THE Z-TRANSFORM AS A GENERAL TOOL IN APPROXIMATION

by

Freeman Luke Pendleton

A Thesis Submitted to the Faculty of the

DEPARTMENT OF ELECTRICAL ENGINEERING

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

1 9 6 0

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in their judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Freeman T Pendleton

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

John L Stewart

John L. Stewart
Professor of Electrical Engineering

9 May, 1960
Date

ABSTRACT

In this paper, several methods of approximation that employ the z-transform are reviewed. In addition, a Taylor approximation technique is introduced which maintains simultaneous control over phase and amplitude. This is accomplished by a method involving Padé approximations.

The method is a straightforward one in which a time function is sampled at constant intervals and a polynomial F(z) is formed from the resulting discrete values. F(z) is then rationalized into F'(z), from which a rational function F'(s) is formed.

The validity of this method depends upon the function being approximated in that the results of some approximations are very close to the exact, while others contain a great amount of error. However, this paper points out the fact that there is a considerable group of functions that occur in practical engineering problems for which this method of approximation gives very nearly exact results.

ACKNOWLEDGMENT

The author wishes to express his sincere appreciation to Dr. John L. Stewart for his interest and helpful suggestions in the preparation of this thesis.

TABLE OF CONTENTS

Chapter One

FUNDAMENTAL CONCEPTS OF THE Z-TRANSFORM IN APPROXIMATION

1.1 Introduction

Engineering like many other disciplines is one of approximation rather than of exactness. In solving a problem, the engineer will obtain an exact solution in only the most trivial of cases; a great deal of time and effort has thus been devoted to methods of approximation concerning a wide variety of engineering problems.

Perhaps a word should be said about the meaning of approximation in the engineering sense. In this paper, approximation or approximate solution will designate any method for obtaining a usable solution to a problem for a defined range of the variable. That is, the solution will generally be valid only to within a specified accuracy for a stated range of values.

The engineering problem being solved in each case controls the selection of the appropriate method of approximation. Probably the single most important specification is error tolerance. Reason alone insures that the least complex approximating function that meets the error requirements is the one to use.

Approximation is normally employed whenever a problem has no exact solution, when it has one that is extremely difficult to obtain, or when other considerations of convenience so dictate. Some examples of well known approximations used in electronics engineering are:

a. Expressions relating operating characteristics for electron tubes and transistors.

b. Interstage design using Chebyshev and Butterworth approximations.

c. Constant k and m derived filter design.

d. Transformer design and related analysis.

Before we leave the general definition and basis for approximation, it must be pointed out that in the definition there is no specification of the degree of likeness to the exact. It should not be concluded that to approximate something leaves a problem somewhat short of the desired. This is simply not true; if we are willing to pay the price in the way of time, effort, and cost of materials, most engineering approximations can be made to reach as close to the exact as desired.

1.2 Scope of This Investigation

As was mentioned in the introduction, approximation in engineering is a vast subject on which much has been written. No attempt will be made in this paper to explore the entire field of engineering approximation, rather the

intent here is to investigate only one method of approximation, namely the z-transform.

1.3 Primary Purpose of z-transform

Until recently, the z-transform has been used primarily as a mathematical tool for analyzing control systems in which the signal is discrete as opposed to continuous. An example is a system of control that contains a digital computer or any other digital operation. This, however, is not the only application to which the z-transform is well adapted; the z-transform is also important to the general field of approximation.

1.4 A Shortcoming of Frequency Domain Analysis

In the design of networks, the most popular techniques for synthesis are related to the frequency domain. A shortcoming common to most frequency domain design techniques is that simultaneous control of magnitude and phase response to a given steady state excitation is hard to achieve. As a result, the engineer may design for one and let the other go as best it will. This, of course, can only be done as long as the ignored factor meets specifications. In many practical problems, one or the other of magnitude or phase is of little concern, in which case the frequency domain method is satisfactory. In the event both phase and magnitude are of concern and both have rigid specifications or limits, the engineer may find it quite advantageous to study alternate methods which can better serve his purpose.

1.5 Analysis in the Time Domain

Designing in the time domain may allow the designer equal (but indirect) emphasis of both magnitude and phase of his system. Why then is design in the time domain not the more popular method? The answer to this is that no one method is without shortcomings. One such shortcoming of time domain design is the fact that graphical solution is almost always needed for some step in the procedure. This often adds work and tedium that the engineer wishes to avoid if possible. Also, many signals are described only in terms of amplitude spectra and hence time-domain methods are not applicable. The choice of which method to use will, as always, depend upon the individual problem.

1.6 The z-transform of a Sampled Time Function

By the very definition of a z-transform, we see that its use is adaptable in determining or approximating the transfer function that, upon excitation, yields a particular time response. Consider the relationship between a function in z and its corresponding time function. We make the following assumptions for an impulse at the origin:[1]

a. The width approaches zero and the height approaches infinity in the limit.

b. The area is unity.

---

[1] J. A. Aseltine, Transform Method in Linear System Analysis, McGraw-Hill Book Company, Inc., 1958, Chapter 3

c. The entire function is in positive time.

These assumptions are valid for the mathematical description even though to physically produce such a thing in the limit is quite impossible. However, a time function with large amplitude compared to its duration can be produced which is adequate for most experimental purposes.

The impulse at the origin is represented as $\delta(t)$, where

$$\delta(t) = \begin{cases} \infty & t = 0 \\ 0 & \text{everywhere else along the time axis} \end{cases} \tag{1.1}$$

The impulse can be moved away from the origin to any selected point by the following representation:

$$\delta(t - t_1) = \begin{cases} \infty & t = t_1 \\ 0 & \text{everywhere else} \end{cases} \tag{1.2}$$

To use a notation that will be more adaptable for our purposes, we assume the positive time axis to be indexed in equal increments of width T and that the impulse is located at any of these indexed points. With this assumption, an impulse function

$$\delta(t - nT) = \begin{cases} \infty & t = nT \\ 0 & \text{everywhere else} \end{cases} \tag{1.3}$$

has the Laplace Transform

$$\mathcal{L}\left[\delta(t - nT)\right] = e^{-nTs} = F(s) \tag{1.4}$$

If we wish to locate impulses at all positive integer values of n, the resulting Laplace transform is thus

$$F(s) = \sum_{n=0}^{\infty} e^{-nTs} \tag{1.5}$$

The area of the impulse does not necessarily have to be unity; a weighted impulse may be interpreted as one with area different from unity. Weighting an impulse can be accomplished with an appropriate constant as follows:

$$F(t) = K\delta(t - nT) \tag{1.6}$$

From the above interpretation, we shall consider the discrete values obtained in sampling a continuous time function at regular intervals as a series of delayed impulses being weighted by constants $K_1, K_2, \ldots K_n, \ldots$ K_n is the magnitude of the time function at the point of the n^{th} sample. Upon making the substitution

$$z = e^{Ts} \tag{1.7}$$

into a series of delayed weighted impulses and summing over all positive time, we obtain

$$F(z) = \sum_{n=0}^{\infty} f(NT) \quad z^{-n} \tag{1.8}$$

which is the desired relationship of the time function in z and is called the z-transform of a time function f(t).

1.7 Numerical Integration[2]

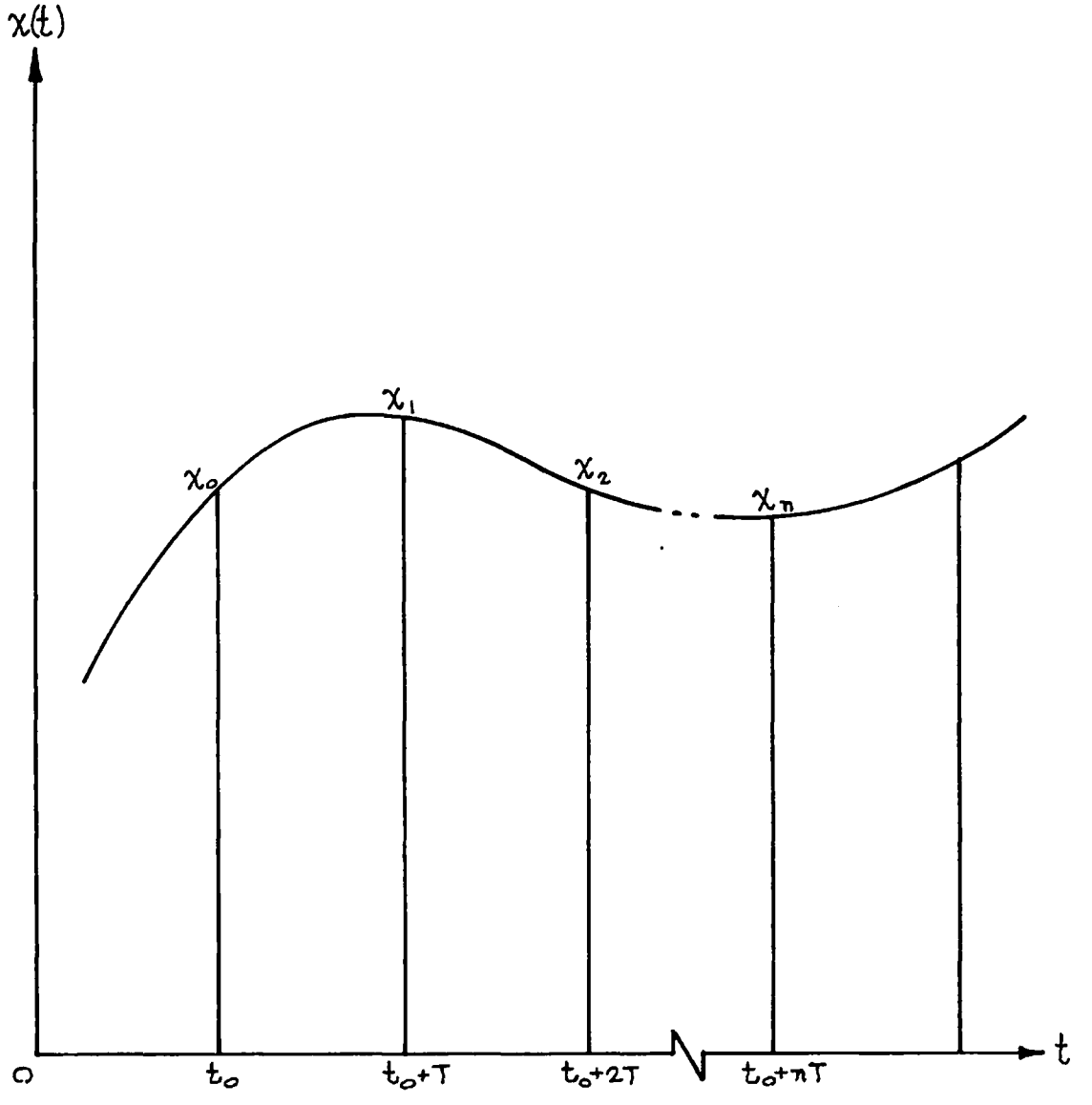
Let us consider some of the known methods of numerical analysis that are related to the z-transform. Evaluation of a definite integral is very often accomplished by numerical integration. The procedure followed in its basic form is a relatively simple one. In Figure (1.1), x(t) is some rational function of t, and x_o, x_1,... x_n are the magnitudes of x(t) values at the intervals t_o, t_1, t_2,...t_n respectively. An approximation to x(t) is made over a finite interval by finding a simpler function P(t) that has the correct values at the sampling times t_o, t_1, t_2,...t_n.

The function of P(t) can have a variety of forms, one which is of particular interest in this paper is a polynomial. When P(t) is a polynomial, the approximating method is known as "Polynomial interpolation", with special cases known as "Newton's", "Stirling's", and "Bessel's" formulas.

According to Newton's formula, x(t) may be represented by the polynomial P(t) in the interval t_o t t_o + nT such that

---

[2] J. T. Tou, Digital and Sampled-Data Control Systems, McGraw-Hill Book Company, Inc., New York, 1959, pp. 198-206



SAMPLES OF THE FUNCTION $x(t)$

FIGURE 1.1

$$P(t) = P(t_o + Tu) = f(u) = x_o + u\Delta x_o$$

$$+ \frac{u(u-1)\Delta^2 x_o}{2!} + \frac{u(u-1)(u-2)\Delta^3 x_o}{3!} + \ldots \qquad (1.9)$$

$$+ \frac{u(u-1)(u-2)\ldots(u-m-1)\Delta^m x_o}{m!} + \ldots$$

where $$u = \frac{t - t_o}{T} \quad \text{or} \quad t = t_o + Tu \qquad (1.10)$$

and $$\Delta x_o = x_1 - x_o \qquad (1.11)$$

$$\Delta^2 x_o = \Delta x_1 - \Delta x_o = x_2 - 2x_1 + x_o \qquad (1.12)$$

$$\Delta^3 x_o = \Delta^2 x_1 - \Delta^2 x_o = x_3 - 3x_2 + 3x_1 - x_o \qquad (1.13)$$

and x_o, x_1, $x_2, \ldots x_n$ are the values of $x(t)$ at the intervals t_o, $t_o + T$, $t_o + 2T, \ldots t_o + nT$. The expressions $\Delta^N x_o$ are differences relating the successive changes of a discrete function. They can be compared to a derivative of order N in a continuous function. That is, $\Delta^2 x_o$ is the second difference which in a continuous function is the second derivative $\frac{d^2x}{dt}$, etc. Equation (1.9) can in fact be recognized as an incremental form of the familiar Taylor series.

The integral $\int_{t_o}^{t + nT} x(t)dt$ can be approximated in the following way:

$$\int_{t_o}^{t_o + nT} x(t)dt \approx \int_{t_o}^{t_o + nT} P(t)dt$$

$$= T\int_0^n P(t_o + Tu)du = T\int_0^n f(u)du \qquad (1.14)$$

where in the above expression

$$dt = Tdu \tag{1.15}$$

and $f(u) = P(t_o + Tu)$ as defined in Eq. (1.9). Upon substituting Eq. (1.9) into Eq. (1.14) and integrating term by term we obtain

$$\int_{t_o}^{t_o + nT} x(t)dt \approx T \left[n\, x_o + \right.$$

$$\frac{n^2}{2}\Delta x_o + \left(\frac{n^3}{3} - \frac{n^2}{2}\right)\frac{\Delta^2 x_o}{2!} +$$

$$\left(\frac{n^4}{4} - n^3 + n^2\right)\frac{\Delta^3 x_o}{3!} + \tag{1.16}$$

$$\left(\frac{n^5}{5} - \frac{3n^4}{2} + \frac{11n^3}{3} - 3n^2\right)\frac{\Delta^4 x_o}{4!} +$$

$$\left(\frac{n^6}{6} - 2n^5 + \frac{35n^4}{4} - \frac{50n^3}{3} + \right.$$

$$\left. \left. 12n^2\right)\frac{\Delta^5 x_o}{5!} + \ldots \right]$$

where n is the number of incremental differences included in a particular approximation.

Eq. (1.16) provides a general formula for numerical integration from which various quadrature formulas may be derived by letting n = 1,2,3,... Examples of a few such formulas along with their z-transform application are next discussed.

a. Integration by the Rectangular Rule

The simplest quadrature formula based on the rectangular rule will be discussed first. It can be seen in Figure (1.2) that a continuous curve x(t) can be approximated by a staircase waveform with the area under the curve approximated by the sum of the rectangular areas thus formed. If in Figure (1.2) we let x_k be the value of x(t) at t = kT and y_{k-1} the area under the curve from t = 0 to t = (k-1)T, then the total area under the curve x(t) between t = 0 and t = kT is approximated by the relationship

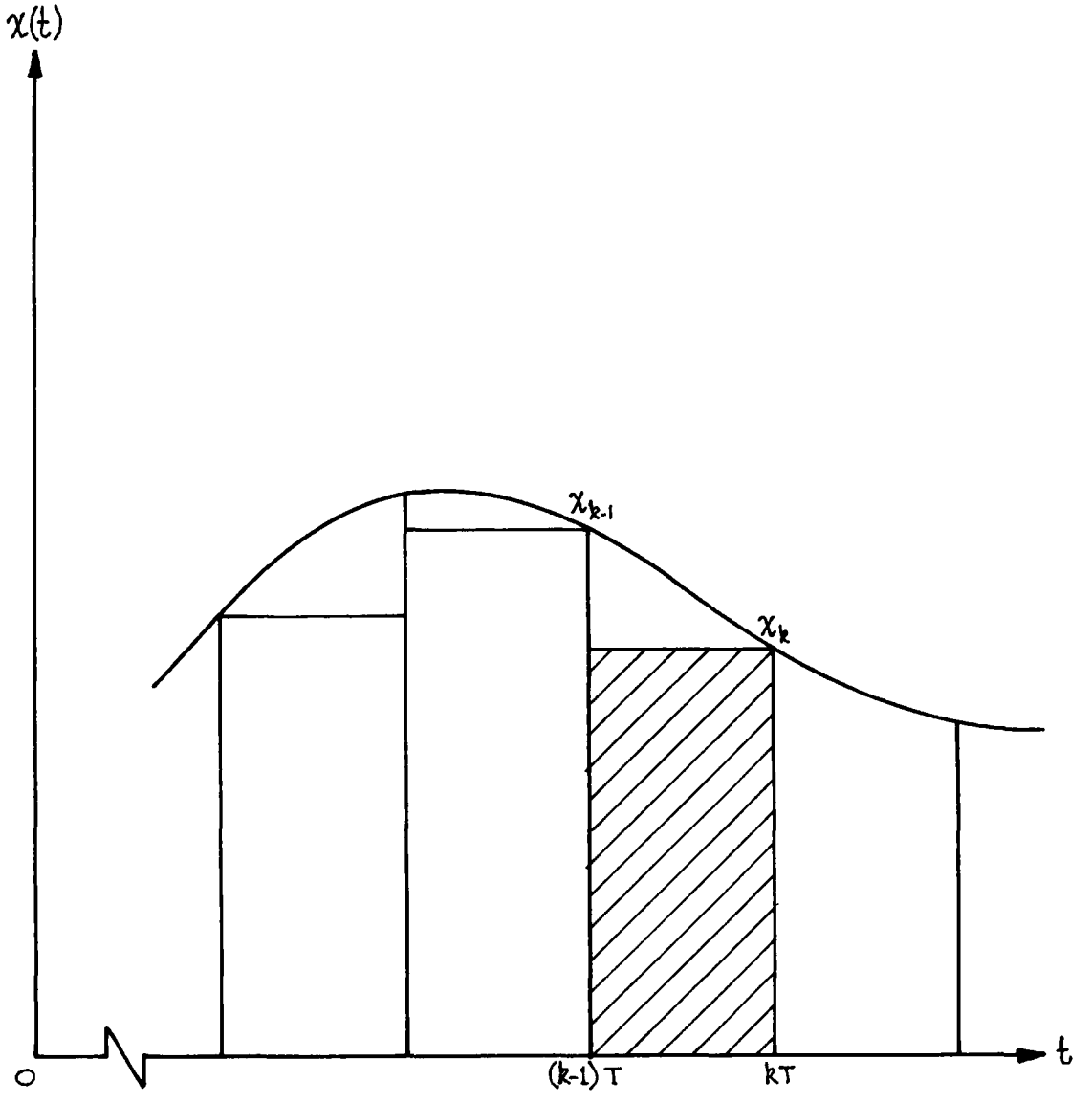
$$y_k = y_{k-1} + Tx_k \tag{1.17}$$

This in turn leads to the sampled function

$$y^*(t) = y^*(t-T) + Tx^*(t) \tag{1.18}$$

Recalling the relationship of the z-transform to the continuous time function, we see that the starred function means that the areas under a continuous curve are represented by weighted impulses. Therefore, in the above function, when t=kT, $y^*(t) = y_k$, $y^*(t-T) = y_{k-1}$, and $x^*(t) = x_k$.

The relationship of this quadrature formula and Newtons general interpolation formula as given in Eq. (1.16) exists when we consider n = 0. The rectangular formula cannot be obtained from Eq. (1.16) by setting n = 0, but rather the interpretation is that by using rectangles,



INTEGRATION BY THE RECTANGULAR RULE

FIGURE 1.2

there will be a 0th difference, hence the incremental area is Tx_k.

When we obtain the z-transform of both sides of Eq. (1.18), there results

$$Y(z) = z^{-1}Y(z) + TX(z) \tag{1.19}$$

rearranging, we have

$$D(z) = \frac{Y(z)}{X(z)} = \frac{T}{1-z^{-1}} = \frac{Tz}{z-1} \tag{1.20}$$

where D(z) is the pulse transfer function (z-transform) of the numerical integration operator. With the substitution $e^{Ts} = z$, we obtain the Laplace transform of the starred or sampled function as

$$D(s) = \frac{T}{1 - e^{-Ts}} \tag{1.21}$$

In order to ascertain how well this integrator performs, we will compare its response to that of the ideal integrator $\frac{1}{s}$. The comparison will be made in the frequency domain. We therefore replace s by jw in Eq. (1.21) to obtain the frequency characteristics. This results in

$$D(jw) = \frac{T}{1 - e^{-jwT}} = \frac{T}{2}\left(1-j\cot\frac{wT}{2}\right) \tag{1.22}$$

which yield the amplitude and phase characteristics as

$$\left|D(jw)\right| = \frac{T}{2}\left|\csc\frac{wT}{2}\right| \tag{1.23}$$

$$\phi(w) = -\frac{\pi}{2} + \frac{wT}{2} \tag{1.24}$$

Plots of Eq. (1.23) and Eq. (1.24) are given in Figure (1.3). It will be noted that for low values of w the amplitude approximation is good, however, the phase starts to depart from the ideal immediately as frequency starts to increase. We in fact, see the same difficulty that is often encountered with frequency domain methods, namely control is not exercised over both amplitude and phase.

b. Integration by the Trapezoidal Rule.

The area under a curve can be approximated by a series of rectangles. It can also be approximated by a series of trapezoids as shown in Figure (1.4). Again let y_{k-1} be the area under the curve from t = 0 to t = (k-1)T, and let the area from t = (k-1)T to t = kT be the k^{th} trapezoid. The total area from t = 0 to t = kT is then represented by

$$y_k = y_{k-1} + \frac{T}{2}(x_k + x_{k-1}) \qquad (1.25)$$

which leads to the quadrature formula based upon the trapezoidal rule as

$$y^*(t) = y^*(t-T) + \frac{T}{2}\left[x^*(t) + x^*(t-T)\right] \qquad (1.26)$$

where the starred notation is as before. Eq. (1.26) can be obtained directly from Eq. (1.16) by letting n = 1. However, because only two points define the area, we can only have the first difference, i.e., $y_1 - y_0$, which gives us only the first two terms of Eq. (1.16)

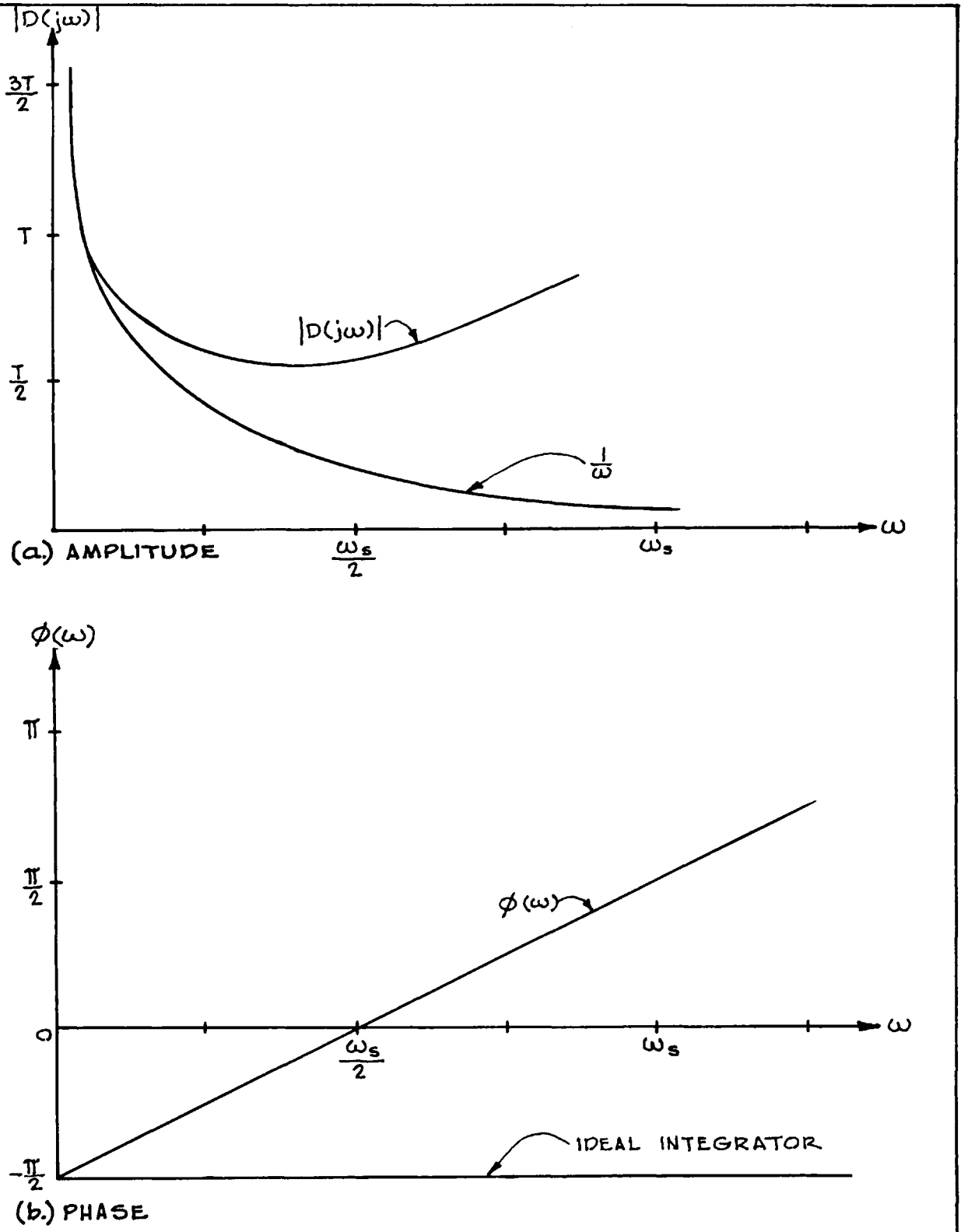
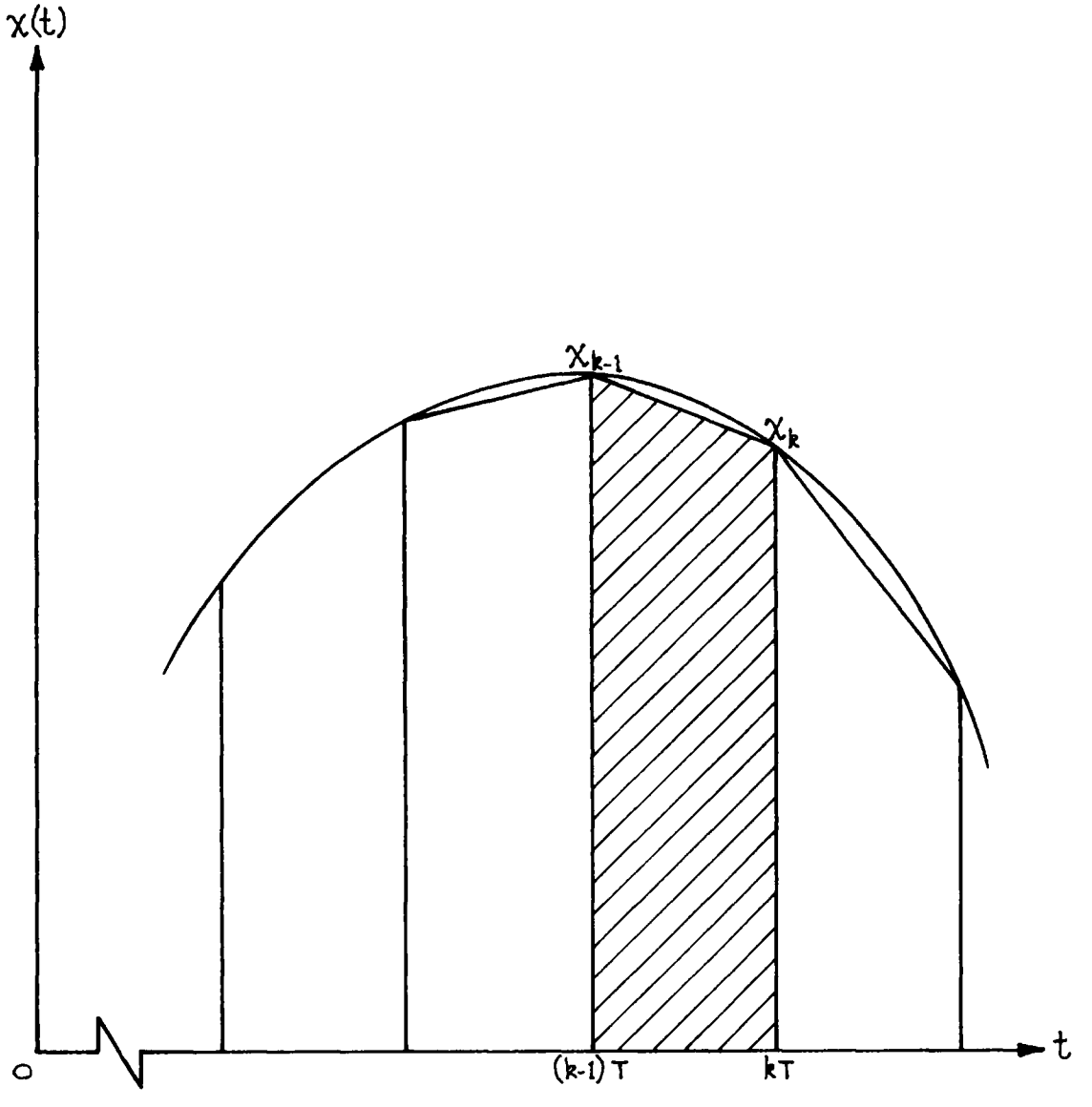


FIGURE 1.3 CHARACTERISTICS OF DIGITAL INTEGRATOR BASED UPON THE RECTANGULAR RULE



INTEGRATION BY THE TRAPEZOIDAL RULE

FIGURE 1.4

Again we take the z-transform of Eq. (1.26), rearrange terms and obtain the pulse transfer function

$$D(z) = \frac{T}{2} \frac{1 + z^{-1}}{1 - z^{-1}} \tag{1.27}$$

When the substitution $z = e^{Ts}$ is made, we have the integral operator

$$D(s) = \frac{T}{2} \frac{1 + e^{-Ts}}{1 - e^{-Ts}} \tag{1.28}$$

The substitution of jw for s yields

$$D(jw) = \frac{T}{2} \frac{1 + e^{-jwT}}{1 - e^{-jwT}} = \frac{-jT}{2} \cot \frac{wT}{2} \tag{1.29}$$

Eq. (1.29) yields amplitude and phase characteristics as

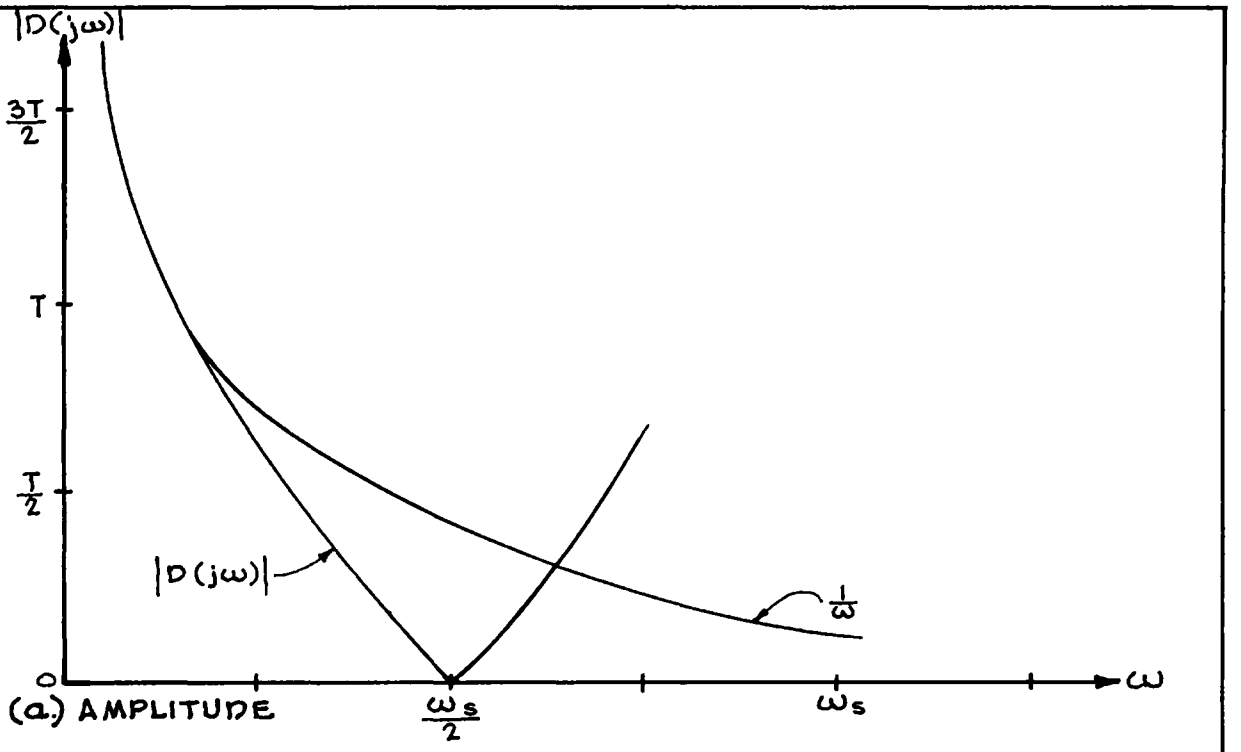
$$\left| D(jw) \right| = \frac{T}{2} \left| \cot \frac{wT}{2} \right| \tag{1.30}$$

$$\phi(w) = -\frac{\pi}{2} \tag{1.31}$$

Again, in order to compare this function with the ideal, Eq. (1.30) and Eq. (1.31) are plotted in Figure (1.5). It is seen that the amplitude characteristic has suffered some as compared with that pertaining to simple rectangular rule integration, but the phase is equal to that of an ideal integrator.

c. Integration by Simpson's One-Third Rule.

By letting n = 2 we obtain from Eq. (1.16) the quadrature formula based upon Simpson's One-Third Rule. In this example we know only the points x_0, x_1, x_2, which means we have only the first and second differences. We



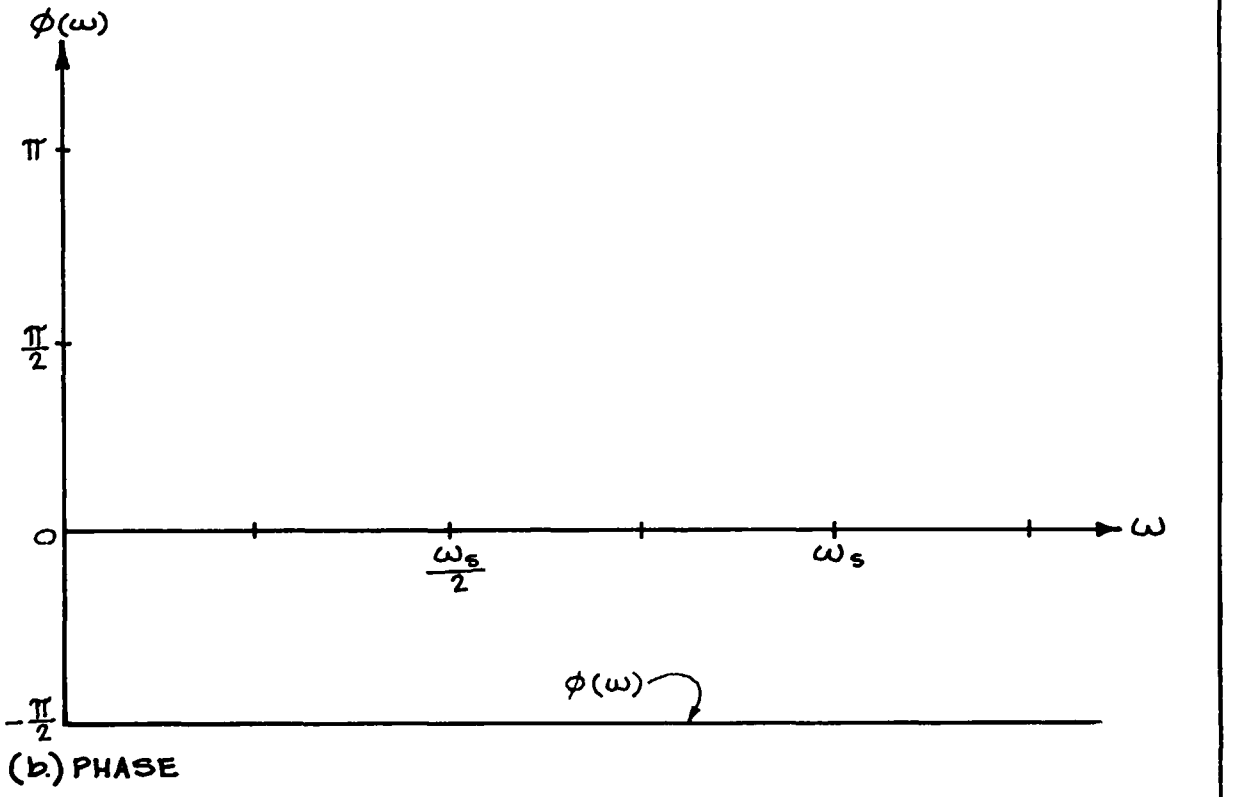


FIGURE 1.5 CHARACTERISTICS OF DIGITAL INTEGRATOR BASED UPON THE TRAPEZOIDAL RULE

formulate the expressions as above, take the z-transform, rearrange terms and obtain

$$D(z) = \frac{T}{3} \frac{1 + 4z^{-1} + z^{-2}}{1 - z^{-2}} \tag{1.32}$$

From Eq. (1.32) we form magnitude and phase characteristics as

$$\left|D(jw)\right| = \frac{T}{3} \left| \frac{2 + \cos wT}{\sin wT} \right| \tag{1.33}$$

$$\phi(w) = -\frac{\pi}{2} \tag{1.34}$$

In this approximation, there is an improvement in the magnitude characteristic while ideal phase behavior is maintained.

Other formulas can be obtained by letting n = 3 and 6 in Eq. (1.16) for quadrature formulas based on Simpson's Three-eights Rule and Weddle's Rule, respectively. Both of these formulas yield ideal phase characteristics with a slight improvement in magnitude over lower order n formulas.

In the above examples, we see the z-transform playing quite a different role than that of an operator in a discrete control system. As an operator in this type of approximation, it tells the designer a great deal more about his function than he would have otherwise obtained.

1.8 Solutions of Linear and Nonlinear Systems Using the z-transform.[3]

The solving of an integrodifferential equation by the use of transform operators such as the Laplace, Mellin, Hankle, Meijer, Legrendre systems has been a welcome relief generally to the engineer for quite some time. However, we find in all cases, the transforms have definite limitations with respect to the type of integrodifferential equations upon which they can operate.

There is no such thing as an unsolvable integrodifferential equation in the sense that not even an approximation can be made. This is because the engineer can always fall back on well known methods of numerical, graphical, or series approximation. However, for sake of generality of solution and ease of algebraic manipulation, it is desirable to solve the equation by purely analytical means whenever possible. To this end the z-transform can be added to the collection of transform operators which are adaptable to the solution of various types of integrodifferential equations.

An example of the above is the approximation derived by Boxer and Thaler which relates in a rational

---

[3] R. Boxer, S. Thaler, "A Simplified Method of Solving Linear and Non-linear Systems", Proc. IRE, Vol. 44, no. 1, pp. 89-101, 1956

equality the complex variable s to the complex variable z and is called the z-form. This approximation is based upon the series expansion of the natural log of the complex variable z in the defining relationship:

$$s = \frac{1}{T} \ln z \tag{1.35}$$

along with other valid assumptions. There have been other approximations that are quite similar to this one of Boxer and Thalers', but the procedure for using one system applies for the most part to all.

The great advantage of using the z-form or others of this type lies in the fact that long division can be used for the inversion from the complex operator domain, which of course is a tremendous aid for large order polynomials that would otherwise have to be factored for residue inversion. The z-transform is adaptable to the solution of any ordinary differential equation with constant coefficients, certain differential equations with time varying coefficients, and certain nonlinear differential equations. The procedure is a straightforward one and in many cases the response of a continuous control system can be obtained with much less time and effort than if the inversion with the continuous transform were made.

1.9 An Approximating Method That Controls Both Magnitude and Phase.

An important feat to accomplish in both analysis and synthesis is to determine the transfer function which, upon excitation by an appropriate time function, results in a desired time response. In the event that a transfer function which will give an exact response cannot be obtained, an approximation thereto is desired.

In continuing this investigation with the objective of finding a method of approximation that is reasonably general in nature, but still simultaneously controls the magnitude and phase of the system, we turn to a form of the Padé approximation.

1.10 Padé Approximations[4]

The Padé approximation deals directly with the operator exp(Ts). This approximation does control the magnitude and phase as desired because the operator exp(Ts) has unity gain and linear phase, ($\phi = \tan^{-1} wT$).

To obtain the Padé approximations to exp(Ts), a ratio of two finite polynomials are equated to the exponential operator in the following manner:

---

[4] J. L. Stewart, Fundamentals of Signal Theory, McGraw-Hill, 1960, Chapter 5.

$$\frac{1 + a_1Ts + a_2 (Ts)^2 +\ldots + a_n(Ts)^n}{1 + b_1Ts + b_2 (Ts)^2 +\ldots + bm(Ts)^m} = e^{-Ts} \qquad (1.36)$$

Here, it is necessary to designate the number of poles and zeros desired in the approximation, which of course will determine (n) and (m) in Eq. (1.36) by a fundamental law of algebra. The infinite series representation of exp(Ts) is then substituted into Eq. (1.36). The expressions are cross multiplied and like powers of s equated and solved for the coefficients a_j and b_k.

An example: Let m = n = 1

$$F(s) = \frac{1 + a_1Ts}{1 + b_1Ts} = 1 + Ts + \frac{(Ts)^2}{2!} + \frac{(Ts)^3}{3!} + \ldots \qquad (1.37)$$

$$1 + a_1Ts = 1 + (b_1 + 1)Ts + (b_1 + \tfrac{1}{2})T^2s^2 + \ldots \qquad (1.38)$$

$$a_1 = b_1 + 1 \qquad 0 = b_1 + \tfrac{1}{2} \qquad (1.39)$$

$$a_1 = \tfrac{1}{2} \qquad b_1 = -\tfrac{1}{2} \qquad (1.40)$$

therefore:

$$F(s) = \frac{2 + Ts}{2 - Ts} \qquad (1.41)$$

The usual Padé approximations are those approximating the negative exponential instead of the positive one as in the above example. These of course are found as above with the only difference being the resulting sign. In the case of the approximation to exp(-Ts), Eq. (1.41) would change to

$$F(s) = \frac{2 - Ts}{2 + Ts} \tag{1.42}$$

When making use of the Padé approximation to exp(-Ts) and if stability of the resulting function is a factor of importance, there are restrictions which must be observed. If there are no finite zeros, the maximum number of poles permissible before instability occurs is four. A system with five or more poles and no finite zeros is definately unstable. If the system has one or more finite zeros it may have more than four finite poles and still remain stable. It is to be noted however, that any Padé approximation with finite zeros is non-minimum phase. Also, any system with equal number of poles and zeros is an all pass function.

Since we are making our approximations using Taylor series about the origin, the Padé functions are low pass. As we will observe later, there seems to be a strong indication that too many poles and zeros in a Padé approximation can lead to wild and erratic responses,

which may result in poor approximation of the time response function for small time.

1.11 Obtaining a Rational Function in Z from an Infinite Series in Z.[5]

Assume we have a time response function which we will designate f(t). Sample f(t) at points nT (n being integer values). Since these samples are nothing more than a series of weighted impulses, we can describe them with a function F(z) as

$$F(z) = b_o + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_n z^{-n} + \ldots \quad (1.43)$$

We now wish to find a rational function in z which, upon dividing denominator into the numerator, will yield exactly or approximately the F(z) in Eq. (1.43) over the range of interest. To accomplish this we employ the technique used in the Padé approximation and obtain the expression

$$F'(z) = \frac{p_o + p_1 z^{-1} + \ldots + p_v z^{-v}}{q_o + q_1 z^{-1} + \ldots + q_m z^{-m}} = b_o + b_1 z^{-1} + \ldots + b_n z^{-n} + \ldots \quad (1.44)$$

---

[5] F. BaHli, "A General Method for Time Domain Network Synthesis", IRE Trans.-Circuit Theory, 1954, pp. 21-28.

By cross multiplying Eq. (1.44) and equating like powers of z, we can solve for the unknown coefficients p_j and q_k. We set $q_o = 1$, and obtain

$$F'(z) = \frac{p_o + p_1 z^{-1} + \ldots + p_v z^{-v}}{1 + q_1 z^{-1} + \ldots + q_m z^{-m}} = b_o + b_1 z^{-1} + \ldots b_n z^{-n} + \ldots \qquad (1.45)$$

At this point, we have the desired function in z. A good check on the algebra at this stage is to divide denominator into numerator of the function F'(z) and see if F(z) is obtained.

There will be some question as to whether F'(z) is an exact function or an approximation to the transfer function that gave us the sampled time response. We can elucidate upon this point with the following observations:

a. If our initial time response was obtained from an irrational transfer function, then of course our F'(z) is an approximation.

b. If our sampled response was obtained from a rational function G(z), and if we select the corresponding number of poles and zeros for F'(z) as are in G(z), then F'(z) is an exact function. There will be no way of telling this from F'(z) alone however.

Most filter design techniques are in terms of the rational complex frequency variable s. We must therefore, convert the function F'(z) to a rational one in s.

We can represent F'(z) as an irrational form in s. immediately merely by substituting the defining relationship $z^N = e^{NTs}$ into the function F'(z), expanding exp(NTs) into its series, and collecting like powers of s. This is accomplished as follows:

$$F(s) = \frac{p_o e^{mTs} + p_1 e^{(m-1)Ts} + \ldots + p_v e^{(m-v)Ts}}{e^{mTs} + q_1 e^{(m-1)Ts} + \ldots + q_m} \tag{1.46}$$

$$= \frac{p_o\left[1+mTs+\frac{(mTs)^2}{2!}+\ldots\right]+\ldots+p_v\left[1+(m-v)Ts+\frac{\overline{(m-v)Ts}^2}{2!}+\ldots\right]}{\left[1+mTs+\frac{(mTs)^2}{2!}+\ldots\right]+q_1\left[1+(m-1)Ts+\frac{\overline{(m-1)Ts}^2}{2!}+\ldots\right]+\ldots+q_m}$$

After making the appropriate multiplications and collecting like powers in s, Eq. (1.46) reduces to

$$F(s) = \frac{n_o + n_1 s + n_2 s^2 + \ldots}{d_o + d_1 s + d_2 s^2 + \ldots} \tag{1.47}$$

The number of terms that must be included in Eq. (1.47) will be determined by the order of the rational function in s approximation.

We must now make an approximation of this irrational function F(s) by a rational one in s which we will designate F'(s). This will be accomplished much in the same manner as in obtaining F'(z) from F(z).

We may divide the denomerator of F(s) into its numerator and obtain an infinite series representation from which to make our rational approximation, or we may make our approximation directly from the F(s) in Eq. (1.47). Both methods give identical results.

If we choose to do the former, we would then equate the resulting infinite series to our rational function, F'(s), cross multiply and solve as we did in Eq. (1.45). If the latter procedure is followed, F'(s) would be equated directly to the F(s) obtained in Eq. (1.47).

We would then assume a solution for F'(s) by equating one of the coefficients to unity. For F'(s), it is convenient to set to unity the coefficient of the highest power of s in the denominator, thus obtaining

$$\frac{n_o + n_1 s + n_2 s^2 + \dots}{d_o + d_1 s + d_2 s^2 + \dots} = \frac{a_o + a_1 s + \dots + a_r s^r}{c_o + c_1 s + \dots + s^w} \qquad (1.48)$$

where the left side of Eq. (1.48) is the F(s) and the right side is F'(s).

This completes our procedure for obtaining a rational function in s from some time response. The function F'(s) is always an approximation to our original function F(z). However, as will be brought out in the

discussion of this point in Chapter Three, the approximation may in some cases approach the exact so closely that for engineering purposes it can pass as the exact.

In case this procedure is being used in a synthesis problem, a suitable filter can be designed directly from the $F'(s)$ by using conventional filter design techniques.

Chapter Two

EXAMPLES OF Z-TRANSFORM APPROXIMATIONS

2.1 Introduction

This chapter will be devoted primarily to examples of the approximation method discussed in Chapter One as applied to various time functions. Conclusions from these results are deferred to Chapter Three where a comprehensive evaluation of this entire method of approximation is made.

2.2 Criterion for Number of Samples

One of the first questions that must be answered when attempting to find a valid approximation to a given time function is how many samples are needed over the interval of interest. The criterion that sets the minimum number is the Sampling Theorem which states that to fully recover the information contained in a signal, we must sample at least twice the rate of the highest frequency component contained in the intelligence carrying portion of the signal.[1]

[1] J. R. Ragazzini, G. F. Franklin, Sampled-Data Control Systems, McGraw-Hill, New York, 1958, Chapt. 2, Section 2.

The Sampling Theorem alone tells us to select a sampling rate as high as possible to insure a high degree of accuracy. This unfortunately is not the only criterion to be considered.

From the coefficient matching technique of the Padé approximation, we note that there is a direct relationship between the number of samples used and the combined number of poles and zeros we have in $F'(z)$. That is

Number of samples = number of poles + number of zeros of $F'(z)$.

This situation will certainly set a practical upper limit on the number of sample points if $F'(z)$ must be factored. This is because of the obvious difficulty of factoring high order polynomials. This presents no problem here however, as we are not required to factor $F'(z)$.

What then is the factor, if one exist, that places an upper limit on the number of samples needed? At this stage of the approximation, there appears to be only one. In Chapter Three, others will be discussed.

The factor that is apparent at this stage is the stability of $F'(z)$. Since the right half-plane of the s-plane maps into the area outside the unit circle in the z-plane it is assured that, if a rational function in z is unstable, a corresponding rational function in s

will also be unstable. Therefore, a check should be made of F'(z) in any approximation of this type to determine its stability, and the number of sample points adjusted accordingly.

2.3 Criterion for Selection Pole-Zero Combination in F'(s).

After we obtain the F'(z), the next stage in the approximation requires a suitable pole-zero combination to be selected in F'(s). In fact, this selection determines the number of terms we must include in F(s).

There are two factors which aid in the selection of a suitable pole-zero combination in F'(s).

a. Stability

b. Physical shape of the initial time response.

In regard to stability, we must select a suitable combination of poles and zeros in F'(s) even though our F'(z) is stable. This is particularly true in using excessive zeros, which very often show up in the right half plane taking with them one or more poles. In any event, a test for stability should be made on F'(s) and appropriate adjustments made.

In considering the physical shape of the time response curve, much information can be gained concerning the number of poles and the pole-zero excess needed in $F'(s)$.[2] For example, in Figure (2.1) we see from (a) and from the Initial Value Theorem

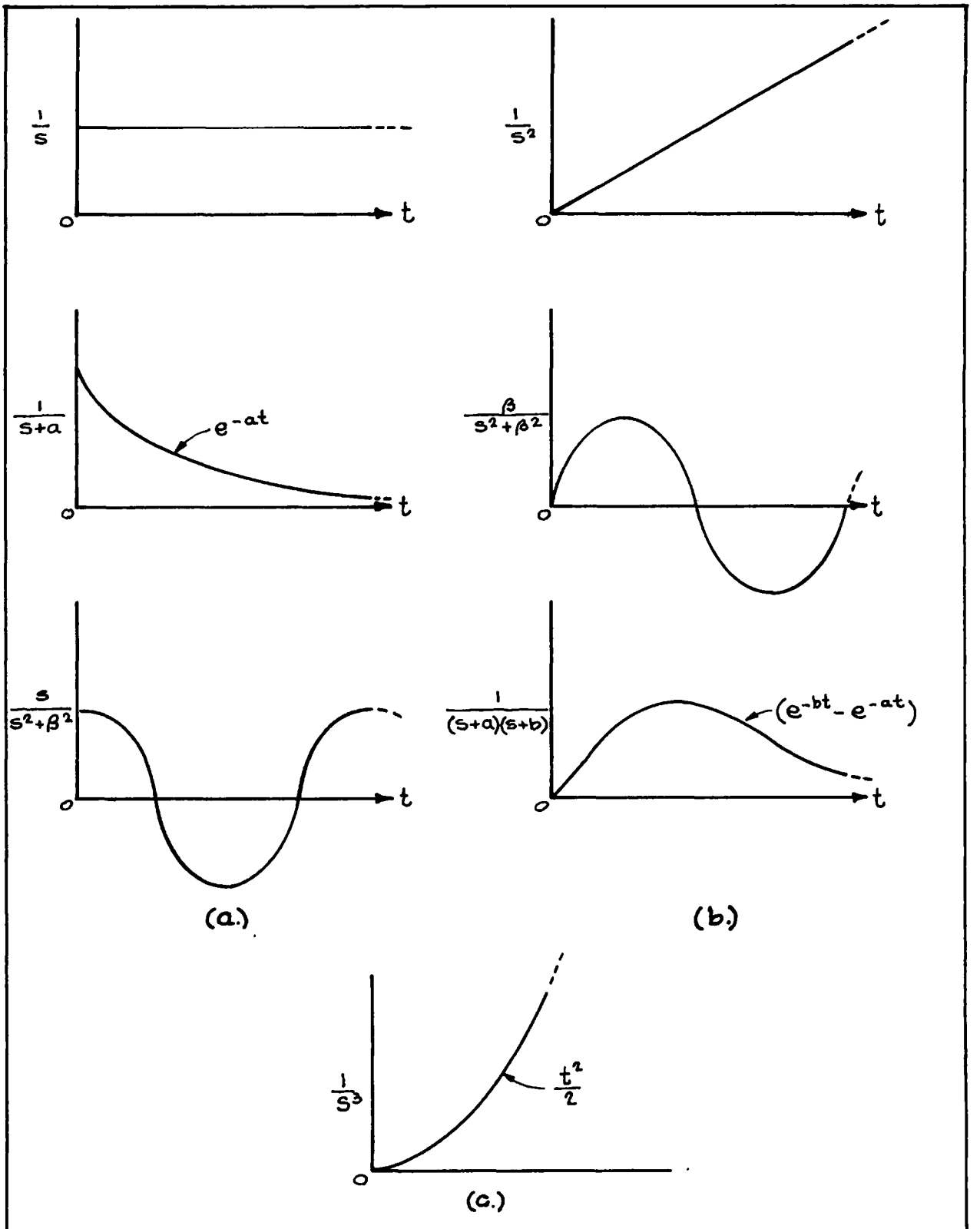
$$\lim_{t\to 0} f(t) = \lim_{s\to\infty} sF(s)$$

that a pole-zero excess of one will always have some finite value at the origin. From (b), we see that a pole-zero excess of two has zero value at the origin, but a finite slope. A pole-zero excess of three as indicated in (c) has zero initial value, zero slope, but a finite second derivative at the origin. This rule continues for higher pole-zero excesses.

The above observations give information for small time, but what about large time? About the only general rule that tells us about large time is the fact that to have a finite value for the function as $t\to\infty$ is for $F'(s)$ to have one pole at the origin. Other than that, we are unable to ascertain any other useful information regarding $F'(s)$ for large time.

---

[2] F. Ba Hli, "A General Method for Time Domain Network Synthesis", IRE Trans.-Circuit Theory, 1954, pp. 21-28.



CHARACTERISTICS AT THE ORIGIN

FIGURE 2.1

2.4 Examples of Approximations Applied.

a. In Figure (2.2), the time function is known to be $f(t) = 1 - e^{-t}$, which has the Laplace transform

$$F'(s) = \frac{1}{s(s + 1)} \tag{2.1}$$

The time axis is normalized so that T = 1 for convenience. (This will be the case for all examples shown in this chapter). Four samples are taken to give us the function

$$F(z) = .393z^{-1} + .632z^{-2} + .776z^{-3} + .864z^{-4} \tag{2.2}$$

This is approximated with

$$F'(z) = \frac{p_0 z^{-1}}{1 + q_1 z^{-1} + q_2 z^{-2} + q_3 z^{-3}} \tag{2.3}$$

When the algebraic manipulation is performed, there results

$$F'(z) = \frac{.393z^2}{z^3 - 1.61z^2 + .650z - .024} \tag{2.4}$$

We now make the substitution of

$$z^N = e^{NTs}$$

expand e^{NTs} into its infinite series, collect terms and obtain

$$F(s) = \frac{.393 + .393s + .196s^2 + \ldots}{.016 + .215s + .401s^2 .294s^3 + \ldots} \tag{2.5}$$

F'(s) is selected to have no finite zeros and two finite

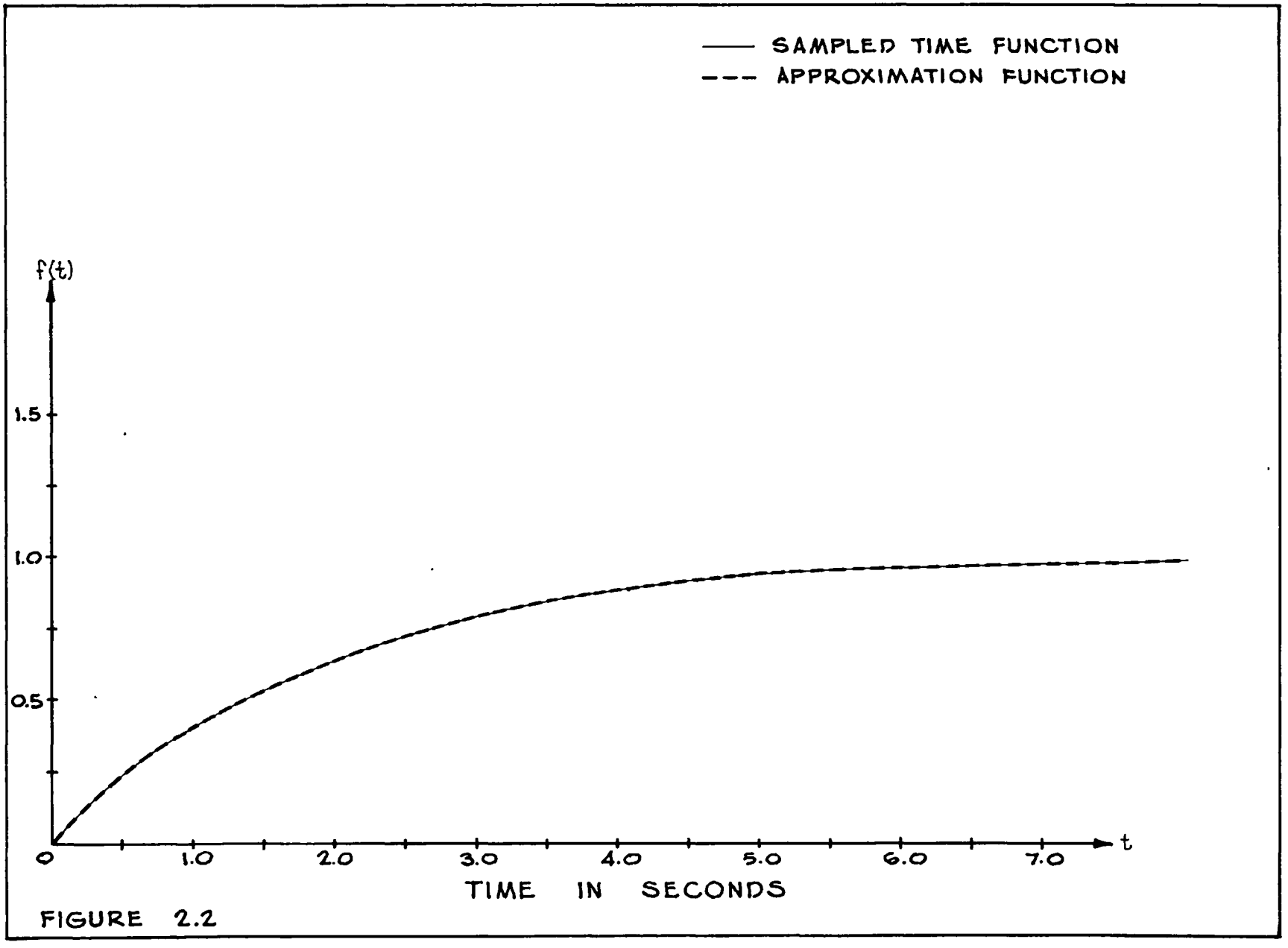

SAMPLED TIME FUNCTION
APPROXIMATION FUNCTION
f(t)
1.5
1.0
0.5
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
t
TIME IN SECONDS

FIGURE 2.2

poles giving

$$F'(s) = \frac{1.06}{s(s + 1.06)} \qquad (2.6)$$

Upon comparing Eq. (2.6) with Eq. (2.1), it is seen to be almost exact. The time responses of the two as seen in Figure 2.2 bears out this fact.

b. In Figures (2.3), (2.4), and (2.5) are time responses to step function excitation of three slightly more complicated transfer functions. By this is meant that the functions contain complex poles which of course yield oscillatory responses, and as seen in Figures (2.3) thru (2.5), shorter rise time and overshoot.

All three functions were sampled first with five samples yielding an F'(z) with two zeros and three poles. Only the approximation algebra for the response of Figure (2.4) will be included here as the accuracy of this function is representative of the other two.

Referring to Figure (2.4), it will be noticed that the time axis has two different sets of sample points, one with five and one with seven. Consider the one with five sample points first. The response was sampled at the five indicated points yielding

$$F(z) = .560z^{-1} + .745z^{-2} + .560z^{-3} + .565z^{-4} + .610z^{-5} \qquad (2.7)$$

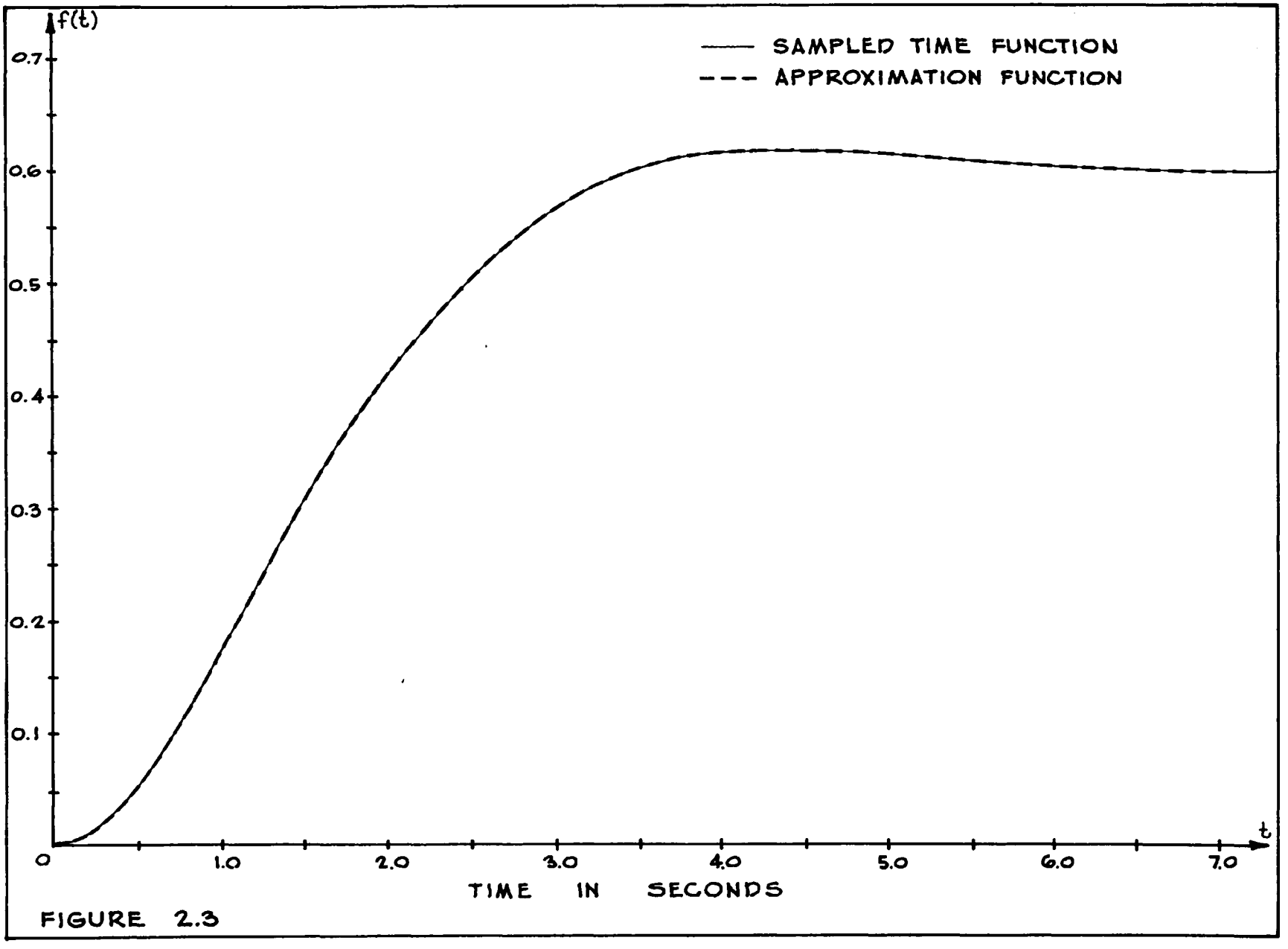

f(t)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
SAMPLED TIME FUNCTION
APPROXIMATION FUNCTION
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
t
TIME IN SECONDS

FIGURE 2.3

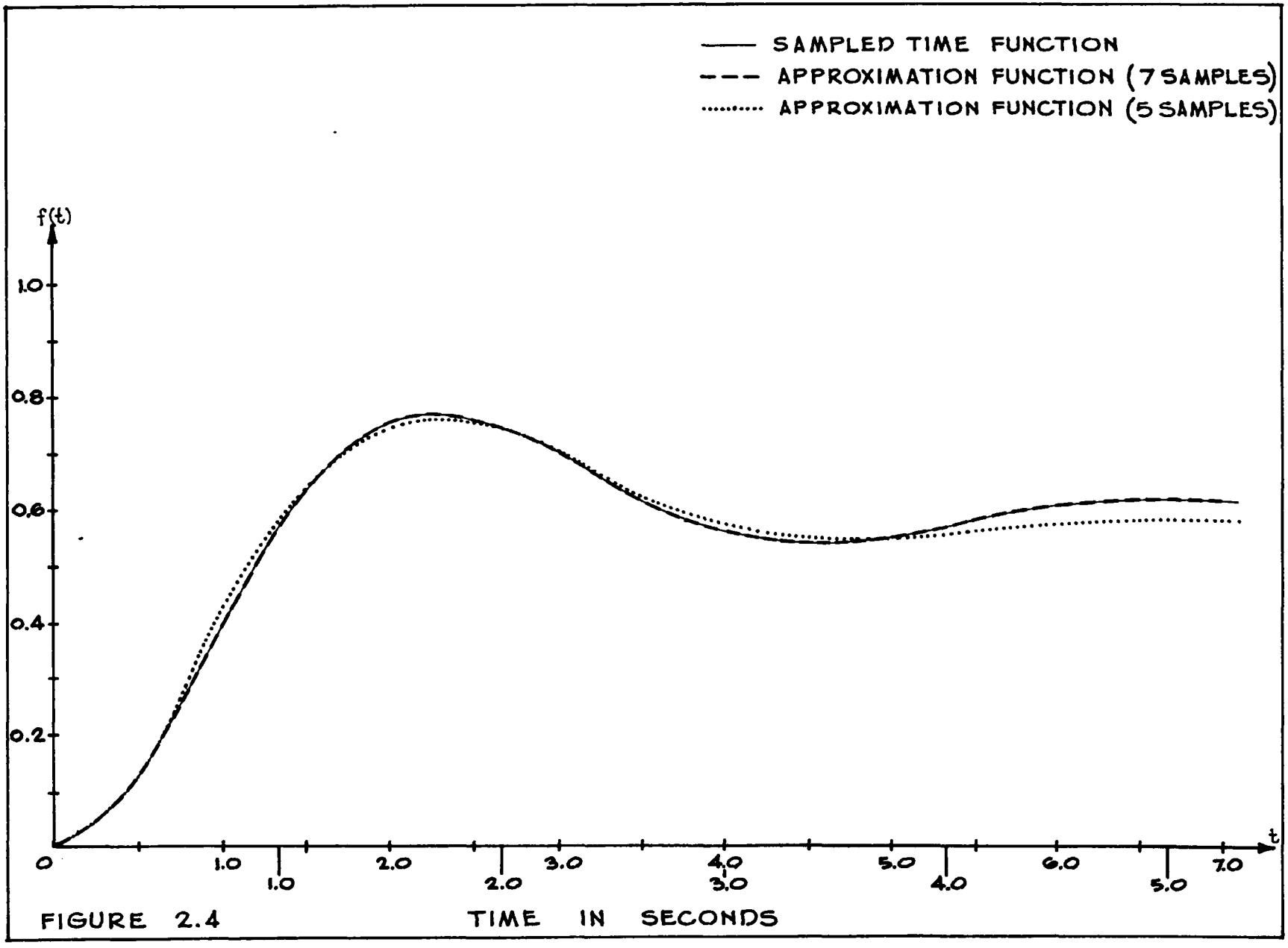

SAMPLED TIME FUNCTION
APPROXIMATION FUNCTION (7 SAMPLES)
APPROXIMATION FUNCTION (5 SAMPLES)
f(t)
1.0
0.8
0.6
0.4
0.2
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
t
1.0
2.0
3.0
4.0
5.0
TIME IN SECONDS

FIGURE 2.4

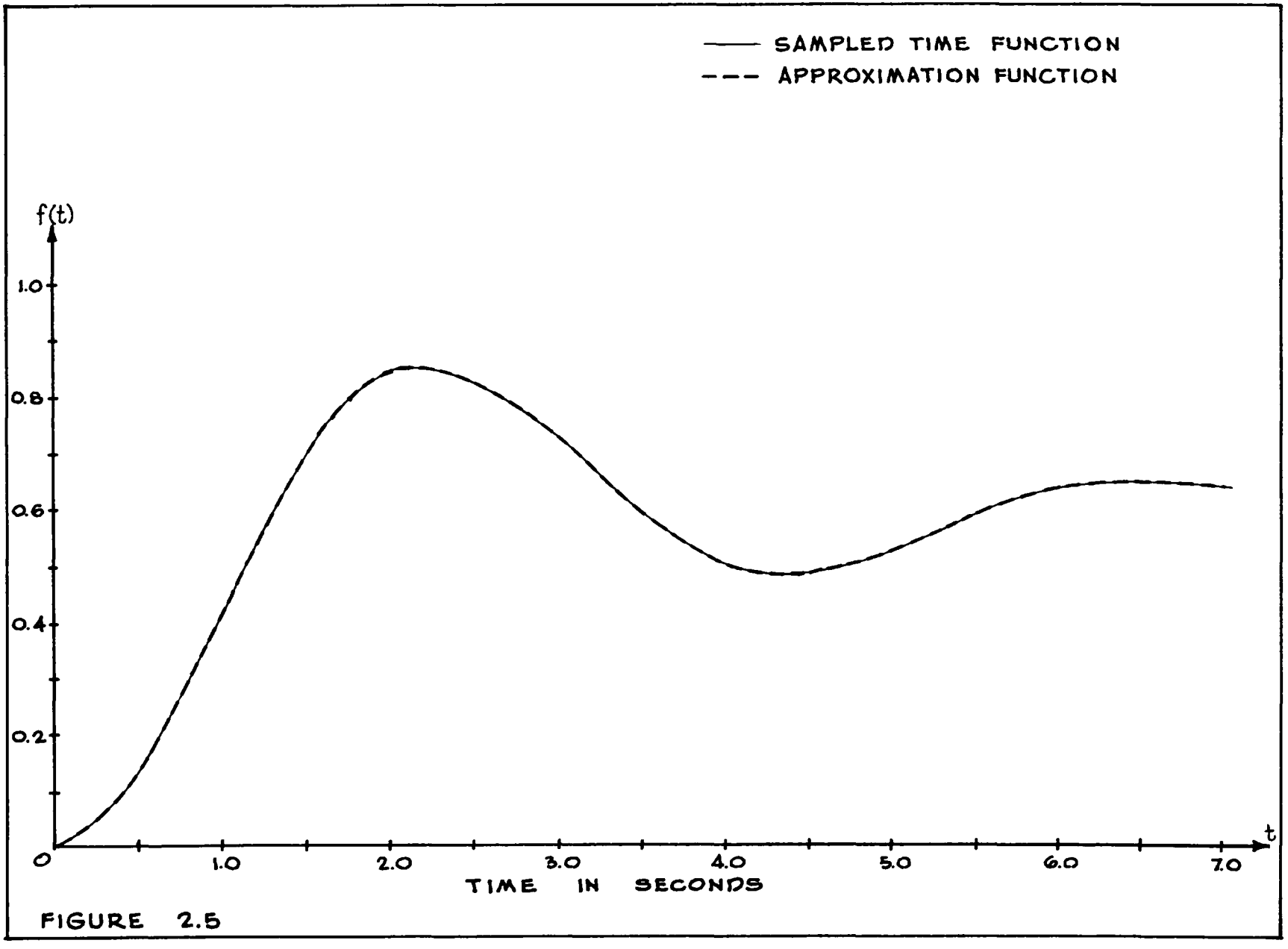

SAMPLED TIME FUNCTION
APPROXIMATION FUNCTION
f(t)
1.0
0.8
0.6
0.4
0.2
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
t
TIME IN SECONDS

FIGURE 2.5

This was approximated by the rational function in z of

$$F'(z) = \frac{.560z^2 + .335z}{z^3 - .733z^2 - .025z - .244} \qquad (2.8)$$

When e^{NTs} in series form is substituted for z^N and like terms collected, the resulting function in s is

$$F(s) = \frac{.895 + 1.455s + 1.288s^2 + \ldots}{1.510s + 3.022s^2 + 3.516s^3 + \ldots} \qquad (2.9)$$

This is in turn approximated with the desired transfer function

$$F'(s) = \frac{2.150}{s(s^2 + 1.350s + 3.630)} \qquad (2.10)$$

These same time response functions were again sampled at seven points as indicated in Figure (2.4). Again, only the approximation algebra for the response of Figure (2.4) is shown here. It is as follows:

$$F(z) = .390z^{-1} + .755z^{-2} + .700z^{-3} + .560z^{-4} + .550z^{-5} + .600z^{-6} + .610z^{-7} \qquad (2.11)$$

$$F'(z) = \frac{.390z^3 + 3.989z^2 + 2.634z}{z^4+8.291z^3-11.092z^2+5.156z-3.387} \qquad (2.12)$$

$$F(s) = \frac{7.013+11.781s+11.049s^2+7.512s^3}{11.845s+25.704s^2+34.047s^3+31.470s^4} \qquad (2.13)$$

$$F'(s) = \frac{1.245}{s(s^2 + 1.031s + 2.103)} \qquad (2.14)$$

c. The next group of time responses are for impulse excitation. These response functions are shown in Figures (2.6) and (2.7) where that of Figure (2.7) is the more underdamped.

The response of Figure (2.6) was sampled at seven sample points and the approximation made as follows:

$$F(z) = .390z^{-1} + .480z^{-2} + .445z^{-3} + .360z^{-4} + .265z^{-5} + .185z^{-6} + .122z^{-7} \qquad (2.15)$$

$$F'(z) = \frac{.390z^3+4.241z^2+.221z}{z^4+9.643z^3-12.444z^2+3.389z+.441} \qquad (2.16)$$

$$F(s) = \frac{4.852+9.872s+10.347s^2+7.446s^3}{2.030+11.431s+28.202s^2+38.035s^3+35.058s^4} \qquad (2.17)$$

$$F'(s) = \frac{.538}{s^2 + .808s + .225} \qquad (2.18)$$

In Figure (2.7) the seven samples were taken from the indicated points and the function in z formulated as

$$F(z) = .425z^{-1} + .560z^{-2} + .525z^{-3} + .420z^{-4} + .290z^{-5} + .180z^{-6} + .090z^{-7} \qquad (2.19)$$

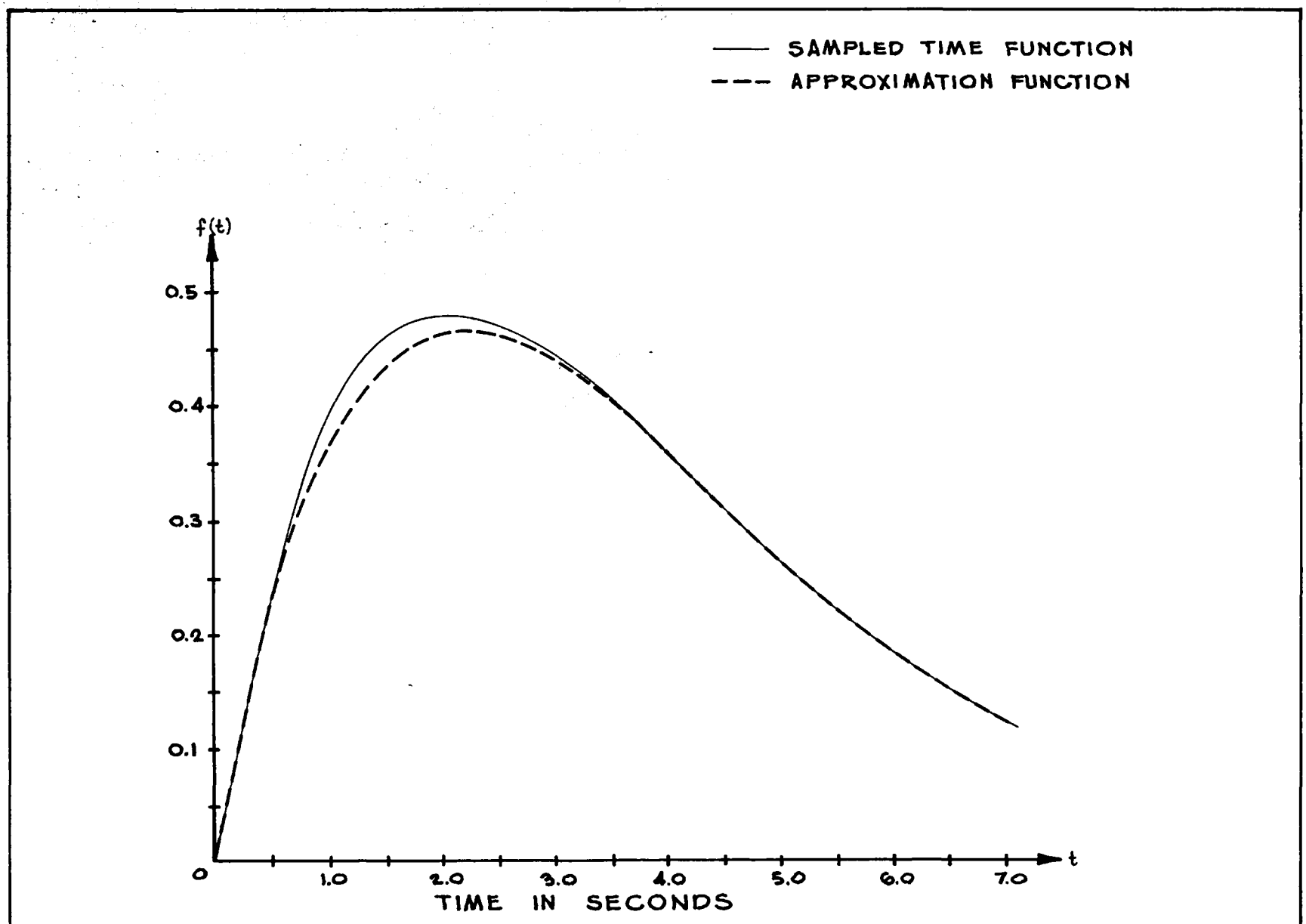

SAMPLED TIME FUNCTION
APPROXIMATION FUNCTION
f(t)
0.5
0.4
0.3
0.2
0.1
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
t
TIME IN SECONDS

FIGURE 2.6

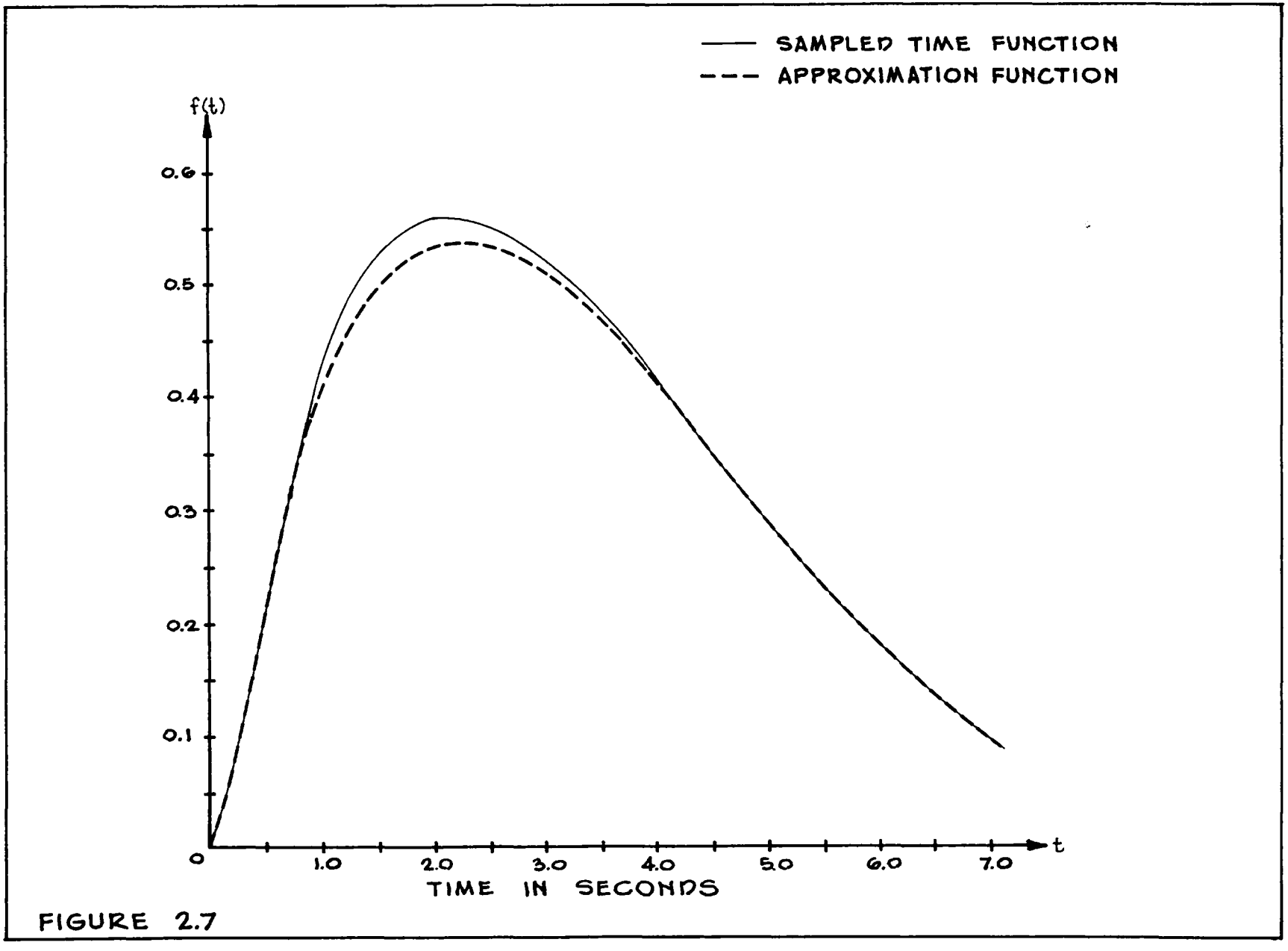

SAMPLED TIME FUNCTION
APPROXIMATION FUNCTION
f(t)
0.6
0.5
0.4
0.3
0.2
0.1
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
t
TIME IN SECONDS

FIGURE 2.7

which yields

$$F'(z) = \frac{.425z^3+.661z^2+.212z}{z^4+.238z^3+.933z^2-.530z+.305} \tag{2.20}$$

$$F(s) = \frac{1.348+2.859s+3.366s^2+2.838s^3+\ldots}{.557+2.796s+7.180s^2+10.468s^3+10.847s^4+\ldots} \tag{2.21}$$

The first approximation of this function by a rational function in s consisted of no finite zeros and two finite poles. This resulted in

$$F'(s) = \frac{.570}{s^2 + .683s + .236} \tag{2.22}$$

The time response of Eq. (2.22) is shown in Figure (2.7).

The second approximation to Eq. (2.21) was an F'(s) with one finite zero and three finite poles. This resulted in

$$F'(s) = .588\,\frac{s - .273}{s^3 + .423s^2 + .051s - .066} \tag{2.23}$$

which has a right half plane pole and hence unstable. However, the right half plane pole is at approximately s = .273, and therefore cancels with the zero leaving

$$F'(s) = \frac{.588}{s^2 + .696s + .241} \tag{2.24}$$

which is very close to the function in Eq. (2.22).

The time response in Figure 2.7 was then sampled first by five and then by nine samples. In both cases, the rational approximation in s was made with no finite zeros and two finite poles. Both approximations were less accurate than the one with seven sample points.

d. Still another group of time responses were approximated which were known to have come from transfer functions consisting of two zeros, real or complex, one real pole and a pair of complex poles. The example considered below has two real zeros.

The function was sampled at seven points giving

$$F(z) = .870z^{-1} + 1.190z^{-2} + 1.180z^{-3} + 1.080z^{-4} + 1.020z^{-5} + .990z^{-6} + 1.000z^{-7} \qquad (2.25)$$

which when rationalized yielded

$$F'(z) = \frac{.870z^3 + .986z^2 - .381z}{z^4 - .234z^3 - 1.474z^2 + 1.092z - .377} \qquad (2.26)$$

and hence

$$F(s) = \frac{1.475 + 4.202s + 5.697s^2 + 5.167s^3 + \ldots}{1.442s + 4.545s^2 + 7.830s^3 + 8.940s^4 + \ldots} \qquad (2.27)$$

Eq. (2.27) was then approximated by an F'(s) that contained the same number of poles and zeros as the exact transfer function. This gave

$$F'(s) = 2.246 \frac{s^2 + .369s + .221}{s(s^3 + 2.783s^2 + .956s + .484} \qquad (2.28)$$

which was a very poor approximation to the original time function.

The F'(s) was then made to have one zero and three poles as

$$F'(s) = .987 \frac{s + 1.040}{s(s^2 + 1.268s + 1.000)} \qquad (2.29)$$

which is shown in Figure (2.8) to be a fair approximation.

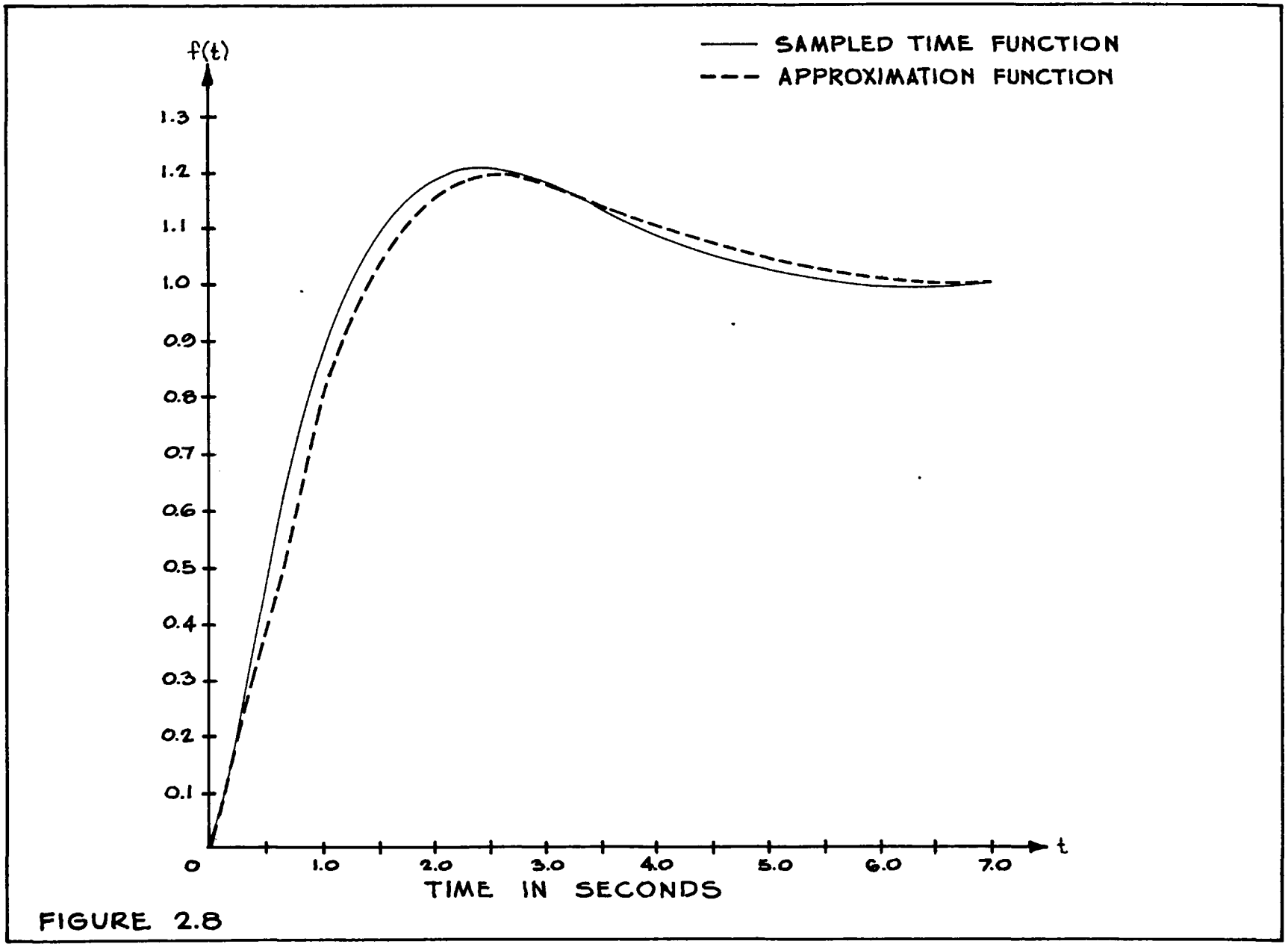

f(t)
SAMPLED TIME FUNCTION
APPROXIMATION FUNCTION
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
t
TIME IN SECONDS

FIGURE 2.8

Chapter Three

DISCUSSION AND CONCLUSIONS

3.1 Introduction

In this chapter, it is intended to consider the results of the approximation method introduced in Chapter One and applied in Chapter Two. The method and the results of the approximations will be analyzed and conclusions made as to its accuracy, area and range of validity, and its over all usefulness.

3.2 Representing a Time Function By a Series of Samples.

As was shown in Chapter One, to represent a time function by a series of weighted impulses is certainly justified in light of the fact that over a finite interval, the coninuous function itself is nothing more than an infinite number of such weighted impulses. However, in this representation, it is impossible to work with an infinite number of such sample points, thus the approximation with a finite number.

We shall now return to the question that was partly answered in Chapter Two regarding the number of sample points that are needed to validly approximate

a function over a specified interval. As was mentioned in Chapter Two, the Sampling Theorem and function stability set definite lower and upper limits respectively, on the number of time function samples required. However, some rule is sought that will specify a more definite number. As for example, consider the function in Figure (2.7), page 44. It was seen that five, seven, and nine samples gave stable functions in s, but the best approximation of the three was with seven samples. In Figure (2.4), page 39, seven samples gave better accuracy than did five samples. On the other hand, the function in Figure (2.2), page 36 was accurate with only four sample points as well as with seven. The indications are that the best accuracy will be obtained generally by sampling somewhere around the frequency halfway between the lowest set by the Sampling Theorem and the highest set by function stability. There probably exist other rules concerning the optimum number of sample points that were not revealed by this investigation.

The next point to consider is the pole and zero combination to use in the rational function of z. There are several factors that bear a direct relationship to this point. First, if the time function is zero at the origin, then F'(z) must by necessity have a pole-zero excess of at least one. Secondly, since the number of poles plus zeros determine the number of sample points

needed, we desire as many poles and zeros as possible. Thirdly, it is seen that upon dividing numerator into denominator of F(z), there is an error accumulation in the terms after the first or second. This error accumulates faster for larger pole-zero excesses.

Considering the above factors, it is obvious that for a time function that at the origin has zero value and finite slope, the best combination is with a pole-zero excess of one.

It has also been observed from the time response approximations included in this paper that any more than four poles in F'(z) leads to a poor approximation in F'(s). Considering this fact in addition to those discussed above leads to the following generalizations:

a. As long as the Sampling Theorem requirements are met with less than seven samples, a satisfactory approximation may be obtained with that number. For example, the function in Figure (2.2), page 36.

b. Good accuracy in approximation persists up to and including seven sample points. From that number up to the point where instability is reached, the accuracy of approximation will drop off accordingly. Hence, a low pass approximation.

A note should be made here regarding the number of significant figures used in the formulation of $F'(z)$ from $F(z)$. Since the error is cumulative, it is essential to use as many significant figures as possible and then round off in the final answer. The calculations in this paper for the most part, consisted of eight significant figures rounded off finally to three decimal places.

3.3 Obtaining a Function in s From the Rational Function in z.

In forming $F(s)$ from $F'(z)$, there is little to be discussed since it is a straightforward substitution of the defining relationship between s and z. Since the series of $\exp(Ts)$ is a rapidly converging series for small values of s, only the first few terms of the series are needed for accuracy. However, the coefficients multiplying the series should have as many significant figures retained as can conveniently be carried.

As was mentioned before, the number of terms needed in $F(s)$ will be determined by the number of poles and zeros desired in $F'(s)$.

3.4 Formulating the Rational Function in s.

The formulation of the rational function $F'(s)$ seems to be the most critical of all stages in the process of approximation. There are of course, several principles to follow in its formulation, but for the most part

experience indicates that trial and error should not be discredited in finding the right function.

As was stated in Chapter Two, the initial and final values of the time function will give information regarding the pole-zero excess and whether or not there is a pole at the origin. As for the number of poles (and consequently the number of zeros) there is not too much information to be had.

With regard to the trial and error method of obtaining the best combination of poles and zeros in $F'(s)$, a word may be said about the appearance of right half-plane poles. As is indicated by Eq. (2.23), page 45, the function in Figure (2.7), page 44, when approximated by an $F'(s)$ with one finite zero and three finite poles is unstable. However, it must be remembered that the function being sought is just an approximation and if, as in the case of Eq. (2.23), page 45, the right half-plane pole cancels with a zero leaving a stable function, a fairly accurate approximation may result. The above possibility should be explored whenever an unstable approximating function results.

3.5 The Apparent Range of Usefulness For This Type of Approximation.

As has been mentioned before in this paper, this method of approximation is valid only for small

values of s. Hence it is limited to low frequency functions. This is not to say this method of approximation is useless, however, as the majority of frequencies encountered in automatic control are in this range.

It is seen by the functions in Figures (2.2), page 36, (2.5), page 40, that if a time function comes from a low order, low frequency, rational transfer function, it can be approximated almost exactly by this method. Also, in the event the order of the exact transfer function of some time response is beyond the order range of this method, such as the one for Figure (2.8), page 48, it may be approximated by a lower order function that has most of the same characteristics. It is very likely that the higher orders of approximations that failed to give satisfactory results are outside the radius of convergence of our Taylor approximation.

3.6 Recommendations

The investigation of this facet of the z-transform use can be carried much further than has been shown in this paper, resulting in a great deal more useful information along this line being revealed. For example, this method of approximation was applied to irrational functions of which some contained discontinuities in the range of interest. Various degrees of success were obtained, but the investigation conducted was too brief

to warrant inclusion in this paper.

Also, it is believed that the number of sample points may be increased over seven and still retain an equal degree of accuracy by introducing artifices such as time delay or possibly hold functions into the approximation. It will certainly warrant further study.

BIBLIOGRAPHY

Aseltine, John A., Transform Method in Linear System Analysis, McGraw-Hill Book Company, Inc. New York, 1958

Ba Hli, F., "A General Method for Time Domain Network Synthesis", IRE Trans, Circuit Theory, 1954

Boxer, R., "A Note on Numerical Transform Calculus", Proc. IRE, vol. 45, no. 10, pp. 1401-1406, Oct. 1957.

Boxer, R; Thaler, S., "A Simplified Method of Solving Linear and Non-Linear Systems", Proc. IRE, vol. 44, no. 1, pp. 89-101, 1956

Ragazzini, J. R.; Franklin, G. F., Sampled-Data Control Systems, McGraw-Hill Book Company, Inc. New York, 1958

Salzer, J. M., "The Frequency Analysis of Digital Computers Operating in Real Time", Proc. IRE vol. 42, No. 1, pp. 457-466, Feb. 1954.

Scarborough, J. B., Numerical Mathematical Analysis Third Edition, The Johns Hopkins Press, Baltimore, Md., 1955

Steffensen, J. F., Interpolation, Chapter 4, The Williams and Wilkins Co., Baltimore, Md. 1952

Stewart, John L., Circuit Theory and Design, John Wiley and Sons, Inc., New York, 1956

Stewart, John L., Fundamentals of Signal Theory McGraw-Hill, 1960

Thaler, S; Boyer, R., "An Operational Calculas for Numerical Analysis", IRE Conv. Rec., pt. 2, pp. 100-105, 1956

Tou, Julius T., Digital and Sampled-Data Control Systems, McGraw-Hill Book Company, Inc., New York, 1959