq k \leq N-1 \end{cases} \quad (A-41)$$

* IBID

Therefore, the b_m 's are given by

$$b_m = \sum_{k=0}^{N-1} \frac{2}{\pi} f_{mk} \beta_k \quad (A-42)$$

For even m :

$$b_m = \sum_{k=\frac{m}{2}}^{\frac{N+1}{2}} \frac{2m}{(2k+m)} \frac{C(2k-1, (2(k-1) + m)/2)}{2^{2k-1}} \beta_{2k-1} \quad (A-43)$$

For odd m :

$$b_m = \sum_{k=\frac{m-1}{2}}^{\frac{N}{2}} \frac{2m}{2k+m+1} \frac{C(2k, (2k+m-1)/2)}{2^{2k}} \beta_{2k} \quad (A-44)$$

APPENDIX B. USEFUL TRIGONOMETRIC IDENTITIES

Trigonometric identities of use are derived in this appendix. The results that are found follow in summary:

$$\sum_{k=0}^N \cos k\theta = \frac{\sin (N+1) \frac{\theta}{2} \cos N \frac{\theta}{2}}{\sin \frac{\theta}{2}}, \quad (\text{B-1})$$

$$\sum_{k=1}^N \sin k\theta = \frac{\sin (N+1) \frac{\theta}{2} \sin N \frac{\theta}{2}}{\sin \frac{\theta}{2}}, \quad (\text{B-2})$$

$$\sum_{k=1}^N \cos k\theta = \frac{\sin N \frac{\theta}{2} \cos (N+1) \frac{\theta}{2}}{\sin \frac{\theta}{2}}, \quad (\text{B-3a})$$

$$= \frac{\sin (N+\frac{1}{2}) \theta}{2 \sin \frac{\theta}{2}}; \quad (\text{B-3b})$$

and

$$\sum_{k=1}^N \frac{\sin k\theta}{k} = \int_0^{\theta} \left[\sum_{k=1}^N \cos k\theta' \right] d\theta' \quad (\text{B-4a})$$

$$= \int_0^{\theta} \frac{\sin (N+\frac{1}{2})\theta'}{2 \sin \frac{\theta'}{2}} d\theta'. \quad (\text{B-4b})$$

Proving B-1: represent $\cos k\theta$ by its exponential form and sum the resultant geometric series:

$$\begin{aligned} \sum_{k=0}^N \cos k\theta &= \frac{1}{2} \sum_{k=0}^N (e^{ik\theta} + e^{-ik\theta}) \quad (\text{B-5}) \\ &= \frac{1}{2} \left[\frac{1 - e^{i(N+1)\theta}}{1 - e^{i\theta}} + \frac{1 - e^{-i(N+1)\theta}}{1 - e^{-i\theta}} \right] \\ &= \frac{1}{2} (e^{iN \frac{\theta}{2}} + e^{-iN \frac{\theta}{2}}) \frac{\sin (N+1) \frac{\theta}{2}}{\sin \frac{\theta}{2}} \end{aligned}$$

or

$$\sum_{k=0}^N \cos k\theta = \frac{\sin(N+1)\frac{\theta}{2} \cos N\frac{\theta}{2}}{\sin \frac{\theta}{2}} \quad (\text{B-6})$$

The calculation of the sum of sine harmonics is performed in an identical manner.

The summation

$$\sum_{k=1}^N \cos k\theta \quad (\text{B-7})$$

is obviously given from the above by

$$\sum_{k=1}^N \cos k\theta = \frac{\sin(N+1)\frac{\theta}{2} \cos N\frac{\theta}{2}}{\sin \frac{\theta}{2}} - 1 \quad (\text{B-8})$$

which can be converted to another form by trigonometric identities:

$$\begin{aligned} \frac{\sin(N+1)\frac{\theta}{2} \cos N\frac{\theta}{2}}{\sin \frac{\theta}{2}} - 1 &= \frac{\sin(N+1)\frac{\theta}{2} \cos N\frac{\theta}{2} - \sin \frac{\theta}{2}}{\sin \frac{\theta}{2}} \quad (\text{B-9}) \\ &= \frac{\sin N\frac{\theta}{2} \cos \frac{\theta}{2} \cos N\frac{\theta}{2} + \sin \frac{\theta}{2} \cos^2 N\frac{\theta}{2} - \sin \frac{\theta}{2}}{\sin \frac{\theta}{2}} \\ &= \frac{\sin \frac{N\theta}{2} (\cos \frac{N\theta}{2} \cos \frac{\theta}{2} - \sin \frac{\theta}{2} \sin \frac{N\theta}{2})}{\sin \frac{\theta}{2}} \end{aligned}$$

or

$$\sum_{k=1}^N \cos k\theta = \frac{\sin \frac{N\theta}{2} \cos(N+1)\frac{\theta}{2}}{\sin \frac{\theta}{2}} \quad (\text{B-10})$$

Also, as

$$\sin(N+1)\frac{\theta}{2} \cos N\frac{\theta}{2} = \frac{1}{2} (\sin(N+\frac{1}{2})\theta + \sin \theta), \quad (\text{B-11})$$

$$\sum_{k=1}^N \cos k\theta = \frac{\sin(N+\frac{1}{2})\theta}{2 \sin \frac{\theta}{2}} \quad (\text{B-12})$$

is obtained.

BIBLIOGRAPHY

- C. Lanczos, "Applied Analysis", Prentice Hall, Inc., 1964
- R. W. Hamming, "Numerical Methods for Scientists and Engineers", McGraw-Hill, 1962
- U. Grenander and M. Rosenblatt, "Statistical Analysis of Stationary Time Series", John Wiley and Sons, Inc., 1957
- J. R. Ragazzini and G. F. Franklin, "Sampled-Data Control Systems", McGraw-Hill, 1958
- R. B. Blackman and J. W. Tukey, "The Measurement of Power Spectra", Dover, 1959
- J. S. Bendat, "Principles and Applications of Random Noise Theory", John Wiley and Sons, Inc., 1958
- A. Papoulis, "The Fourier Integral and its Applications", McGraw-Hill, 1962
- J. F. A. Ormsby, "Design of Numerical Filters with Applications to Missile Data Processing", J. ACM, vol. 8, pp 440-466, July 1961
- R. M. Golden and J. F. Kaiser, "Design of Wideband Sampled-Data Filters", Bell System Tech. J., vol. 48, July 1964
- R. B. Blackman, "Linear Data-Smoothing and Prediction in Theory and Practice", Addison-Wesley, 1965
- J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series", Math. of Comput., vol. 19, pp 279-301, April 1965
- G-AE Subcommittee on Measurement Concepts, "What is the Fast Fourier Transform?", IEEE Trans. on Audio and Electroacoustics, vol. AU-15, pp 45-55, June 1967
- H. D. Helms, "Fast Fourier Transform Method of Computing Difference Equations and Simulating Filters", IEEE Trans. on Audio and Electroacoustics, vol. AU-15, pp 85-90, June 1967
- J. W. Cooley, P. A. W. Lewis, and P. D. Welch, "Application of the Fast Fourier Transform to Computation of Fourier Integrals, Fourier Series and Convolution Integrals", IEEE Trans. on Audio and Electroacoustics, vol. AU-15, pp 79-84, June 1967
- J. W. Cooley, P. A. W. Lewis, and P. D. Welch, "Historical Notes on the Fast Fourier Transform", IEEE Trans. on Audio and Electroacoustics, vol. AU-15, pp 76-79, June 1967
- C. M. Rader and B. Gold, "Digital Filter Design Techniques in the Frequency Domain", Proc. IEEE, vol. 55, No. 2, pp 149-171, February 1967