THE INVESTIGATION OF OPTIMAL DISCRETE APPROXIMATIONS FOR REAL TIME FLIGHT SIMULATIONS

Final Technical Report
Grant No. NASA NSG 1151

Submitted to:

NASA Scientific & Technical Information Facility
P. O. Box 8757
Baltimore/Washington International Airport
Maryland 21240

Submitted by:

E. A. Parrish

E. S. McVey

G. Cook

K. Henderson

SCHOOL OF ENGINEERING AND APPLIED SCIENCE

RESEARCH LABORATORIES FOR THE ENGINEERING SCIENCES

UNIVERSITY OF VIRGINIA

CHARLOTTESVILLE, VIRGINIA 22901

Report No. EE-4041-102-76

March 1976

THE INVESTIGATION OF OPTIMAL DISCRETE APPROXIMATIONS FOR REAL TIME FLIGHT SIMULATIONS

Final Technical Report
Grant No. NASA NSG 1151

Submitted to:

NASA Scientific & Technical Information Facility
P. O. Box 8757
Baltimore/Washington International Airport
Maryland 21240

Submitted by:

E. A. Parrish

E. S. McVey

G. Cook

K. Henderson

Department of Electrical Engineering
RESEARCH LABORATORIES FOR THE ENGINEERING SCIENCES
SCHOOL OF ENGINEERING AND APPLIED SCIENCE
UNIVERSITY OF VIRGINIA
CHARLOTTESVILLE, VIRGINIA

Report No. EE-4041-102-76
March 1976

Copy No. 1

I. INTRODUCTION

This report summarizes the results obtained during the first year of a continuing effort on the general topic of discrete approximations for real-time flight simulation. The report is divided into five major topics as follows:

1. Digital Autopilot Modelling--consideration of the particular problem of approximation of continuous autopilots by digital autopilots.

2. Frequency Domain Synthesis of Discrete Representations--use of Bode plots and synthesis of transfer functions by asymptotic fits in a warped frequency domain.

3. Substitutional Methods--an investigation of the various substitution formulas, including the effects of nonlinearities.

4. Use of Pade approximation to the solution of the matrix exponential arising from the discrete state equations.

5. An Analytical Integration of the State Equation Using Interpolated Input--uses polynomial approximations to the input signal and integrates the state equations analytically.

Each of the sections is self-contained in that the developments and conclusions are contained in each section.

The work presented in this report represents the initial efforts and preliminary investigations of the topics covered. The actual application of the various techniques has been limited to some rather simple linear and nonlinear systems, while the emphasis has been placed on theoretical investigations. This emphasis will shift somewhat in future efforts as more complex and nonlinear systems are simulated with the techniques which appear most promising as a result of the work presented in this report.

II. DIGITAL AUTOPILOT DESIGN

Introduction

The purpose of this part of the research is to establish criteria for approximating continuous autopilots with digital autopilots. It is desired that the performance of an aircraft with a digital autopilot be similar enough to the continuous autopilot which it is to replace that pilots cannot perceive a difference between them. This is the first of three levels of work being considered in the investigation of digital autopilots. The three levels are:

1. The problem of replacing existing continuous autopilots with a digitally implemented autopilot that is as nearly indistinguishable as practical from the continuous autopilot.

2. Design of autopilots using digital control methods based on original performance specifications.

3. Respecification of aircraft control systems and use of modern methods to control and optimize system performance for flight situations, such as optimal fuel control which existing autopilots do not provide.

Although the ideal situation is for the performance of new digital autopilots to be identical to the continuous systems they replace, it is easy to prove that, in general, it is not possible to find exact discrete equivalents to continuous systems. The principal differences in the output signals will be:

1. phase shift and attenuation due to the zero order hold at the computer output.

2. phase shift and gain due to the digital representation of the continuous autopilot transfer function.

3. aliasing of high frequency input components.

4. distortion components due to sample-hold operations.

Although these can result in major differences in the results of the two implementations, it should be possible to make them conform to commonly used control specifications to within any accuracy desired if the sampling frequency of the digital system is made high enough; but,

unfortunately, this parameter is usually restricted because it is desired to keep the sampling frequency as low as practical to minimize computer capacity. Sample frequency is a major parameter in the study.

A logical approach and the one used here is to use design specifications of continuous autopilot systems, but keep in mind also that the origiral design was based on continuous control technology and should take into account any new factors introduced by a digital controller. The original specifications are not available but they undoubtedly [2.1] consisted of such factors as phase margin, gain margin, magnification, closed loop pole locations, rise time, overshoot, delay time, settling time, mean squared error to stochastic inputs, final value of error, dynamic tracking error, and bandwidth.

Investigation of Gain and Phase Shift Errors Due to Discretization

Discrete approximation techniques to represent continuous functions in digital computers are studied and compared in the following. As already stated, the objective is to design digital controllers which will substitute for existing analog controllers, such that the digital controllers are as similar in function as practical to the hardware they replace. Evaluation on an input-output basis will make use of phase and gain specifications.

Consider the autopilot transfer function for a commercial aircraft as shown in Fig. 2.1. It is

$$\frac{Y(s)}{U(s)} = \frac{36(s + 1.65)(s^2 + 2.31s + 2.72)}{(s + 0.62)(s^2 + 5.62s + 3.1)(s^2 + 8.4s + 36)} \qquad (2.1)$$

where U(s) is the Laplace transform of the pitch rate input signal and Y(s) is the Laplace transform of the control signal. The Bode diagram for the above transfer function is shown in Fig. 2.2. It is inferred from the figure that the crossover frequency for the closed loop system should be located at about $\omega = 3$ rad/sec. The crossover frequency is of special importance and will be used in later discussions.

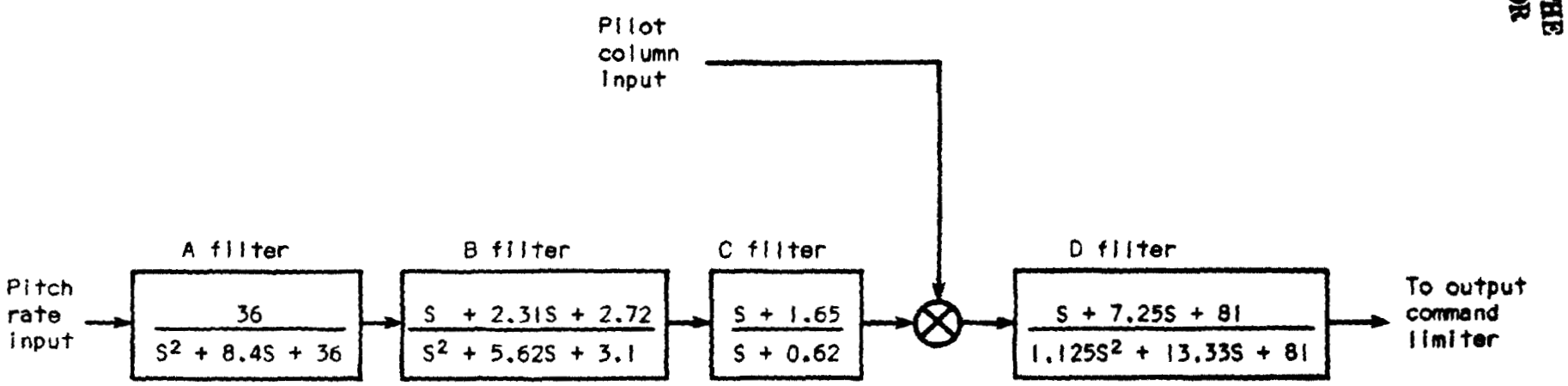


Figure 2.1 Autopilot Transfer Function for a Commercial Aircraft

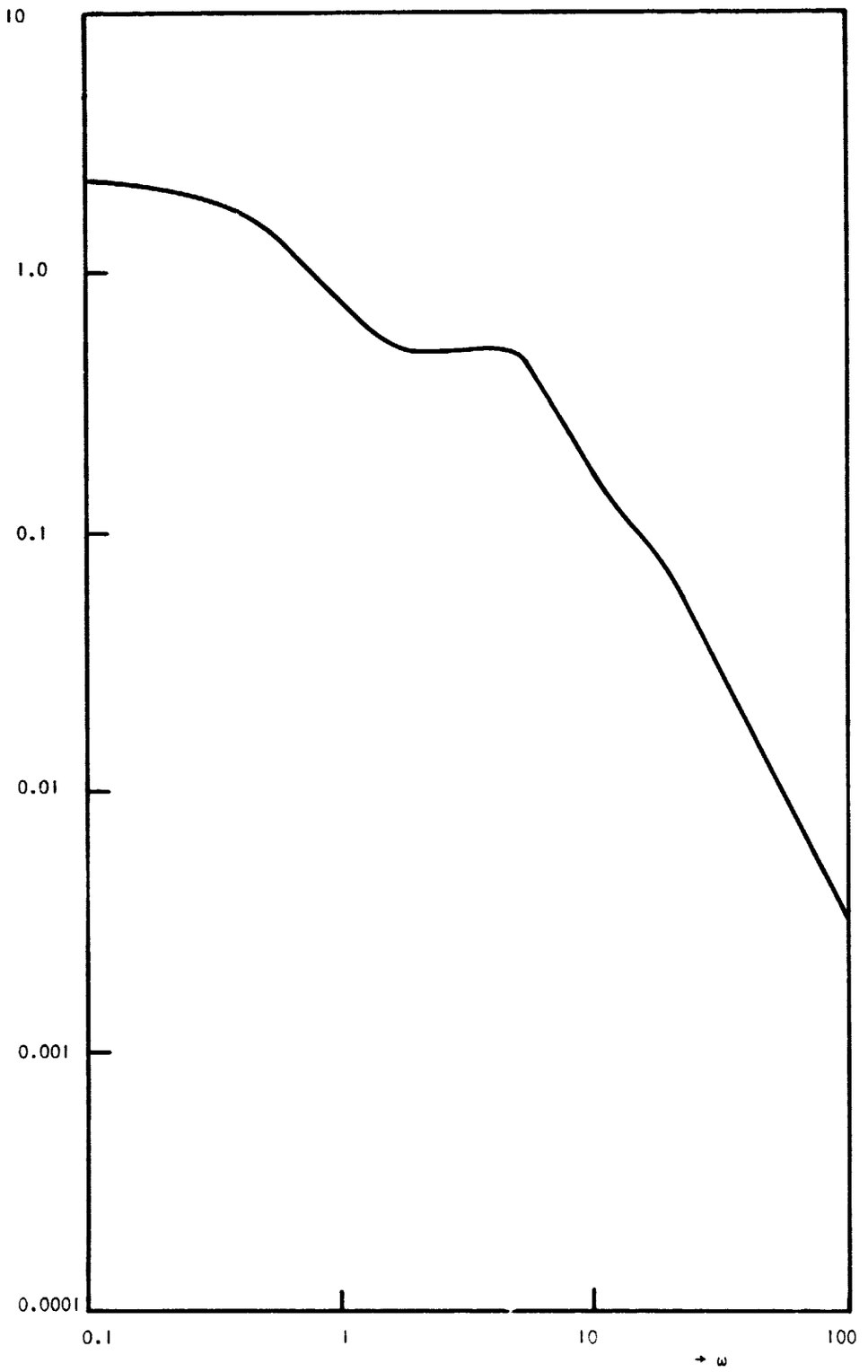


Figure 2.2 Attenuation Diagram for Continuous Autopilot

Discrete Approximation

If G(s) is a continuous transfer function as shown in Fig. 2.3(a), its discrete counterpart would appear as in Fig. 2.3(b) where

$$Y(s) = G_c(s)E(s) \tag{2.2}$$

$$Y^*(s) = [G_c(s)E(s)]^* \tag{2.3}$$

$$Y_d^*(s) = G_A^*(s)E^*(s) \tag{2.4}$$

The ideal discrete system for the applications here would be one that yields

$$y'(t) = y(t) \quad \text{for all } t \geq 0 \tag{2.5}$$

However due to the discrete nature of the digital controller, it is impossible to realize Eq. (2.5) exactly. The best possible discrete system could give, instead,

$$y'(t) = y(t) \quad \text{for } t = 0,T,2T, \ldots \tag{2.6}$$

where T is the sampling period. Because

$$y'(nT) = y_d(nT) \quad n = 0,1,2, \ldots \tag{2.7}$$

in this case Eq. (2.6) is satisfied

$$y(nT) = y_d(nT) \quad n = 0,1,2, \ldots \tag{2.8}$$

or

$$Y^*(s) = Y_d^*(s) \tag{2.9}$$

Using Eqs. (2.3), (2.4), and (2.9),

$$[G_c(s)E(s)]^* = G_A^*(s)E^*(s). \tag{2.10}$$

Taking the z-transform of both sides of Eq. (2.10),

$$Z\{G_c(s)E(s)\} = G_A(z)E(z) \tag{2.11}$$

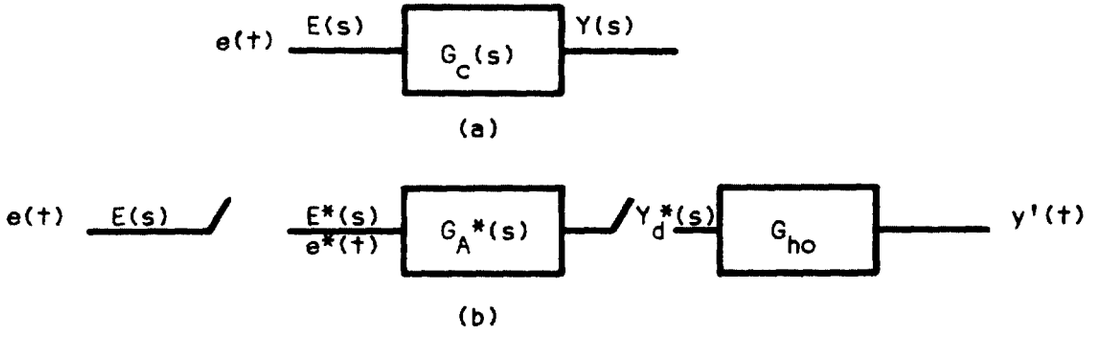


Figure 2.3 (a) Continuous Transfer Function and (b) Its Discrete Counterpart

or

$$G_A(z) = \frac{Z\{G_c(s)E(s)\}}{E(z)} \tag{2.12}$$

where $Z\{\cdot\}$ indicates the z-transform.

In Eq. (2.12), D(z) cannot be obtained independent of the input function e(t). This justifies the previous assertion that an exact digital equivalent of the continuous autopilot cannot be obtained. Thus one must settle for a difference equation transfer function that gives an acceptable approximation to the solutions for an arbitrary input, e(t).

Various discrete approximation techniques have been proposed to transform a previously designed continuous controller into a discrete equivalent. These discrete approximation techniques can be divided into three broad categories: (1) numerical methods (2) operational methods (3) input approximation methods.

Numerical methods generally provide a means of obtaining accurate approximation. However, since these methods generally take a great deal of calculation time, their application to discrete approximations is often limited. In the operational methods, every integrating operator of the continuous transfer function is replaced by a discrete integrating operator in order to obtain the discrete transfer function. In the input approximation methods, the input is assumed to be approximated in a certain way, for example, by a stair-step function, or by straight line segments, etc.

In the literature, Tustin's method is generally preferred among the operational methods and the linear segment input approximation method is considered to be one of the best input approximation methods. These two methods will be used in the following discussion to show that the degree to which the digital control system approximates the continuous system is determined primarily by the zero-order hold rather than by the discrete representation (i.e., discretization) of the continuous controller transfer function. Various methods for discretizations are compared later.

Tustin's Approximation

Tustin's method enjoys the merit of simplicity. This method is usually quite accurate, does not introduce spurious solutions, and is ideally suited for operational calculus operations due to its cascading property. To apply Tustin's method, first divide the numerator and denominator by the highest power of s in the denominator, s^n, then replace $\left(\frac{1}{s}\right)^i$ by $\left(\frac{T(z+1)}{2(z-1)}\right)^i$ to obtain the discrete transfer function.

For G(s) consisting of a simple pole at -a,

$$G_{AT}(z) = Z_{Tustin}\left\{\frac{1}{s+a}\right\}$$

$$= Z_{Tustin}\left\{\frac{\frac{1}{s}}{1+\frac{a}{s}}\right\} = \frac{\frac{T}{2}\cdot\frac{z+1}{z-1}}{1+a\cdot\frac{T}{2}\cdot\frac{z+1}{z-1}} = \frac{T(z+1)}{2(z-1)+aT(z+1)}$$

$$= \frac{T(z+1)}{(aT+2)z+(aT-2)} \tag{2.13}$$

Straight Line Approximation

This technique assumes that the excitation can be approximated by a series of straight-line segments as shown in Fig. 2.4(a) for an arbitrary function. The piecewise linear input may be applied to any system by placing a sample and a triangular hold (unrealizable first order hold) before the normal input as shown in Fig. 2.4(b). For the above example

$$G_{AS}(z) = Z\left\{G_{h_1}(s)G(s)\right\} = Z\left\{\frac{e^{TS}(1-e^{-TS})^2}{TS^2}\cdot\frac{1}{s+a}\right\}$$

$$= \frac{z(1-z^{-1})^2}{T}Z\left\{\frac{1}{s^2(s+a)}\right\}$$

$$= \frac{z(1-z^{-1})^2}{T}Z\left\{\frac{1/a}{s^2}+\frac{(-1/a^2)}{s}+\frac{1/a^2}{s+a}\right\}$$

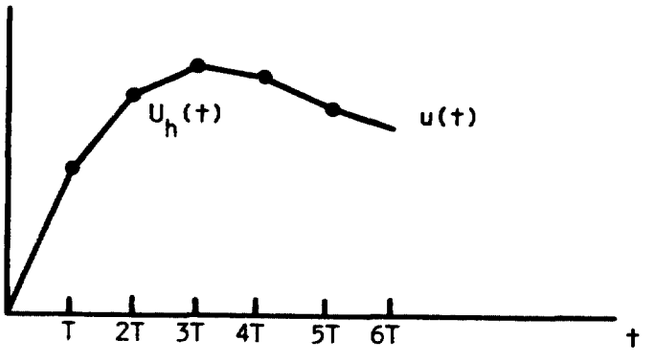


Figure 2.4 (a) Straight-line Segment Approximation

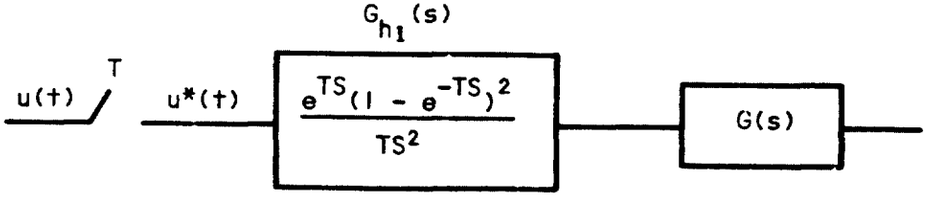


Figure 2.4 (b) Application of Triangular Hold, $G_{h_1}(s)$

$$= \frac{(z-1)^2}{Tz}\left[\frac{1}{a}\cdot\frac{Tz}{(z-1)^2} - \frac{1}{a^2}\cdot\frac{z}{z-1} + \frac{1}{a^2}\cdot\frac{z}{z-e^{-aT}}\right]$$

$$= \frac{(z-1)^2}{Tz}\cdot\frac{[a(z-e^{-aT})Tz - z(z-1)(z-e^{-aT}) + z(z-1)^2]}{a^2(z-1)^2(z-e^{-aT})}$$

$$= \frac{aT(z-e^{-aT}) - (z-1)(z-e^{-aT}) + (z-1)^2}{Ta^2(z-e^{-aT})}$$

$$= \frac{(e^{-aT} + aT - 1)z + (1 - e^{-aT} - aTe^{-aT})}{Ta^2(z-e^{-aT})} \qquad (2.14)$$

for the straight-line approximation.

Gain and Phase Shift

An approximation of continuous autopilots with digital autopilots should include matching the gain, phase shift, and steady state errors of the systems.

The gain and phase shift of the digital and analog controllers in the frequency domain will now be considered.

Referring to Fig. 2.3(b) and using Eq. (2.4),

$$Y'(s) = G_{ho}(s)Y_d^*(s) \qquad (2.15)$$

$$= G_{ho}(s)G_A^*(s)E^*(s) \qquad (2.16)$$

or

$$Y'(j\omega) = G_{ho}(j\omega)G_A^*(j\omega)E^*(j\omega) \qquad (2.17)$$

The frequency range of interest is

$$\omega < \omega_s/2$$

where ω is the input frequency and ω_s is the sampling frequency.

If the sampling theorem is satisfied (and it will be for cases of interest here)

$$E^*(j\omega) = \frac{1}{T} E(j\omega) \quad \text{for } \omega < \omega_s/2 \tag{2.18}$$

Thus Eq. (2.17) becomes

$$Y'(j\omega) = G_{ho}(j\omega)G_A^*(j\omega)\frac{1}{T}E(j\omega) \tag{2.19}$$

or

$$\frac{Y'}{E}(j\omega) = \frac{G_{ho}}{T}(j\omega)G_A^*(j\omega) \tag{2.20}$$

For the continuous transfer function in Fig. 2.3(a)

$$\frac{Y}{E}(j\omega) = G(j\omega) \tag{2.21}$$

The transfer function obtained from Eq. (2.20) is to be matched as closely as possible to the transfer function of Eq. (2.21).

For the gain

$$\left|\frac{Y'}{E}(j\omega)\right| = \left|\frac{G_{ho}(j\omega)}{T}\right| \left|G_A^*(j\omega)\right| \tag{2.22}$$

and

$$\left|\frac{Y}{E}(j\omega)\right| = \left|G(j\omega)\right| \tag{2.23}$$

Therefore

$$\left|\frac{\frac{Y'}{E}(j\omega)}{\frac{Y}{E}(j\omega)}\right| = \left|\frac{G_{ho}(j\omega)}{T}\right| \cdot \left|\frac{G_A^*(j\omega)}{G(j\omega)}\right| \qquad \omega < \omega_s/2 \tag{2.24}$$

For the phase shift

$$\angle\frac{Y'}{E}(j\omega) = \angle G_{ho}(j\omega)/T + \angle G_A^*(j\omega) \tag{2.25}$$

and

$$\angle \frac{Y}{E}(j\omega) = \angle G(j\omega) \tag{2.26}$$

Therefore

$$\angle \frac{Y'}{E}(j\omega) - \angle \frac{Y}{E}(j\omega) = \angle G_{ho}(j\omega)/T + \angle (G_a^*(j\omega) - G(j\omega)), \quad \omega < \omega_s/2 \tag{2.27}$$

As indicated in Eq. (2.24) and Eq. (2.27) the differences in the gain and the phase shift of the digital controller and the analog controller can be divided into two terms; one is due to $G_{ho}(j\omega)$; the other is due to $G_A^*(j\omega)/G(j\omega)$. The former depends only on the sampling frequency, whereas the latter would depend on the specific continuous transfer function to be discretized and on the specific discretization technique used as well as on the sampling frequency. It is interesting to compare the differences due to the two terms above in a specific example.

Consider the transfer function

$$G(s) = \frac{1}{s + 1}$$

Tustin's approximation from Eq. (2.13) with a = 1 is

$$G_{AT}(z) = \frac{T(z + 1)}{T(z + 1) + 2(z - 1)}$$

For $f_s = 5$, $T = .2$

$$G_A(z) = \frac{0.2(z + 1)}{0.2(z + 1) + 2(z - 1)} = \frac{0.090909(z + 1)}{z - 0.818182}$$

$$G_A^*(j\omega) = \frac{0.090909(e^{j0.2\omega} + 1)}{e^{j0.2\omega} - 0.818182} \tag{2.28}$$

For $f_s = 10$, $T = .1$

$$G_{AT}^*(j\omega) = \frac{0.047619(e^{j0.1\omega} + 1)}{e^{j0.1\omega} - 0.904762} \tag{2.29}$$

For $f_s = 20$, $T = 0.05$

$$G_{AT}^*(j\ ) = \frac{0.02439(e^{j0.05} + 1)}{e^{j0.05} - 0.95122} \tag{2.30}$$

The straight-line approximation from Eq. (2.14) with a = 1 is

$$G_{AS}(z) = \frac{(T - 1 + e^{-T})z + 1 - (T + 1)e^{-T}}{T(z - e^{-T})}$$

For $f_s = 5$, $T = .2$

$$G_{AS}(z) = \frac{0.018731z + 0.017523}{0.2(z - 0.818731)} = \frac{0.093655z + 0.087615}{z - 0.818731}$$

$$G_{AS}^*(j\omega) = \frac{0.093655(e^{j0.2\omega} + 0.935508}{e^{j0.2\omega} - 0.818731} \tag{2.31}$$

For $f_s = 10$, $T = .1$

$$G_{AS}^*(j\omega) = \frac{0.04837(e^{j0.1\omega} + 0.96734)}{e^{j0.1\omega} - 0.904837} \tag{2.32}$$

For $f_s = 20$, $T = 0.05$

$$G_{AS}^*(j\omega) = \frac{0.02458(e^{j0.05\omega} + 0.98373)}{e^{j0.05\omega} - 0.951229} \tag{2.33}$$

For the zero-order hold

$$G_{ho}(j\omega) = \frac{2\pi}{\omega_s} \cdot \frac{\sin \pi(\omega/\omega_s)}{\pi(\omega/\omega_s)} e^{-j\pi(\omega/\omega_s)}$$

$$|G_{ho}(j\omega)| = T \cdot \left|\frac{\sin \Pi(\omega/\omega_s)}{\Pi(\omega/\omega_s)}\right| \tag{2.34}$$

and

$$\angle G_{ho}(j\omega) = -\Pi\left(\frac{\omega}{\omega_s}\right) \operatorname{sgn}[\sin \Pi(\omega/\omega_s)] \tag{2.35}$$

And, from Eq. (2.34)

$$\left|\frac{G_{ho}(j\omega)}{T}\right| = \left|\frac{\sin \Pi(\omega/\omega_s)}{\Pi(\omega/\omega_s)}\right| \tag{2.36}$$

For the frequencies of interest,

$$\operatorname{sgn}[\sin \Pi(\omega/\omega_s)] = 1.$$

Therefore,

$$\angle G_{ho}(j\omega) = -\Pi\left(\frac{\omega}{\omega_s}\right) \tag{2.37}$$

The results are shown in Figs. 2.5, 2.6, 2.7, 2.8, and 2.9. Fig. 2.5 shows $|G(j\omega)|$ and $|G_A^*(j\omega)|$ as a function of frequency for different sampling frequencies. The distinction between the two different approximation methods has been ignored because the differences are negligible for the present purpose. This is discussed later in more detail. It is noted that in the low frequency range, $|G_A^*(j\omega)|$ follows $|G(j\omega)|$ almost perfectly and in the high frequency range, $|G_A^*(j\omega)|$ follows $|G(j\omega)|$ more closely as the sampling rates get higher as expected. Figure 2.6 shows $\left|\frac{G_{ho}}{T}\right|$ and $|G_A^*(j\omega)/G(j\omega)|$ as a function of frequency. From Fig. 2.7 to Fig. 2.9 $\angle G_{ho}(j\omega)$ and $\angle G_A^*(j\omega) - \angle G(j\omega)$ are shown for different sampling frequencies. It is noted that $\angle G_{ho}(j\omega)$ is much larger than $\angle G_A^*(j\omega) - \angle G(j\omega)$ over most of the frequency range for all sampling frequencies.

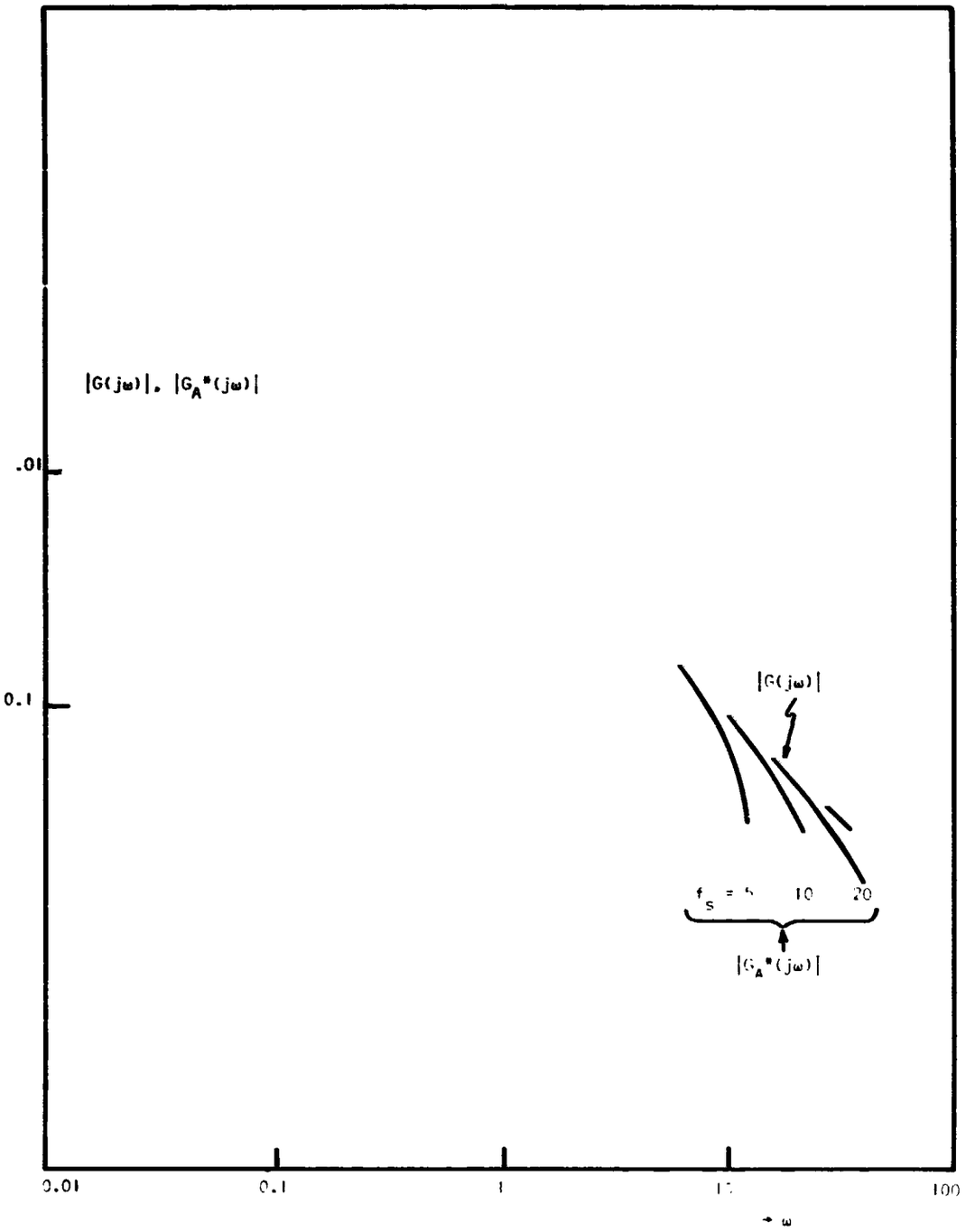


Figure 2.5 Magnitude Plot of the Continuous Transfer Function $\frac{1}{s+1}$ and its Discrete Transfer Functions for Different Sampling Frequencies where Tustin and Straight-Line Approximation Methods Were Used

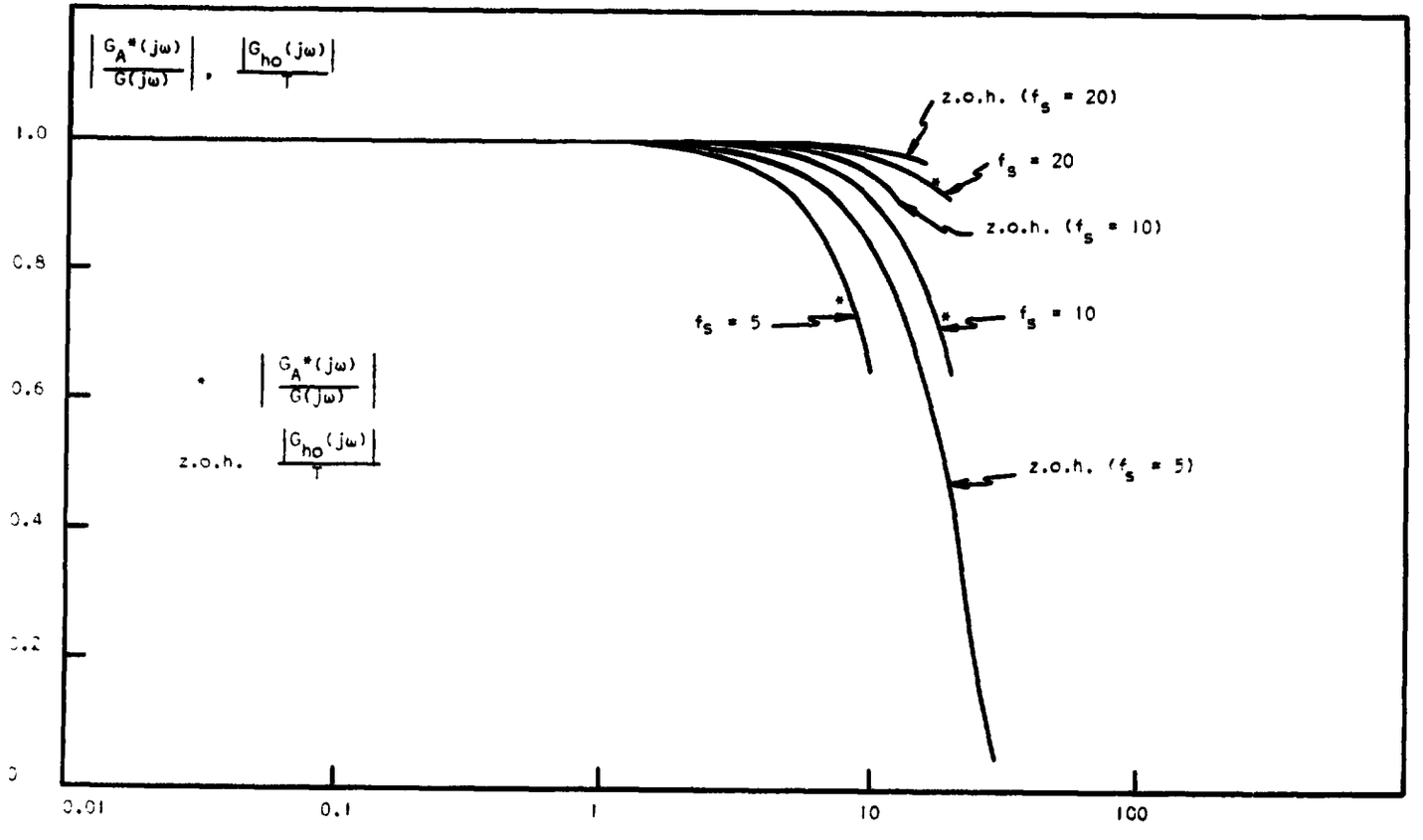


Figure 2.6 Attenuation Due to Discretization Compared with that due to Zero-Order Hold. Tustin and Straight-Line Approximation Methods are used for the Discretization

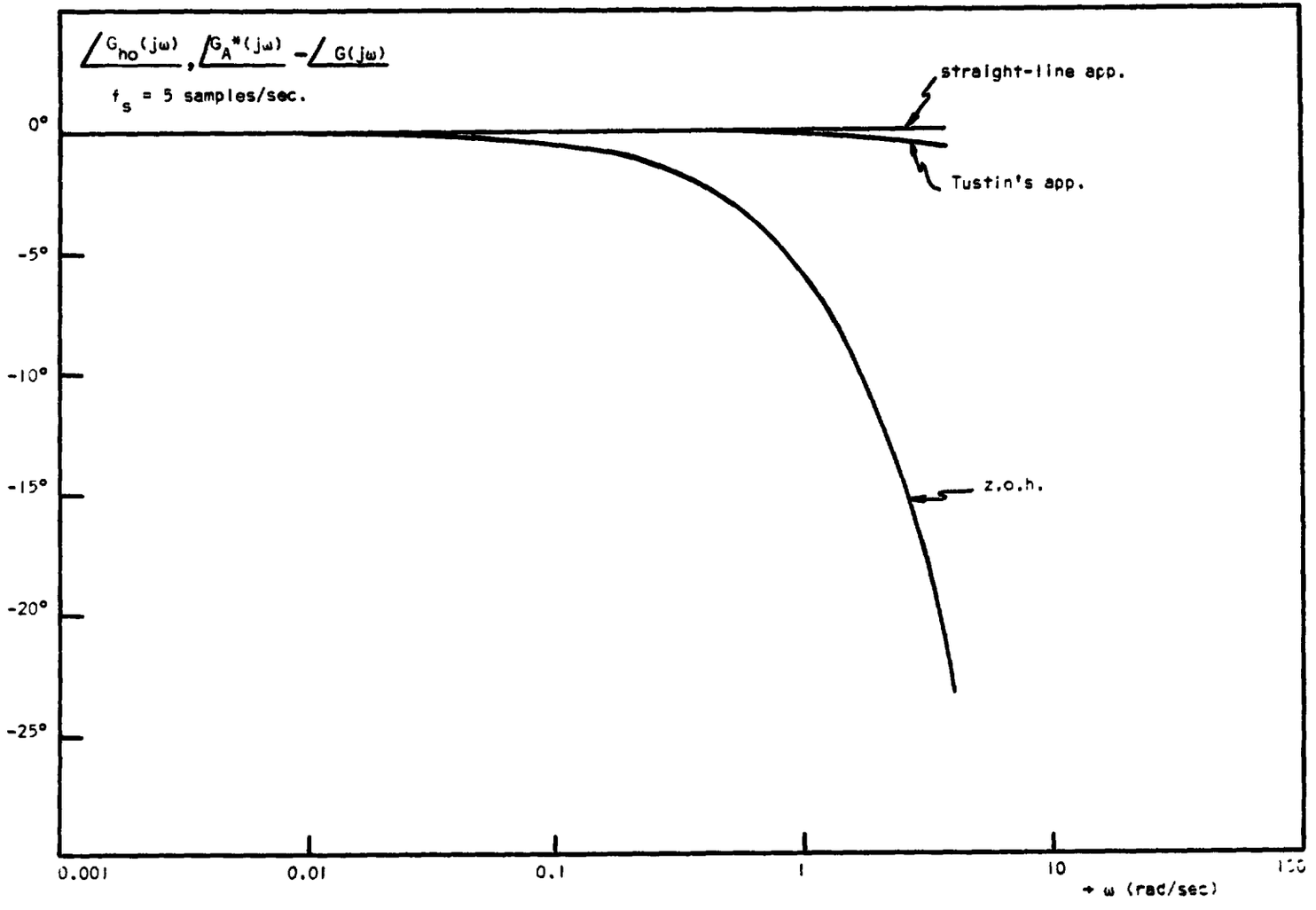


Figure 2.7 Phase Shift due to Discretization Compared with that due to Zero-Order Hold for f_s = 5 samples/sec

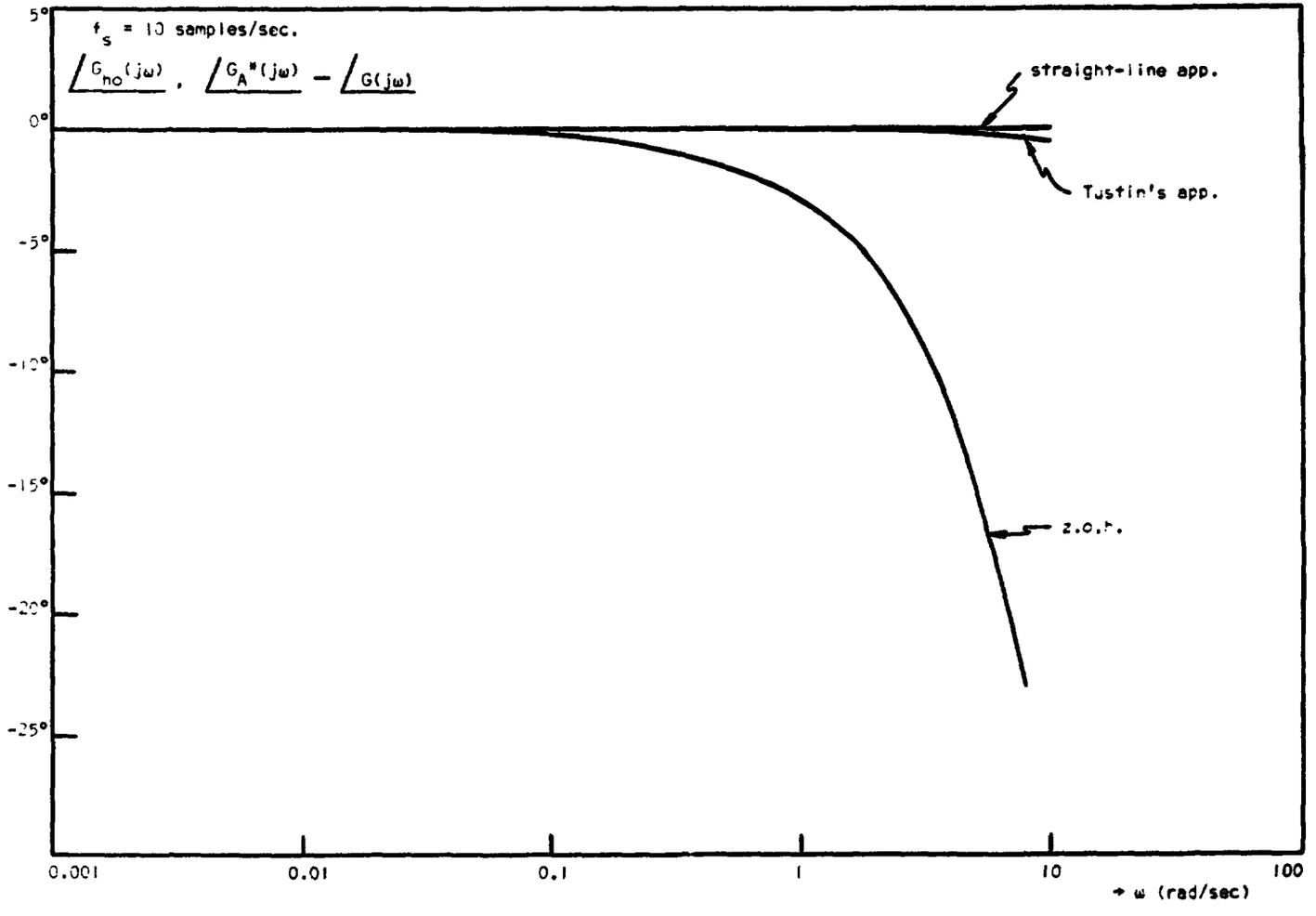


Figure 2.8 Phase Shift due to Discretization Compared with that due to Zero-Order Hold for f_s = 10 samples/sec

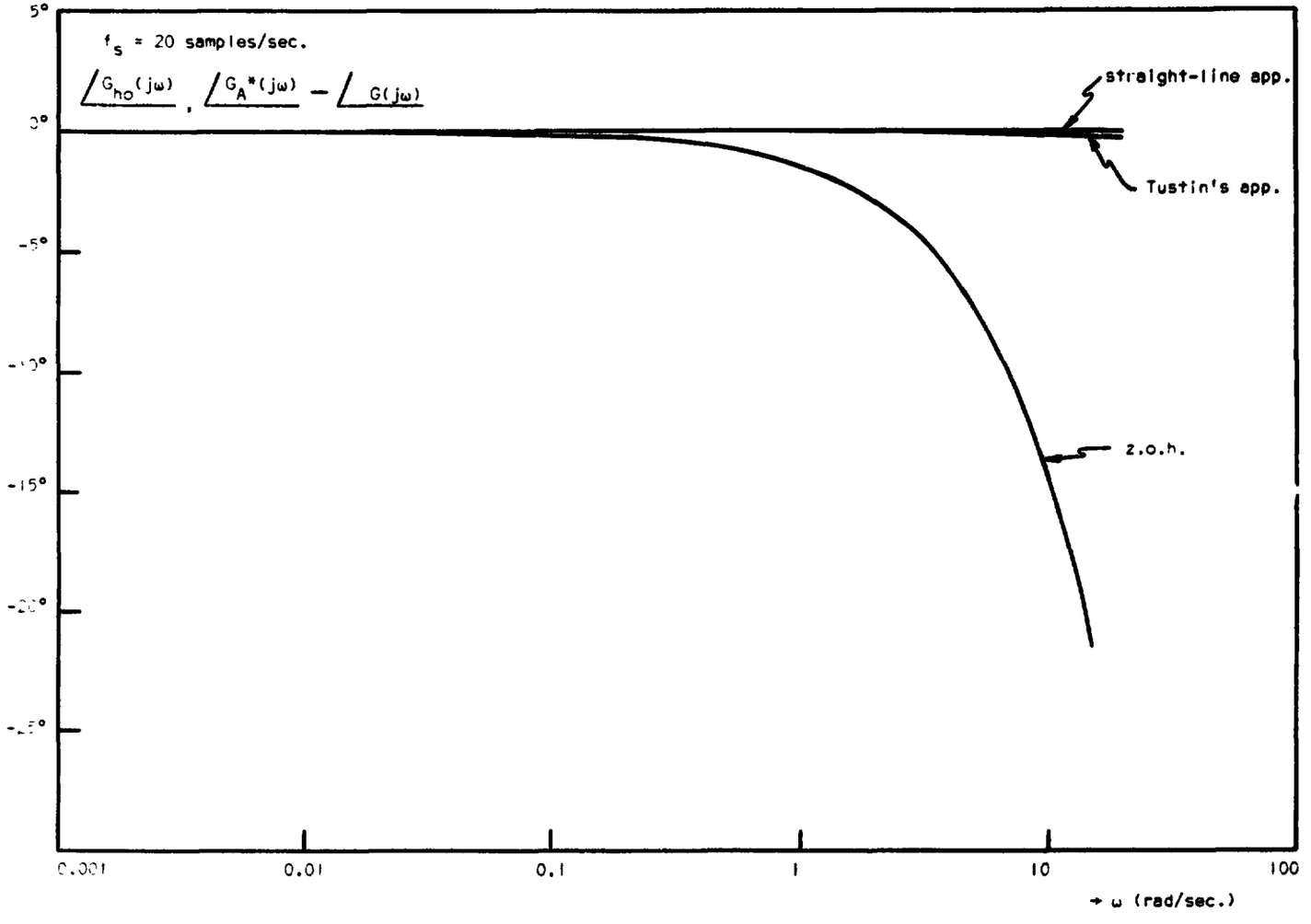


Figure 2.9 Phase Shift due to Discretization Compared with that due to Zero-Order Hold for f_s = 20 samples/sec

Consideration of Figs. (2.5-2.9) shows that in this discretization process the phase shift would be a more important determining factor than the gain. For example, for f_s = 5 samples/sec, at ω = 3 rad/sec

$$\angle G_{ho}(j\ ) = -17.3^\circ$$

$$\angle G_A^*(j\omega) - \angle G(j\omega) \cong 0^\circ \quad \text{(straight-line app.)}$$

$$\cong -0.5^\circ \quad \text{(Tustin's app.)}$$

and

$$\left|\frac{G_{ho}(j\omega)}{T}\right| \cong 0.99$$

$$|G_A^*(j\omega)/G(j\omega)| \cong 0.975$$

Thus

$$\angle G_{ho}(j\omega) + \angle G_A^*(j\omega) - \angle G(j\omega) \cong -17.3^\circ$$

and

$$\left|\frac{G_{ho}(j\omega)}{T}\right| \times \left|\frac{G_A^*(j\omega)}{G(j\omega)}\right| \cong 0.97$$

The change in phase shift by -17.3° would produce more significant effects than the change in gain by 0.97 in most practical systems. In fact, it is found that over most of the frequency range of interest the phase shift is more important than the gain.

It is important to note that in the range of useful frequencies the change in phase shift due to discretization comes from the zero order hold and one cannot choose between the above two approximation techniques on the basis of performance. A designer would be justified in using the simpler of the techniques.

It can also be demonstrated for general cases using Tustin's approximation that the value of $\angle G_A^*(j\omega) - \angle G(j\omega)$ is negligible compared to $\angle G_{ho}(j\omega)$. The demonstration proceeds as follows.

Assume a transfer function with a simple pole,

$$G(s) = \frac{1}{s + a}$$

$$\underline{/\ G(j\omega)} = -\tan^{-1}\frac{\omega}{a} \tag{2.38}$$

Then, using Tustin's method

$$G_A(z) = \frac{\frac{T}{2}\,\frac{z+1}{z-1}}{1 + a \cdot \frac{T}{2} \cdot \frac{z+1}{z-1}} = \frac{T(z+1)}{2(z-1) + aT(z+1)}$$

$$G_A^*(j\omega) = G_A(z)\Big|_{z = e^{j\omega T}} = \frac{T(e^{j\omega T} + 1)}{2(e^{j\omega T} - 1) + aT(e^{j\omega T} + 1)}$$

$$\frac{T}{2 \cdot \frac{e^{j\omega T/2} - e^{-j\omega T/2}}{e^{j\omega T/2} + e^{-j\omega T/2}} + aT} = \frac{T}{2j\tan\left(\frac{\omega T}{2}\right) + aT} \tag{2.39}$$

The phase angle of the approximation is given by

$$G_A^*(j\omega) = -\tan^{-1}\left[\frac{2}{aT}\tan\left(\frac{\omega T}{2}\right)\right] \tag{2.40}$$

Now, Eq. (2.40) should be compared wtih Eq. (2.38). Let θ, and θ_2 be defined as

$$\theta_1 = -\tan^{-1}\frac{\omega}{a} \qquad \theta_2 = -\tan^{-1}\left[\frac{2}{aT}\tan\left(\frac{\omega T}{2}\right)\right] \tag{2.41}$$

Then

$$\tan(\theta_2 - \theta_1) = \frac{\tan\theta_2 - \tan_1}{1 + \tan\theta_2\tan\theta_1}$$

$$= \frac{-\frac{2}{aT}\tan\left(\frac{\omega T}{2}\right) + \left(\frac{\omega}{a}\right)}{1 + \left(\frac{2}{aT}\tan\frac{\omega T}{2}\right)\left(\frac{\omega}{a}\right)} \tag{2.42}$$

Now show that the numerator of Eq. (2.42) is negligible. First,

$$\omega T/2 = \frac{\omega}{2} \cdot \frac{1}{f} = \frac{2\pi\omega}{2\omega_s}$$

$$= \pi\left(\frac{\omega}{\omega_s}\right)(\text{rad})$$

$$= \frac{f}{f_s} \times 180^\circ \qquad (2.43)$$

Referring to Eq. (2.37), $\omega T/2$ is the phase lag due to the zero-order hold.

If $\left(\frac{\omega T}{2}\right) \simeq 5^\circ$ (which should be done to realize a digital controller that approximates the continuous case), the tangent of the angle may be replaced by the angle so

$$\tan\frac{\omega T}{2} = \omega T/2$$

and from the numerator of Eq. (2.42)

$$-\frac{2}{aT}\tan\left(\frac{\omega T}{2}\right) + \frac{\omega}{a} \simeq -\frac{2}{aT}\frac{\omega T}{2} + \frac{\omega}{a}$$

$$= -\frac{\omega}{a} + \frac{\omega}{a} \simeq 0$$

i.e., $\theta_1 - \theta_2 \simeq 0$

or

$$\theta_1 \simeq \theta_2 \qquad (2.44)$$

For the case where the transfer function has a pair of complex poles, a similar development is possible.

Because Tustin's approximation has a cascading property, i.e.

$$Z_{Tustin}[G_1(s)G_2(s)] = \{Z_{Tustin}[G_1(s)]\}\{Z_{Tustin}[G_2(s)]\} \qquad (2.45)$$

the above approximation holds for general transfer functions.

Further investigation of gain and phase differences introduced by the discretization process for the example of the simple first order pole is summarized in Figs. 2.10 through 2.16.

Figure 2.10 shows the ratio $|G_A^*(j\omega)/G(j\omega)|$ for several values of system time constant as a function of the ratio of frequency to sampling frequency. Tustin's method was used for the approximation. The normalized magnitude of the zero-order hold is also shown for comparison.

Figure 2.11 is a plot of the phase for the same conditions. This figure reiterates the conclusion that the phase lag introduced by the zero-order hold predominates.

Figures 2.12 and 2.13 differ from Figs. 2.10 and 2.11 only in that the linear segment input approximation was used for the system. Again the phase lag of the zero-order hold predominates.

Figures 2.14 through 2.15 compare the gain and phase shift of several discretization methods for two sampling frequencies. These figures show that one of the simpler methods (Tustin) is also one of the more accurate methods considered for this application.

Selection of the Simulation Increment (Computer Sample Period)

One of the most important criteria in the choice of a particular digital simulation method is the length of the simulation increment required to produce a given accuracy. It has been well documented in the literature and in our efforts that some methods require a smaller simulation increment than others to produce equivalent results, and , in fact, some methods are unstable for increments that produce good results in other methods. Therefore the choice of the smallest simulation increment to produce a given accuracy has been the subject of recent investigation. This topic is of particular importance when the simulation must be performed in real time and on the smallest possible machine.

The structure of the problem of finding the smallest possible simulation increment which will yield a specified accuracy for a given simulation method suggests the possibility of using classic optimization theory. This

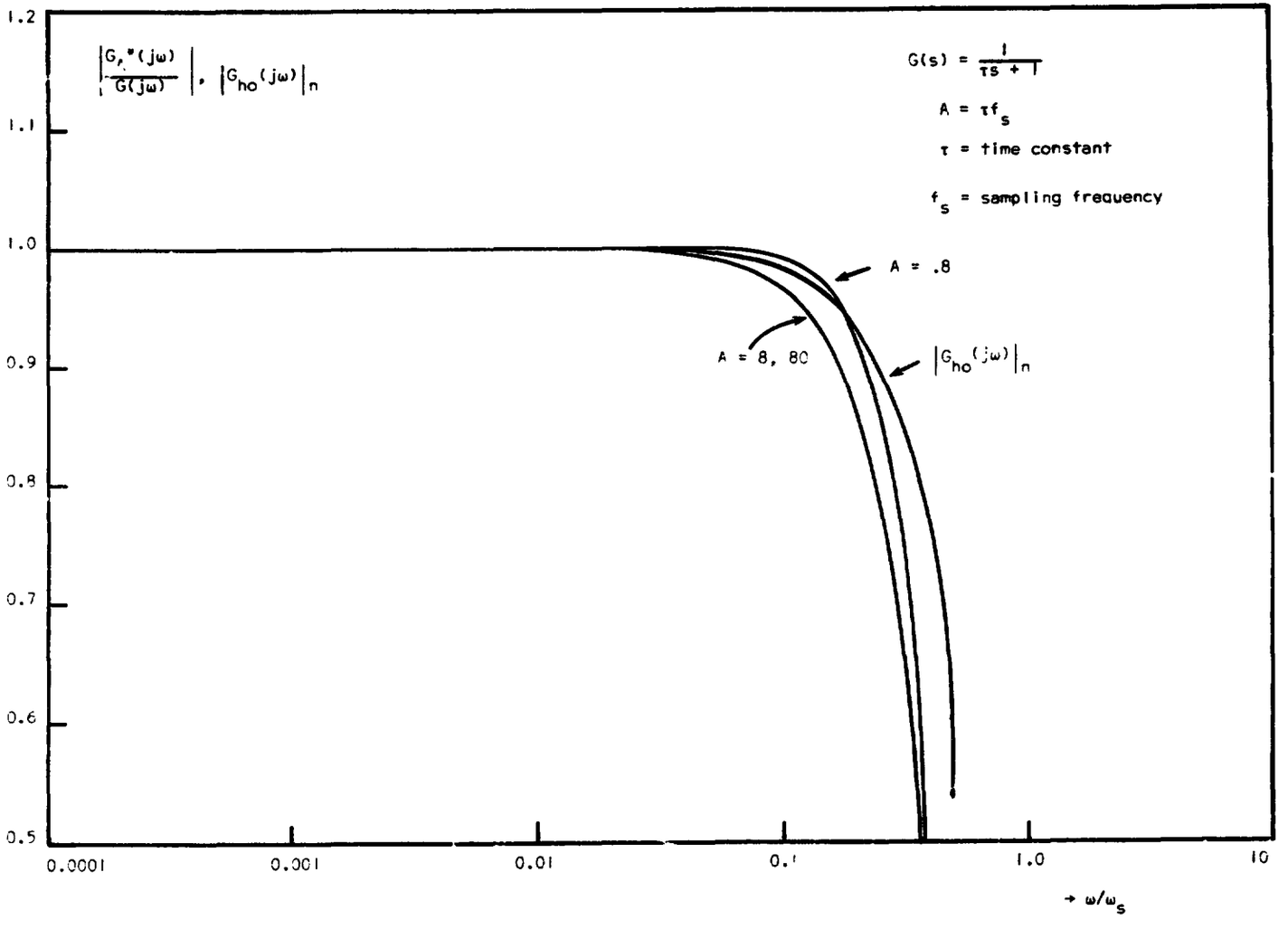


Figure 2.10 Attenuation due to Discretization by Tustin's Method as a Function of the Ratio of Frequency to Sampling Frequency with the Product of System Time Constant and Sampling Frequency as Parameter

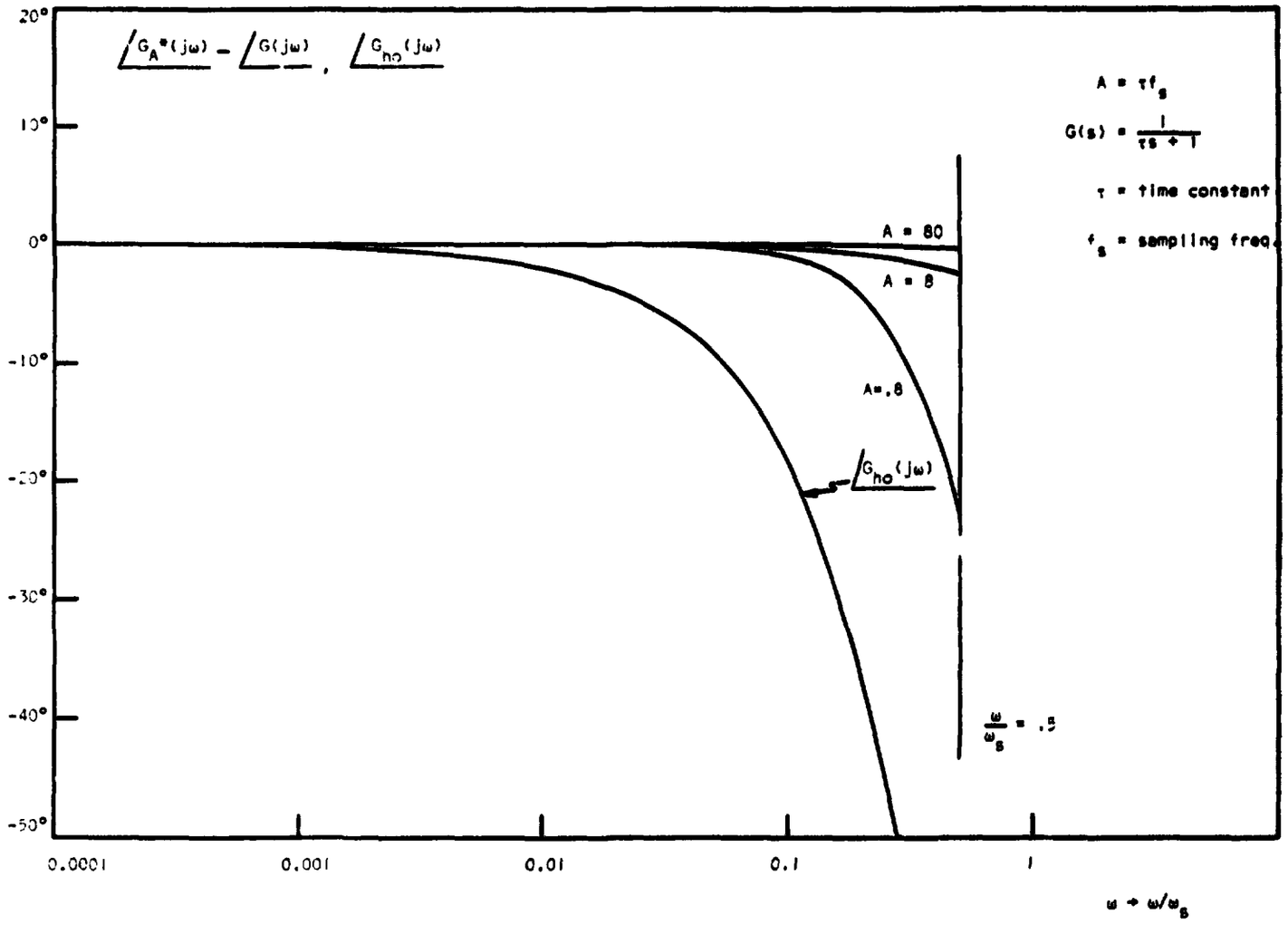


Figure 2.11 Phase Difference for the Same Conditions as Figure 2.10

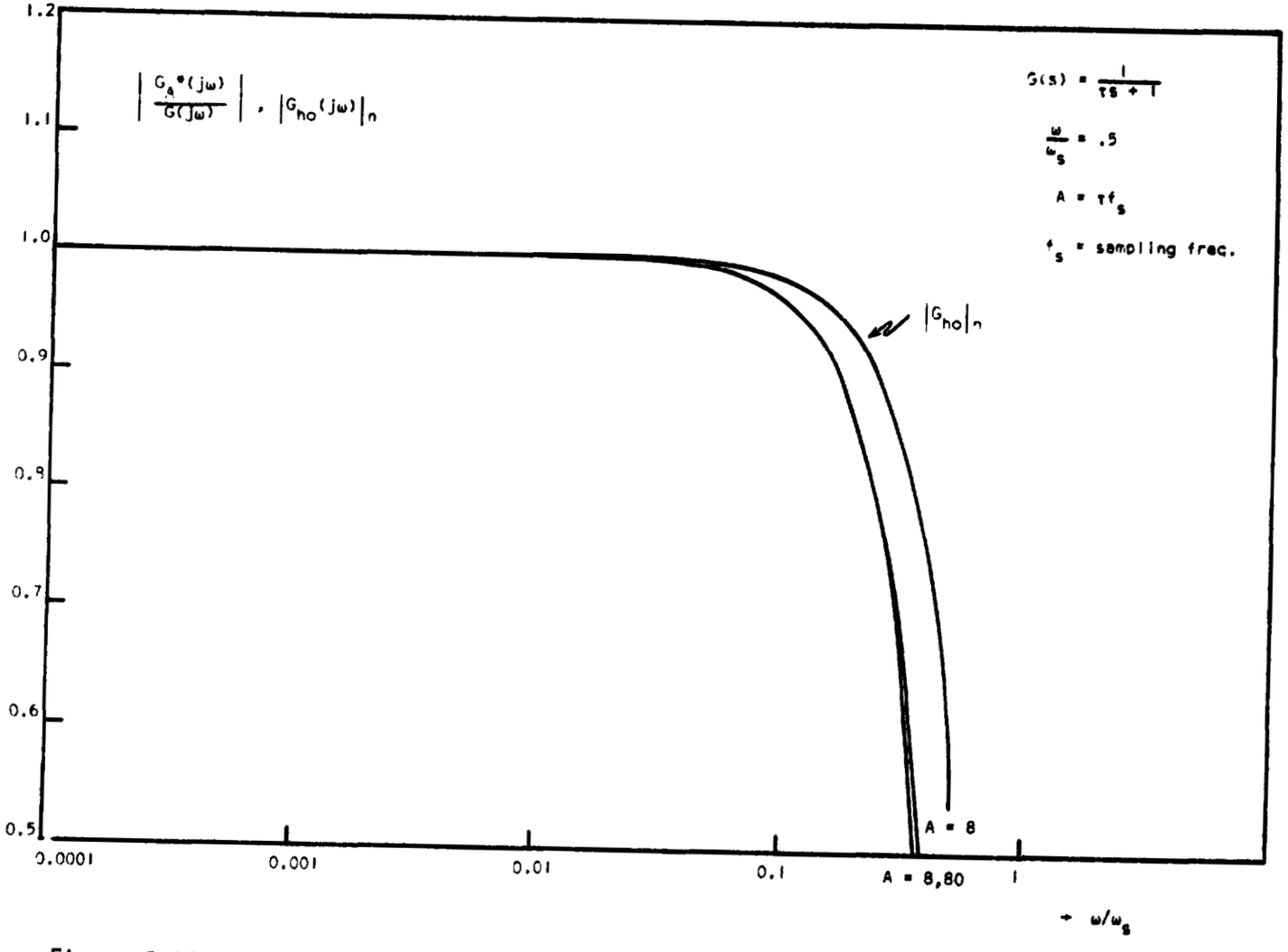


Figure 2.12 Attenuation due to Discretization by Linear Segment Input Approximation for Same Conditions as Figure 2.10

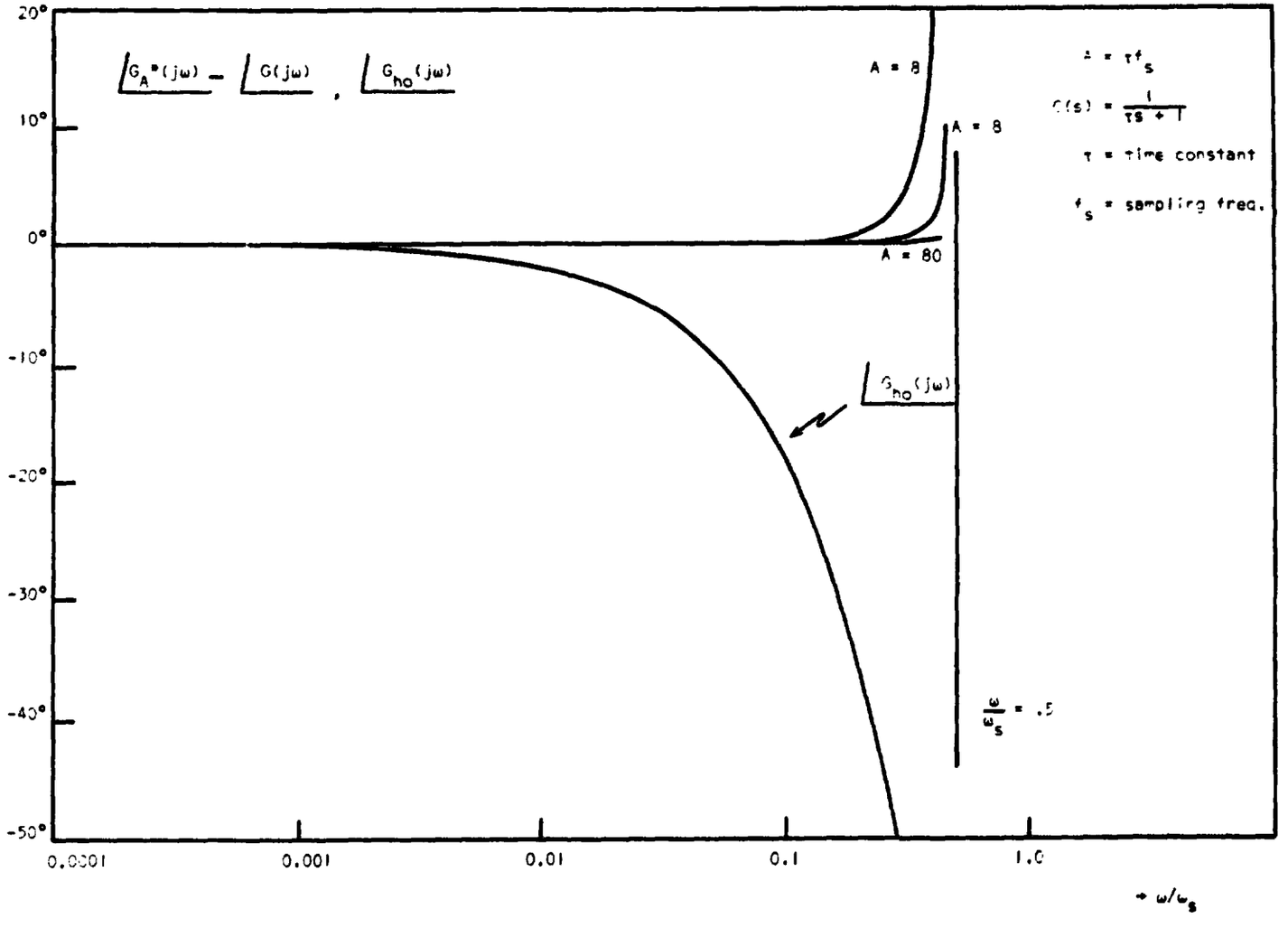


Figure 2.13 Phase Difference for Same Conditions as Figure 2.12

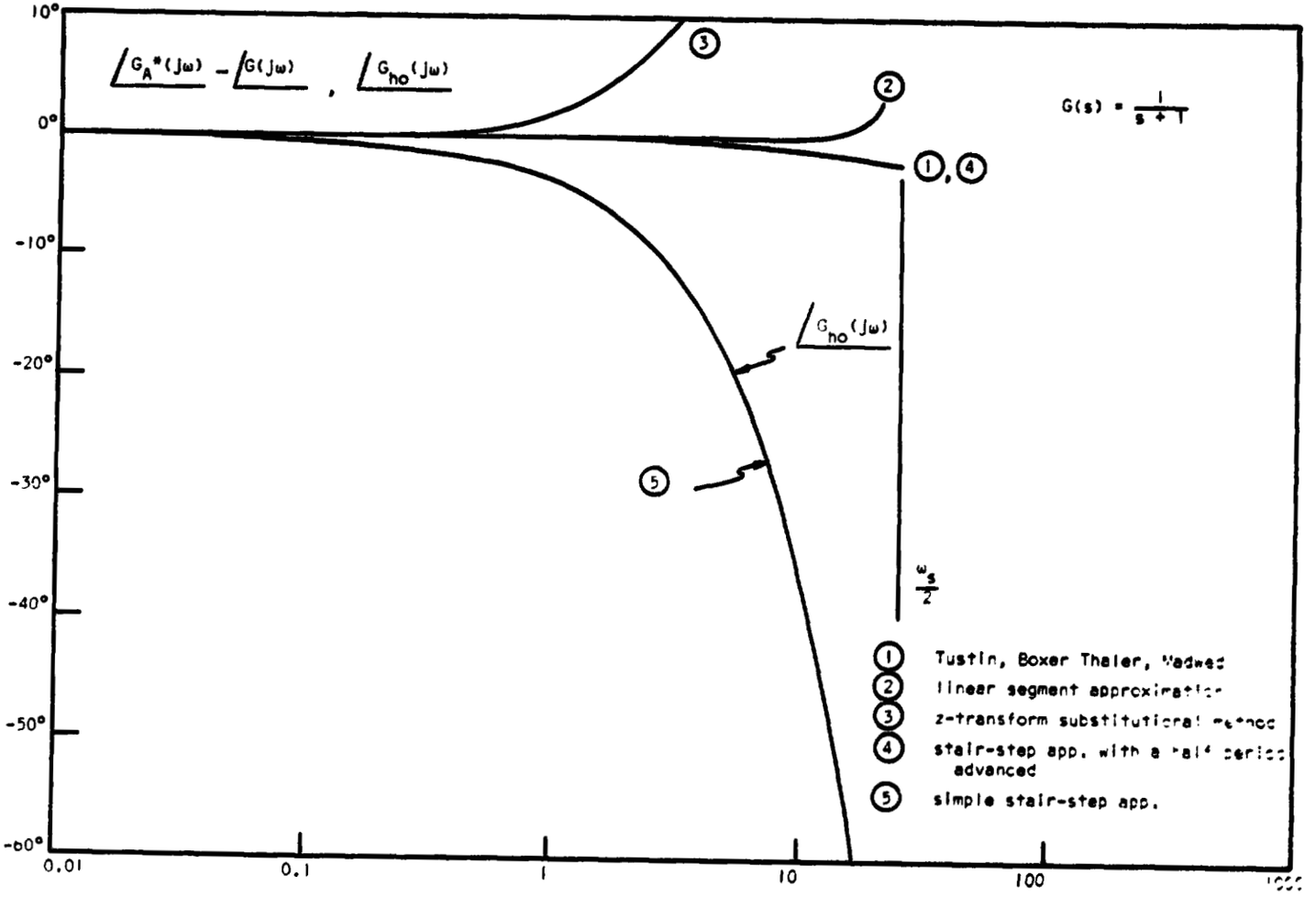


Figure 2.14 Phase difference due to Discretization for Several Methods (Sampling Frequency of 5 Hz)

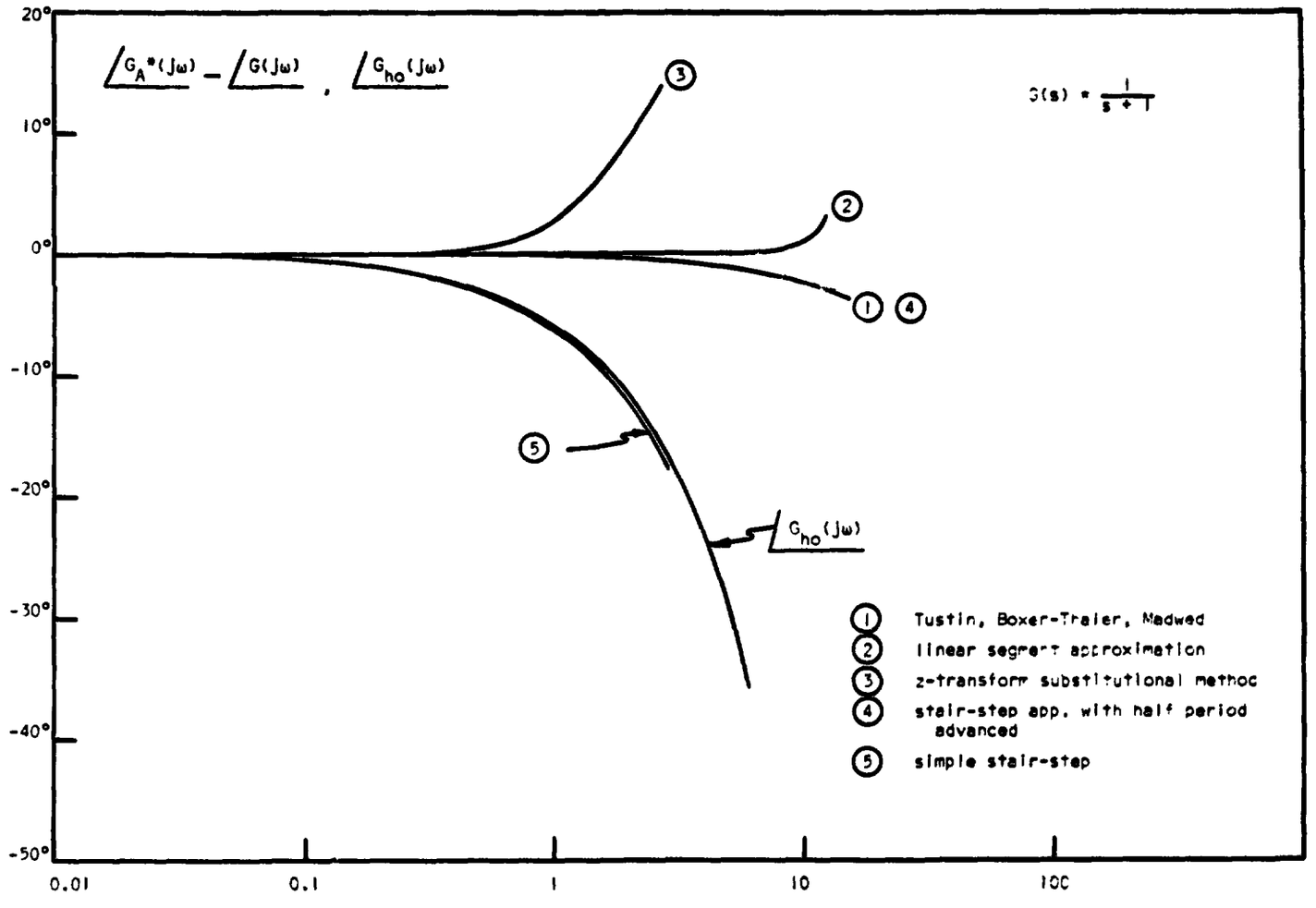


Figure 2.15 Phase Difference due to Discretization by Several Methods (Sampling Frequency of 8 Hz)

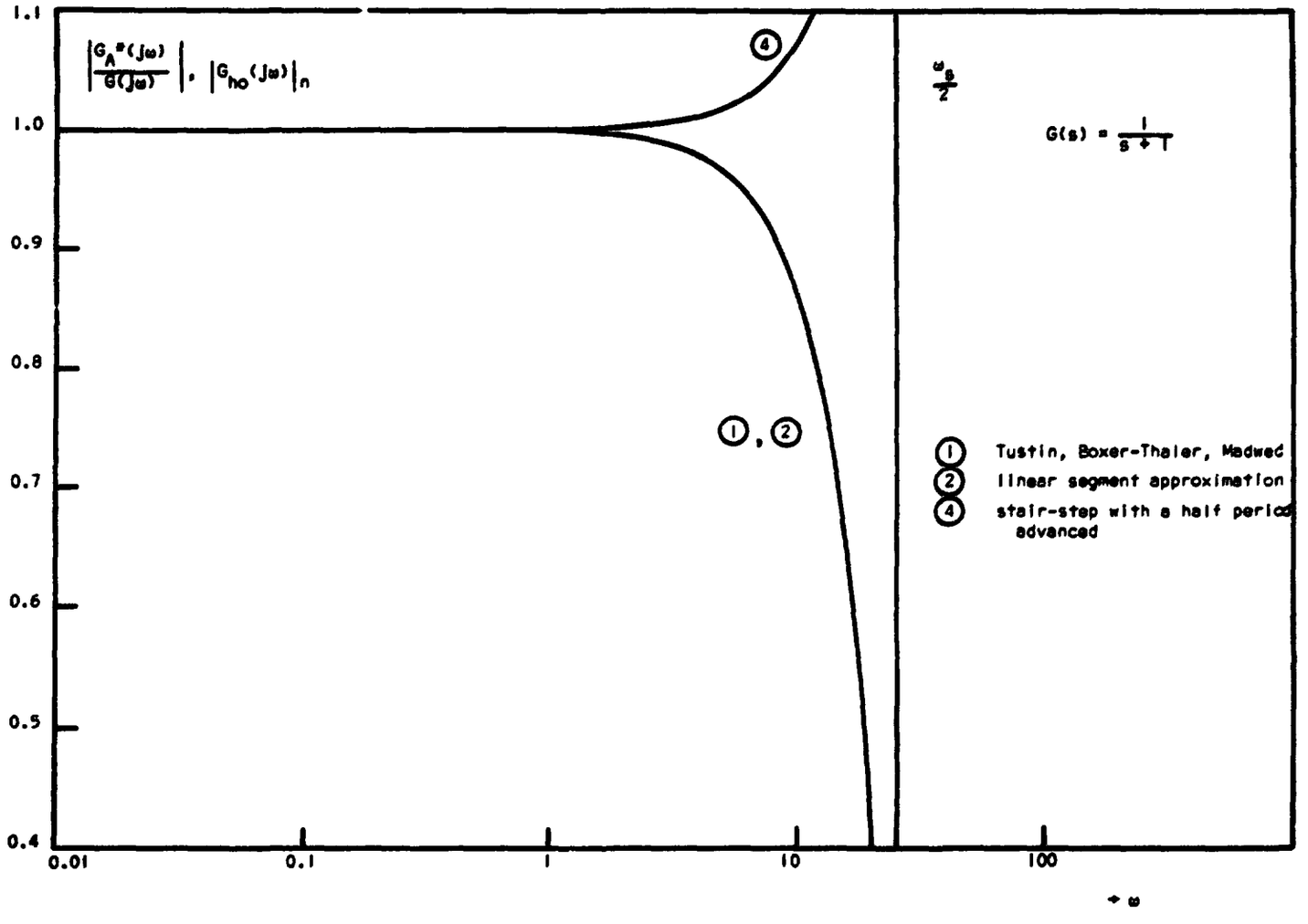


Figure 2.16 Attenuation due to Discretization by Several Methods (Sampling Frequency of 8 Hz)

was the approach taken by Rosko and Durling [2.2] and Rosko [2.3]. Their results are currently being implemented to aid in the choice of simulation method and for possible inclusion in the Simulation Design Package.

Rosko and Durling's method is based upon the assumption that round-off errors have been or can be made negligible for the time increment to be determined. The problem then is to find a compromise between computation time and simulation error. The method will now be summarized.

Formulate a performance index which will mathematically reflect the parameters to be minimzed and their relative importance. Functionally,

$$J = f(E,T, \ldots) \tag{2.46}$$

where E is a measure of the simulation error and T is the simulation increment to be determined. An example of such a performance measure is

$$J = E + \frac{\beta}{T} \tag{2.47}$$

where β is a constant weight.

For deterministic inputs the integral squared error for a system simulated with interpolated intervals is given by [2.2, 2.3]

$$E(t) = e^2(t) = \int_0^\infty [y_i(t) - y_A(t)]^2 dt \tag{2.48}$$

where $y_i(t)$ and $y_A(t)$ are the outputs of the ideal and approximated systems respectively. Using Parseval's theorem we may express the integral squared error as

$$E(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_{ee}(\omega) d\omega \tag{2.49}$$

where $\Phi_{ee}(\omega)$ is the spectral density of the error.

For the case of discrete simulation where the error is considered only at the sampling intervals, the summation of squared error is given by [2.2, 2.3]

$$E(nT) = \sum_{n=0}^{\infty} e^2(nT) = \frac{1}{2\pi j} \oint_\Gamma E(z)E(z^{-1})z^{-1} dz \tag{2.50}$$

which may be evaluated by summing the residues of the poles of the integrand inside the unit circle in the z-plane.

Select a reasonable increment value, $T = T_1$, and calculate the initial value of the error using the appropriate error criterion. This gives an initial value of the performance index.

$$J_1 = E_1 + \frac{\beta}{T_1} . \tag{2.51}$$

Calculate an estimated value of the error and performance index by

$$\hat{E}_2 = \Psi(T) = \frac{T}{T_1} E_1 \Big|_{T=T_2} \tag{2.52}$$

$$\hat{J}_2 = \hat{E}_2 + \frac{\beta}{T} \Big|_{T=T_2} \tag{2.53}$$

A necessary condition for optimality is

$$\hat{J}_2' = \frac{\partial \hat{J}_2}{\partial T} = \frac{\partial}{\partial T} \left(\frac{T}{T_1} E_1 + \frac{\beta}{T} \right) \Big|_{T=T_2} = 0 \tag{2.54}$$

which yields

$$T_2 = +\left(\frac{\beta T_1}{E_1} \right)^{\frac{1}{2}} \tag{2.55}$$

Use $T = T_2$ to predict E using the appropriate error formulation. Then calculate

$$J_2 = E_2 + \frac{\beta}{T_2} \tag{2.56}$$

subject to the constraint

$$\frac{\Delta J_2}{J_2} = \left| \frac{J_2 - J_1}{J_2} \right| \leq \Gamma_0 \tag{2.57}$$

where Γ_0 is a constant to be chosen.

For all succeeding steps, $m \geq 3$, the following procedure applies:

1) $$\hat{E}_m = \Psi(T) = \frac{\sum_{i=1}^{m} \Pi_m(T) E_i}{(T-T_i)\Pi_m'(T)} \qquad (2.58)$$

where Lagrangian interpolation has been used, and

$$\Pi_m = T(T-T_1)(T-T_2) \cdots (T-T_m) \qquad (2.59)$$

and

$$\Pi_m' = \frac{d}{dT} \Pi_m(T). \qquad (2.60)$$

2) Now calculate

$$\hat{J}_m = \hat{E}_m + \frac{\beta}{T_k}\Big|_{min} \qquad (2.61)$$

3) Calculate $T_m = T_{m,k}$ using Newton-Raphson quasilinearization where

$$T_{m,1} = T_{m-1} - \frac{\hat{J}'(T_{m-1})}{\hat{J}''(T_{m-1})} \quad \text{for } k = 1 \qquad (2.62)$$

and

$$T_{m,k} = T_{m,k-1} - \frac{\hat{J}'(T_{m,k-1})}{\hat{J}''(T_{m,k-1})} \quad \text{for } k > 1 \qquad (2.63)$$

subject to the constraint

$$\left| \frac{T_{m,k} - T_{m,k-1}}{T_{m,k}} \right| \leq \tau_0 \qquad (2.64)$$

where τ_0 is a chosen constant.

4) Calculate the predicted value E_m using $T = T_m$ and the appropriate error formulation.

5) Calculate the performance index

$$J_m = E_m + \frac{\beta}{T_m} \qquad (2.65)$$

subject to

$$\frac{\Delta J_m}{J_m} = \left| \frac{J_m - J_{m-1}}{J_m} \right| \leq \Gamma_0 \qquad (2.66)$$

The calculation terminates whenever Eq. (2.66) is satisfied.

It has been shown that the discretization method is unimportant from a performance view based on phase shift and gain considerations. This conclusion will also be checked using the procedure presented above.

Conclusions and Work in Progress

The performance of a continuous autopilot which is implemented digitally is affected so much by the zero order hold that the discretization method is a secondary consideration. Tustin is a relatively simple method that is satisfactory. Computer speed (sampling rate) may be established on the basis of the phase shift a designer will allow to be introduced by the zero order hold. Phase lag decreases as sample speed increases. Gain constants should be maintained to keep steady state errors constant. The relative stability of the aircraft will be decreased by the digital implementation so the value of computer sampling time should be based on the decrease in phase margins a designer is willing to allow. This value can be determined through sensitivity studies. Work is beginning on an example sensitivity simulation to demonstrate how one can approach the problem of specifying an allowable change in phase margins. Optimum simulation increment work is also underway to give credance to the above conclusions from another point of view. Round off error work is to be done during the next year. Further studies of discretization methods considered in this section will be undertaken for more complex (higher order) systems with a view toward possibly utilizing phase lead introduced by some of the methods to offset phase lag introduced by the zero-order hold.

REFERENCES

[2.1] Gibson, J. E., McVey, E. S., et al., "A Set of Standard Specification for Linear Automatic Control Systems," Trans. AIEE 80, Part 2, 1961.

[2.2] Rosko, J. S. and A. E. Durling, "Optimal Simulation of Linear Systems," Proc. Nat. Elec. Conf., 23 (1967), pp. 170-174.

[2.3] Rosko, J. S., Digital Simulation of Physical Systems, Addison-Wesley, Reading, Mass. 1972.

III. FREQUENCY DOMAIN SYNTHESIS OF DISCRETE REPRESENTATIONS

Analysis of linear, continuous time systems is frequently done in the frequency domain. By use of the Laplace Transform, differential equations are replaced by algebraic equations which, in general, are simpler to solve. In designing a discrete time system to approximate the performance of a particular continuous time system, it would be helpful if some of the familiar analysis techniques could be utilized.

A design procedure is presented by which a discrete time transfer function can be developed in the frequency domain. The resulting system will have frequency domain characteristics similar to the continuous time system from D.C. to one half the sampling frequency. It will also be shown that the time domain performance of the two systems will be similar. The characteristics of the type of continuous time system which can be most closely approximated using this design method will also be discussed.

A simple design example will be presented to explain the methodology of the design procedure. Following that, the frequency domain design and time domain evaluation for the autopilot will be discussed.

Explanation of Design Procedure

The transfer function for this example is shown in Figure 3.1

$$F(s) = \frac{2}{s + 2} \qquad (3.1)$$

and the sampling frequency is chosen (arbitrarily) to be 20 rad/sec. The design objective is to develop a discrete time transfer function F(z), whose frequency domain characteristics will closely approximate those of F(s) from 0 to 10 rad/sec. A graphical version of the design procedure will be presented to illustrate the method. Part or all of the procedure can be automated.

The first step is to plot the magnitude of the transfer function versus frequency, i.e., 20 log $F(j\omega)$ versus log ω. The plot need not extend higher in frequency than one half the sampling frequency and may go as low in frequency as desired.

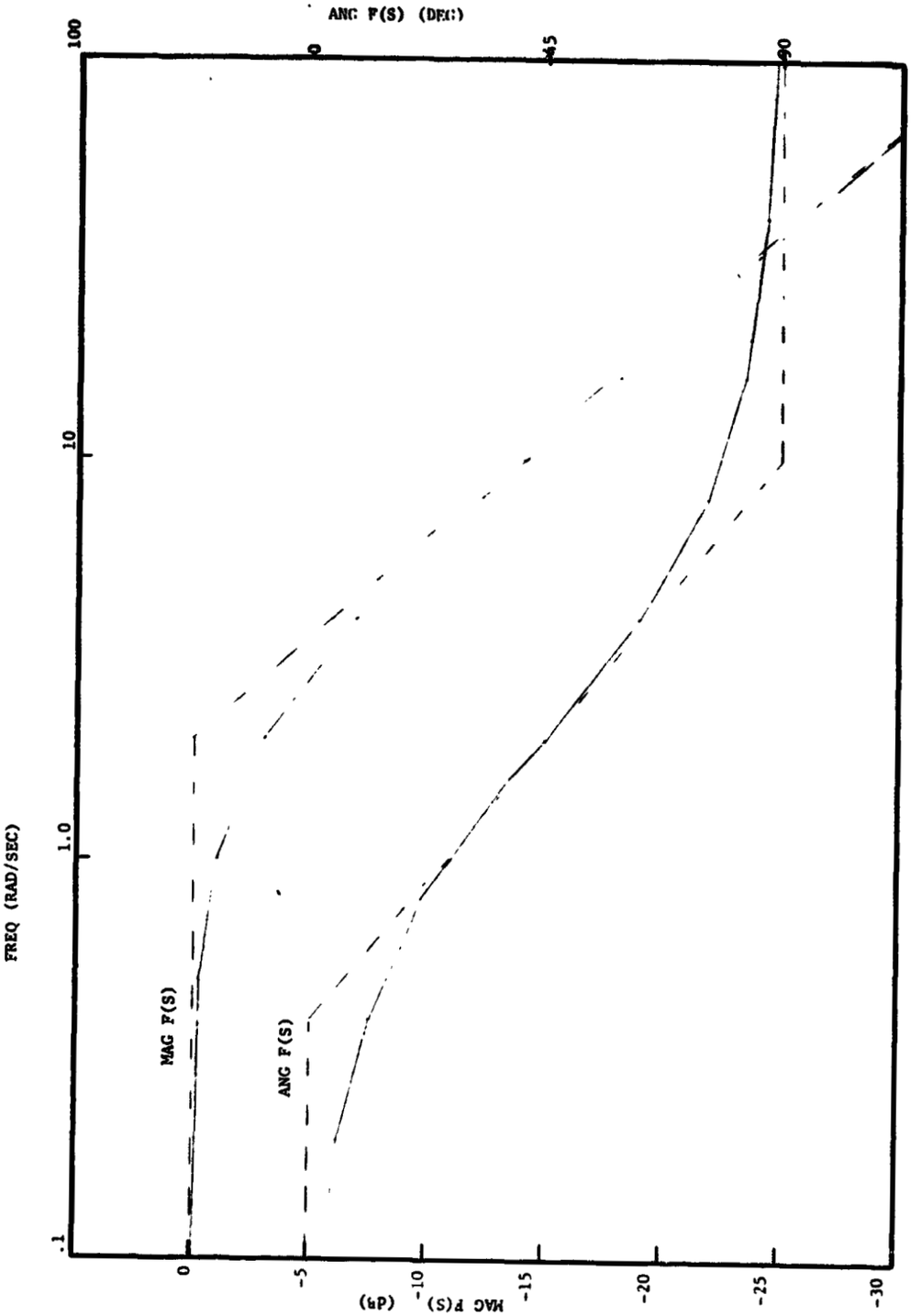


Figure 3.1 Bode Plot for Sample Problem

The next step is to define a complex variable p such that

$$p = j\lambda = \frac{Z - 1}{Z + 1} \tag{3.2}$$

From Z-transform theory

$$z = \exp(sT)$$

and for real frequencies this becomes

$$z = \exp(j\omega T) \tag{3.3}$$

where $T = 2\pi/\omega_s$ is the sampling period, and ω_s is the sampling frequency in radians per second. Substituting Eq. (3.3) into Eq. (3.2) yields

$$p = j\lambda = (\exp(j\omega T) - 1)/(\exp(j\omega T) + 1) \tag{3.4}$$

By factoring and utilizing trigonometric identities, Eq. (3.4) can be expressed as

$$p = j\lambda = j\sin(\omega T/2)/\cos(\omega T/2) = j\tan(\omega T/2) \tag{3.5}$$

The magnitude plot of Eq. (3.1) is now re-plotted versus the frequency λ such that the following relation is true (see Figure 3.2):

$$F(j\lambda)\Big|_{\lambda = \lambda_i = \tan\omega_i T/2} = F(j\omega T) \tag{3.6}$$

In other words the magnitude of $F(j\omega)$ at a particular frequency ω_i is mapped into a point on the $F(j\lambda)$ plot, having the same magnitude and occurring at a frequency λ_i, where

$$\lambda_i = \tan(\omega_i T/2) \tag{3.7}$$

When $\omega = 0$, $\lambda = \tan(0) = 0$, and when $\omega = \omega_s/2$, $\lambda = \tan(\pi/2) = \infty$. Thus, the frequency range $0 \leq \omega \leq \omega_s/2$ maps into the frequency range $0 \leq \lambda \leq \infty$.

The next step is to synthesize an equation for F(p) by inspecting the magnitude plot just made. The plot of $F(j\lambda)$ versus λ will look similar to the plot of $F(j\omega)$ versus ω at low frequencies. As λ increases, $F(j\lambda)$ will approach the value that $F(j\omega)$ has at $\omega_s/2$ along a horizontal asymptote.

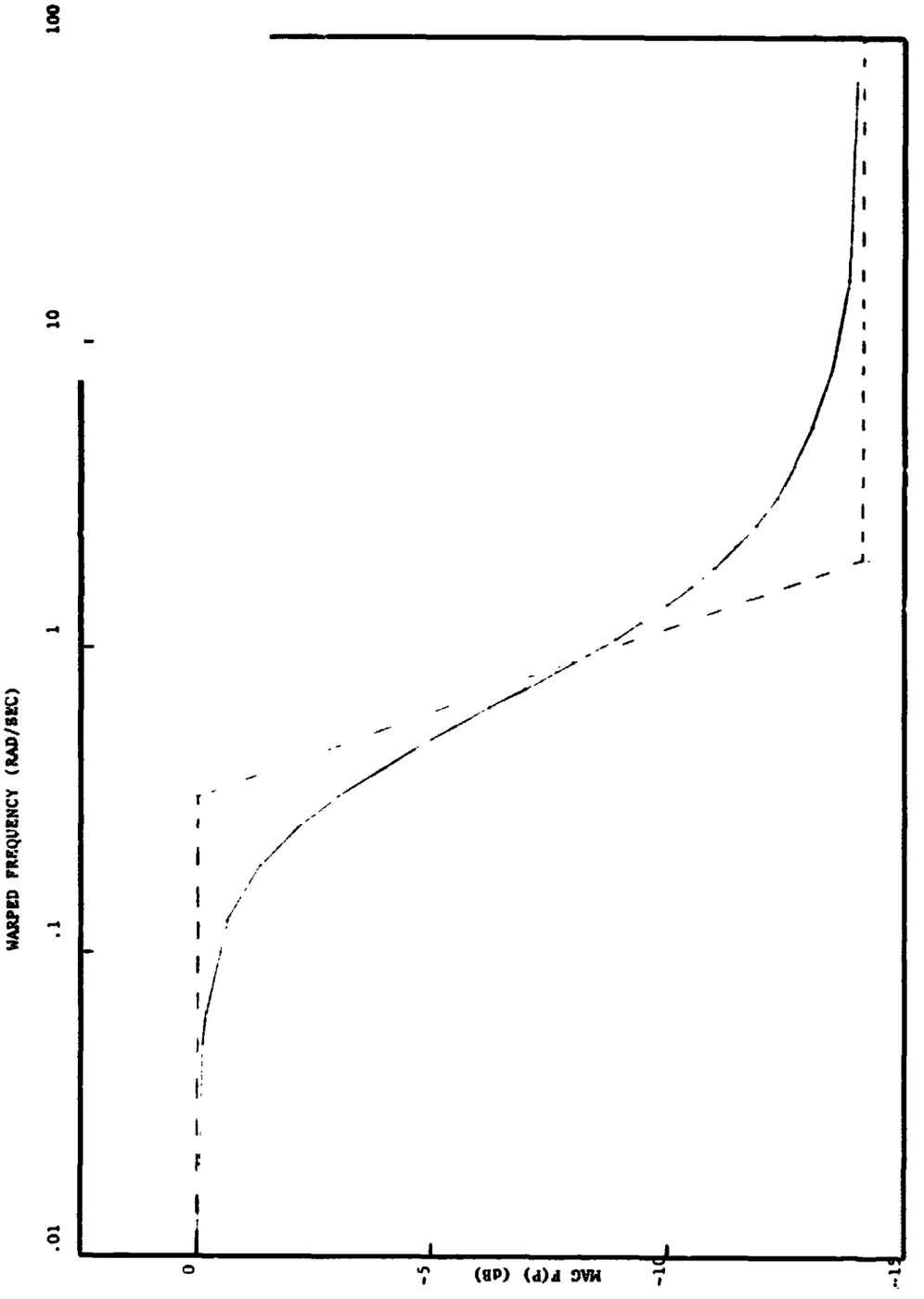


Figure 5.2 Magnitude Plot for Sample Problem Warped Frequency Domain

By applying the relation that the actual magnitude curve differs from the asymptotic curve by 3 dB at the break frequency, F(p) can be seen to have a lag break frequency at $\lambda_1 = .325$ and a lead break at $\lambda_2 = 1.91$. Therefore, F(p) has the form $(p + \lambda_2)/(p + \lambda_1)$. Normalizing to unity gain at DC, we have

$$F(p) = \left(\frac{\lambda_1}{\lambda_2} \right) \frac{(p + \lambda_2)}{(p + \lambda_1)} \tag{3.8}$$

It can be noted that the break frequency $\lambda_1 = .325$ corresponds to $\omega = 2$ rad/sec., the break frequency of the original F(s).

The next step is to make the substitution

$$p = \frac{z - 1}{z + 1}$$

into Eq. (3.8). Simplifying, this results in the expression:

$$F(z) = \frac{\lambda_1}{\lambda_2} \cdot \frac{(\lambda_2 + 1)z + (\lambda_2 - 1)}{(\lambda_1 + 1)z + (\lambda_1 - 1)} \tag{3.9}$$

Equation (3.9) can be written in the form

$$F(z) = K_1 \frac{z + K_2}{z + K_3} \tag{3.10}$$

where $K_1 = (\lambda_1/\lambda_2)(\lambda_2 + 1)/(\lambda_1 + 1)$, $K_2 = (\lambda_2 - 1)/(\lambda_2 + 1)$, and $K_3 = (\lambda_1 - 1)/(\lambda_1 + 1)$.

The frequency response of F(z) can be determined by setting $Z = \exp(j\omega T) = \cos(\omega T) + j\sin(\omega T)$. This results in the following two expressions:

$$\text{mag}[F(z)] = 20 \log\{K_1[(\cos\omega T + K_2)^2 + (\sin\omega T)^2]^{\frac{1}{2}}/[(\cos\omega T + K_3)^2 + (\sin\omega T)^2]^{\frac{1}{2}}\}$$

$$\text{ang}[F(z)] = \tan^{-1}[\sin\omega T/(\cos\omega T + K_2)] - \tan^{-1}[\sin\omega T/(\cos\omega T + K_3)]$$

Since $z = \exp(j\omega T)$, these two functions can be plotted versus the original frequency ω. The magnitude plot of F(z) should be quite close to the magnitude plot of F(s) from DC to $\omega_s/2$. Since the numerator and denominator of F(z) are of the same order in z, the phase shift will be zero degrees at $\omega_s/2$. To try and shape the phase shift of F(z) to more closely

match F(s), an all-pass filter can be added to F(p). This results in the new transfer function

$$F(p) = \frac{\lambda_1}{\lambda_2}\left(\frac{p + \lambda_2}{p + \lambda_1}\right)\left(\frac{\lambda_3 - p}{\lambda_3 + p}\right) \tag{3.11}$$

making the substitution $p = (z - 1)/(z + 1)$ into Eq. (3.11) gives the following:

$$F(z) = K_1 \frac{(z + K_2)(z + K_3)}{(z + K_4)(z + K_5)} \tag{3.12}$$

where

$$K_1 = (\lambda_1/\lambda_2)(\lambda_2 + 1)(\lambda_3 - 1)/(\lambda_1 + 1)(\lambda_3 + 1),$$

$$K_2 = (\lambda_2 - 1)/(\lambda_2 + 1),$$

$$K_3 = (\lambda_3 + 1)/(\lambda_3 - 1),$$

$$K_4 = (\lambda_1 - 1)/(\lambda_1 + 1),$$

and

$$K_5 = (\lambda_3 - 1)/(\lambda_3 + 1) = 1/K .$$

The phase of this new F(z) will go to -180 degrees at $\omega = \omega_s/2$. The actual phase characteristics of the original F(s) will determine how high in frequency the phase of F(z) matches that of F(s). Figure 3.3 shows the magnitude of F(z), as well as the phase with and without an all-pass filter.

To evaluate the time domain performance, F(z) can be expressed as a ratio of polynomials

$$F(z) = K_1[z^2 + (K_2 + K_3)z + K_2K_3]/[z^2 + (K_4 + K_5)z + K_4K_5] \tag{3.13}$$

Multiplying the numerator and denominator of Eq. (3.13) by z^{-2} yields

$$F(z) = \frac{C(z)}{R(z)} = K_1[1 + (K_2 + K_3)z^{-1} + K_2K_3z^{-2}]/[1 + (K_4 + K_5)z^{-1} + K_4K_5z^{-2} \tag{3.14}$$

where C(z) and R(z) represent the output and input of the transfer function, respectively. Cross multiplying Eq. (3.14) produces the following expression.

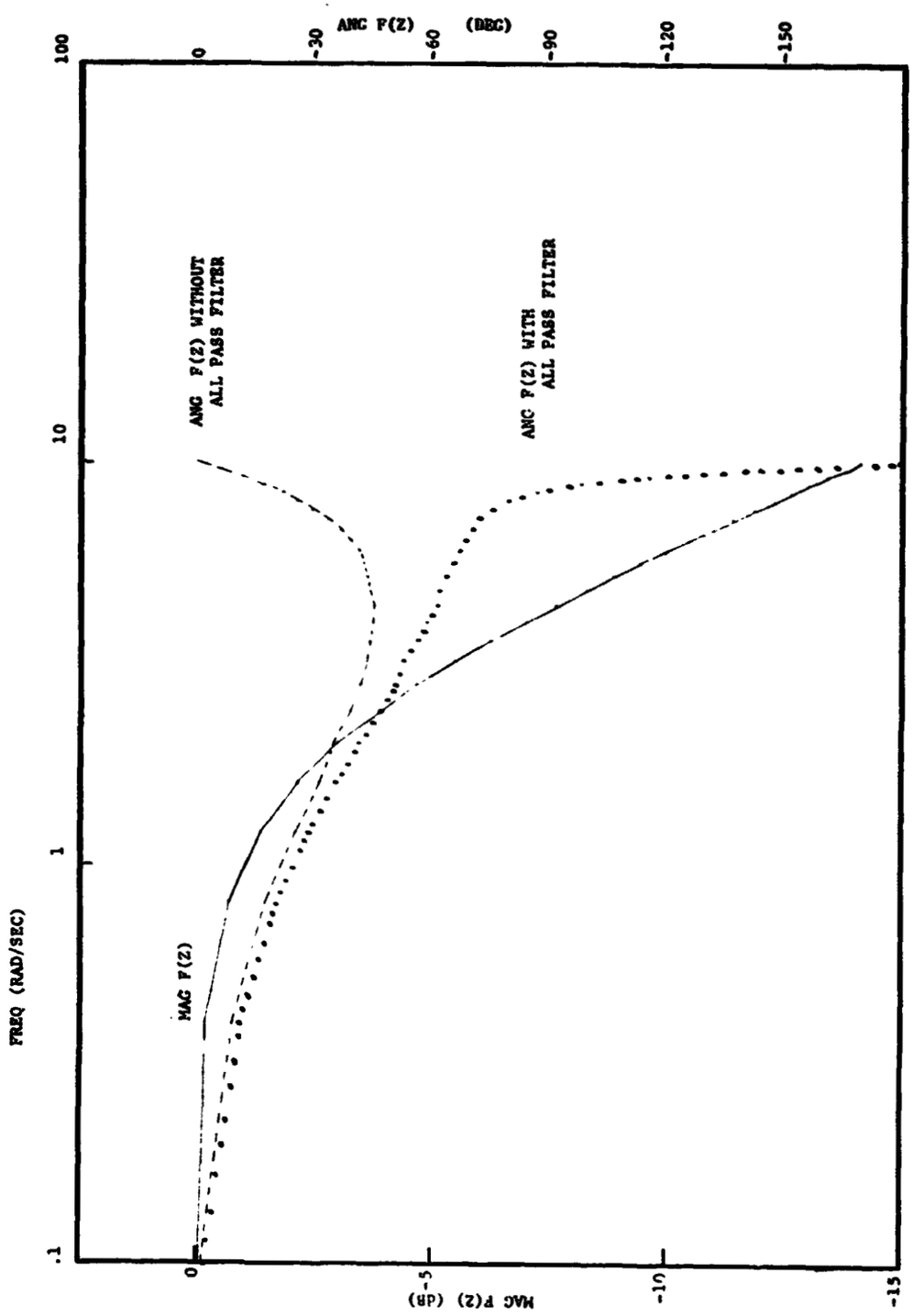


Figure 3.3 Magnitude and Phase for Discrete Approximation with and without All-Pass Filter

$$[1 + (K_4 + K_5)z^{-1} + K_4K_5z^{-2}]C(z) = K_1[1 + (K_2 + K_3)z^{-1} + K_2K_3z^{-2}]R(z) \tag{3.15}$$

Applying the Real Translation Theorem from Z-Transform theory yields the following time domain recursive equations:

$$C[nT] + (K + K)C[(n - 1)T] + K K C[(n - 2)T]$$
$$= K_1[R(nT) + (K_2 + K_3)R[(n - 1)T] + K_2K_3R[(n - 2)T] \tag{3.16}$$

$$C[nT] = -\sum_{i=1}^{2} a_i C[(n - i)T] + K_1 \sum_{i=0}^{2} b_i R[(n - i)T] \tag{3.17}$$

where the a's and b's are obvious from inspection of Eq. (3.16). By means of Eq. (3.17), the output C(nT) can be found for any input by merely specifying the sequence R(nT) as the integer n varies over the range of interest.

The location of the poles and zeros of F(p) and the location of the all-pass filter break frequency can be optimized, using whatever performance criterion desired.

Application to Autopilot

The design procedure outlined on the previous pages will now be applied to the pitch portion of the continuous time autopilot. The open-loop transfer function for this system is

$$H(s) =$$

$$\frac{36}{s^2 + 8.4s + 36} \cdot \frac{s^2 + 2.31s + 2.72}{s^2 + 5.62s + 3.1} \cdot \frac{s + 1.65}{s + .62} \cdot \frac{s^2 + 7.25s + 81}{1.125s^2 + 13.33s + 81} \tag{3.18}$$

The sampling frequency was chosen to be 40 rad/sec. This allows the range of interest to extend to approximately twice the highest critical frequency of the autopilot.

The fictitious transfer function H(p) was synthesized in the following manner:

1. mag[H(jω)] for $0 \leq \omega \leq 20$ was calculated, as well as the frequency λ, using the relation $\lambda = \tan(\omega T/2)$.

2. A plot of mag[H(jλ)] versus λ was made (Fig. 3.4).

3. Real poles and zeros of H(s) at ω_i were transformed into real poles and zeros of H(p) at λ_i, where λ_i is given by Eq. (3.7).

4. Complex roots of H(s) with undamped natural frequencies of ω_n were transformed into complex roots of H(p) with natural frequencies also given by Eq. (3.7). The values of the damping ratios were preserved in synthesizing H(p).

5. An initial value for a second order real zero was made by inspection of the magnitude plot made in step 2. This additional term is due to the asymptotic approach of H(p) to the value of H(s) at $\omega = \omega_s/2$.

6. The substitution $p = (z - 1)/(z + 1)$ was made in the expression for H(p).

7. The magnitude and phase plots of H(z) versus ω are made.

8. An initial choice for the all-pass filter break frequency was made, and the phase curve re-plotted.

At this stage the magnitude curve of H(z) differs from that of H(s) by a maximum of 1.7 dB at a frequency of 6.5 rad/sec. The phase curve of H(z) has the same general shape as that of H(s) up to approximately 15 rad/sec, but has an error of about 17-22 degrees from 8 to 15 rad/sec. It was decided at this point to optimize several parameters of H(z).

The locations of the second order real zero added during the design process and the term resulting from the single real pole of H(s) at $\omega = 5$ rad/sec were simultaneously optimized. The performance criterion used was the mean absolute error between the frequency domain magnitude

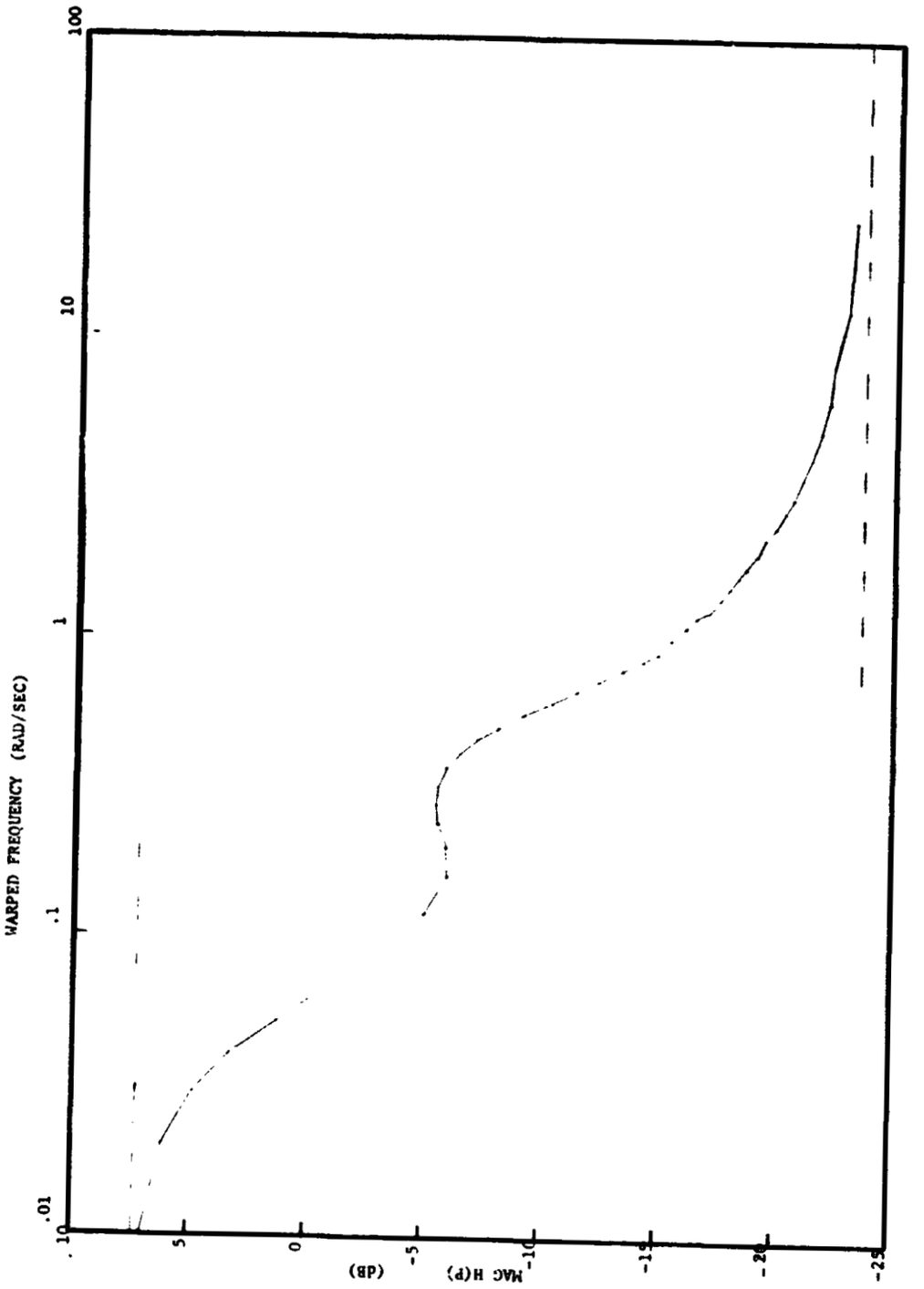


Figure 5.1 Magnitude Plot for Autopilot Transfer Function in Warped Frequency Domain

of H(s) and the magnitude of H(z) over the frequency range 0 to $\omega_s/2$. The values of these two parameters were independently stepped through the range 0 to 20 in intervals of 0.08 so that essentially all space was spanned. Using the expression for H(z) with the two optimized parameters, the break frequency for the all-pass filter was optimized in a similar fashion. The performance criterion used was the mean absolute error between the frequency domain phase curve for H(s) and that of H(z).

The frequency characteristics of the final design are shown in Figures 3.5 and 3.6. Figure 3.5 shows the magnitudes of H(s) and H(z) plotted versus the real frequency ω. The maximum magnitude error is approximately 1.28 dB which occurs at 20 rad/sec., and the mean absolute magnitude error is .342 dB. Figure 3.6 shows the phase shift curve for H(s) and the phase shift curve for H(z) with and without the all-pass filter. For the curve with the all-pass filter over the frequency range 0-16 rad/sec. the following error information was obtained. The maximum phase error between H(s) and H(z) was 12.74 degrees, occurring at 16 rad/sec., and the mean absolute phase error was 3.86 degrees. Over the full frequency range 0-20 rad/sec., the maximum error was 42.4 degrees, occurring at 20 rad/sec., and the mean absolute phase error was 8.61 degrees.

The optimization procedure was carried out again; this time, using the mean squared error criterion on the magnitude curve and the mean absolute error on the phase curve. The maximum magnitude error was 1.17 dB, and the mean absolute magnitude error was .364 dB. Over the frequency range 0-16 rad/sec., the maximum and mean absolute phase errors were 12.34 degrees and 3.52 degrees, respectively. For the frequency range 0-20 rad/sec. the maximum and mean absolute phase error were 42.4 degrees and 8.30 degrees, respectively. In each case the maximum error occurred at the highest frequency in the range specified.

Trying to use the mean squared error criterion to optimize the location of the all-pass filter produced poor results when applied to this system. With an all-pass filter, this design procedure produces a

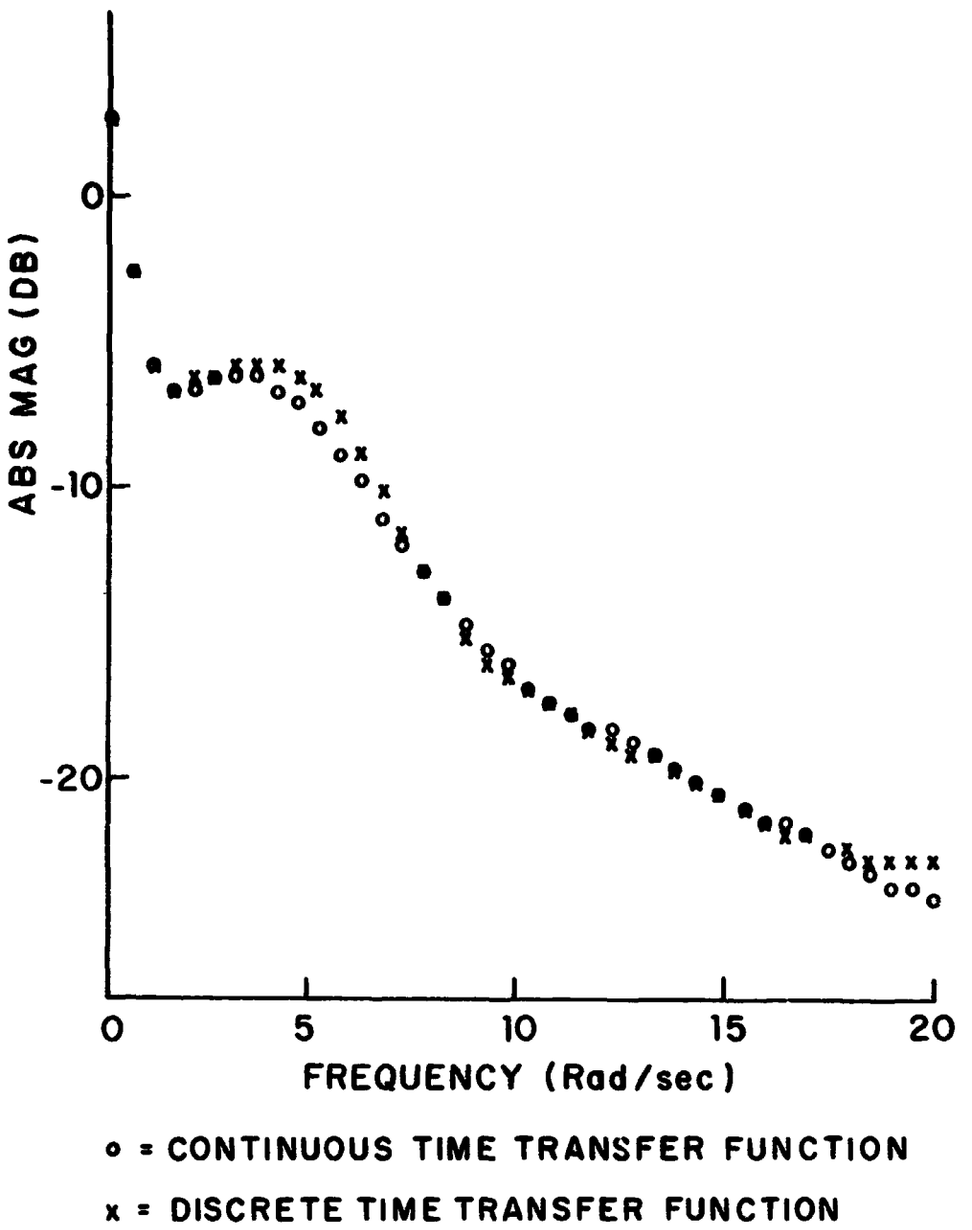


Figure 3.1 Magnitude of Autopilot and its Discrete Approximation

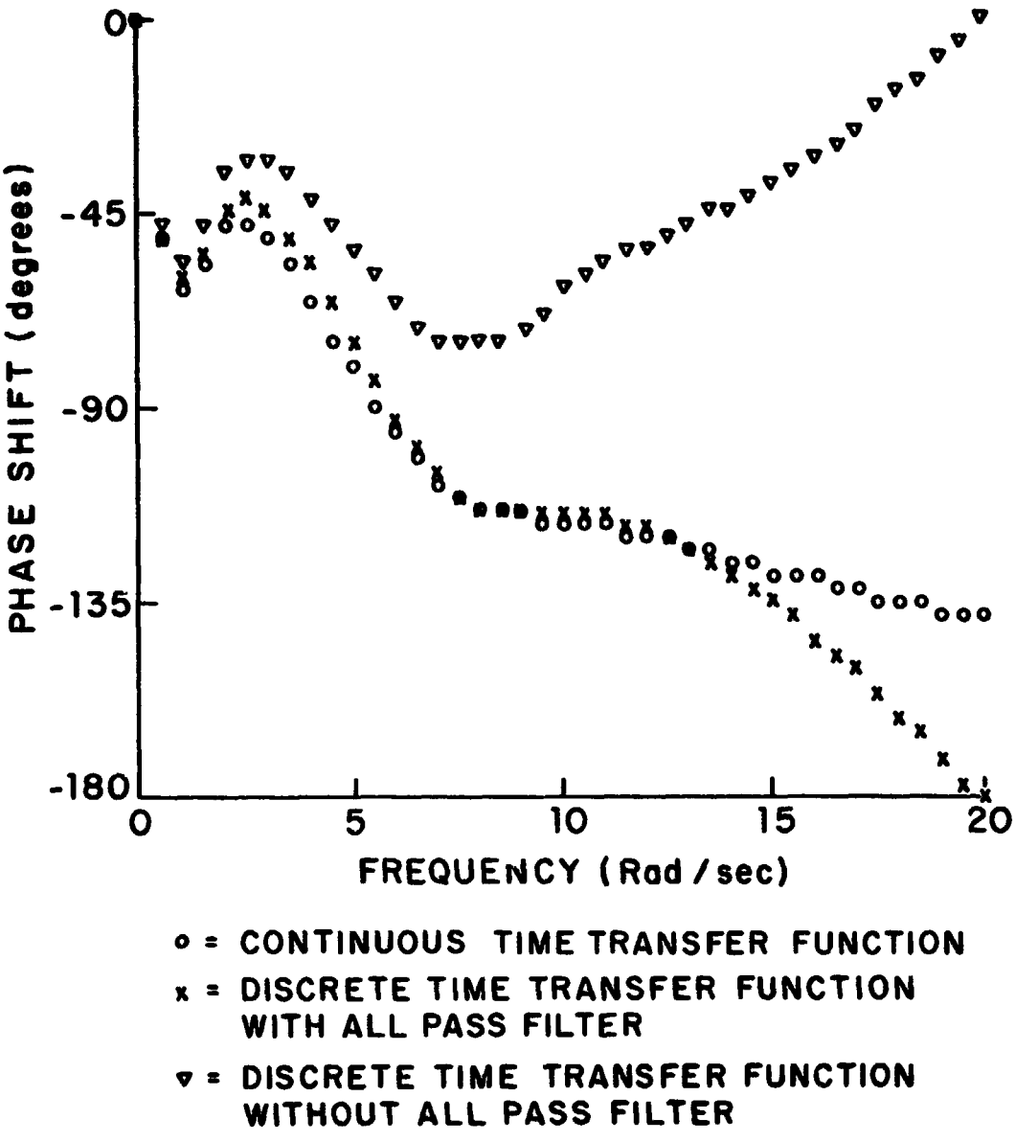


Figure 3.6 Phase of Autopilot and its Discrete Approximation with and without All-Pass Filter

phase shift of -180 degrees at $\omega_s/2$, regardless of the phase shift of H(s). For the continuous autopilot the phase shift at 20 rad/sec. is -137.6 degrees. With the sampling frequency fixed, nothing can be done to reduce this large error at $\omega_s/2$. Since the mean squared error criterion tends to accentuate large errors, applying this criterion to the phase curve results in increased errors at low frequencies without substantially reducing the errors near $\omega_s/2$.

For the optimization schemes tried it appears that for this system the mean squared magnitude error and mean absolute phase error criteria produced slightly smaller errors than the mean absolute magnitude and phase error criteria. The transfer functions for the two realizations are shown below. In each case they are of the form:

$$H(z) = \frac{KN(z)}{D(z)}$$

$$N(z) = H_1(z) \cdot H_2(z) \cdot H_3(z) \cdot H_4(z) \cdot H_9(z)$$

$$D(z) = H_5(z) \cdot H_6(z) \cdot H_7(z) \cdot H_8(z)$$

In the table below the heading MAM/MAP indicates the mean absolute magnitude and mean absolute phase error criteria, and MSM/MAP indicates the mean squared magnitude and mean absolute phase error criteria.

	MAM/MAP	MSM/MAP
K	7.88489 E-02	7.73779 E-02
$H_1(z)$	$z - .769409$	$z - .769409$
$H_2(z)$	$z^2 - 1.63918z + .69577$	$z^2 - 1.63918z + .69577$
$H_3(z)$	$z^2 - .223826z + .4308$	$z^2 - .223826z + .4308$
$H_4(z)$	$(z + .230133)^2$	$(z + .223527)^2$
$H_5(z)$	$(z - .907063)^2$	$(z - .907063)^2$
$H_6(z)$	$z - .470564$	$z - .486006$
$H_7(z)$	$z^2 - .750535z + .276885$	$z^2 - .750535z + .276885$
$H_8(z)$	$z^2 - .280832z + .191517$	$z^2 - .280832z + .191517$
$H_9(z)$	$(z + 2.72379)/(z + .367135)$	$(z + 2.72379)/(z + .367135)$

$H_4(z)$ represents the second order real zero added during design, $H_6(z)$ is the single real pole of H(s) whose location in H(z) was optimized; and $H_9(z)$ is the all-pass filter.

Design Considerations

Several factors concerning this design procedure should be mentioned. First, when H(z) is expressed as a ratio of polynomials, the degree of the numerator and denominator in powers of z will be equal. The degree of the polynomials in z will be equal to one, plus the highest degree of the polynomials in H(s). The additional power of z is due to the all-pass filter. In this example H(s) was of seventh degree in the denominator and fifth degree in the numerator. H(z) was of eighth degree in both numerator and denominator. Because of the equal degrees of the numerator and denominator in H(z), the phase shift at $\omega_s/2$ will be 0 degrees if no all-pass filter is used, and -180 degrees if a phase lag all-pass filter is included. A phase lead all-pass filter, $(\lambda + p)/(\lambda - p)$, produces an unstable condition. In the frequency domain a good fit can be obtained for the magnitude curve over the range $0 \leq \omega \leq \omega_s/2$. For the phase curve the closeness of fit depends on the phase characteristics of the continuous system.

Time Domain Evaluation

In the following discussion concerning time domain performance the H(z) obtained from the mean absolute magnitude and phase error criteria will be used. This H(z) will be referred to as the DISCRETE approximation to H(s).

To obtain the time domain response of the continuous time autopilot a fourth order Runge-Kutta-Gill numerical integration was performed. The integration step size was chosen to be T/10. This step size provides approximately 48 samples per period of the highest damped natural frequency in H(s) and 12 samples per time constant for the shortest exponential time constant. The inputs chosen were the step function and sine waves of each of the integer radian frequencies from 1 to 20 rad/sec., inclusive.

The time response of the Tustin approximation to the autopilot was obtained by expressing H(s) as a ratio of polynomials and making the following substitution:

$$\left(\frac{1}{s}\right)^K = \left[\frac{T}{2}\frac{(z+1)}{(z-1)}\right]^K$$

The resulting expression was put in the form of a recursive equation and solved for the same set of inputs previously mentioned.

The H(z) obtained from the design procedure described was also expressed in recursive form and solved for the time response. For each of the three sets of solutions the outputs were calculated over the time range 0-100 T. For the Runge-Kutta solutions this means that ten integrations will be performed per output sample.

Using the Runge-Kutta solution as a reference, the mean squared errors of the Tustin and Discrete approximations were calculated for each of the inputs and averaged over the sinusoidal inputs. Figure 3.7 shows the errors as a function of input sinusoidal frequency. Each of the approximations had negligible steady state error for the step input. The table below lists the error data for the two approximations.

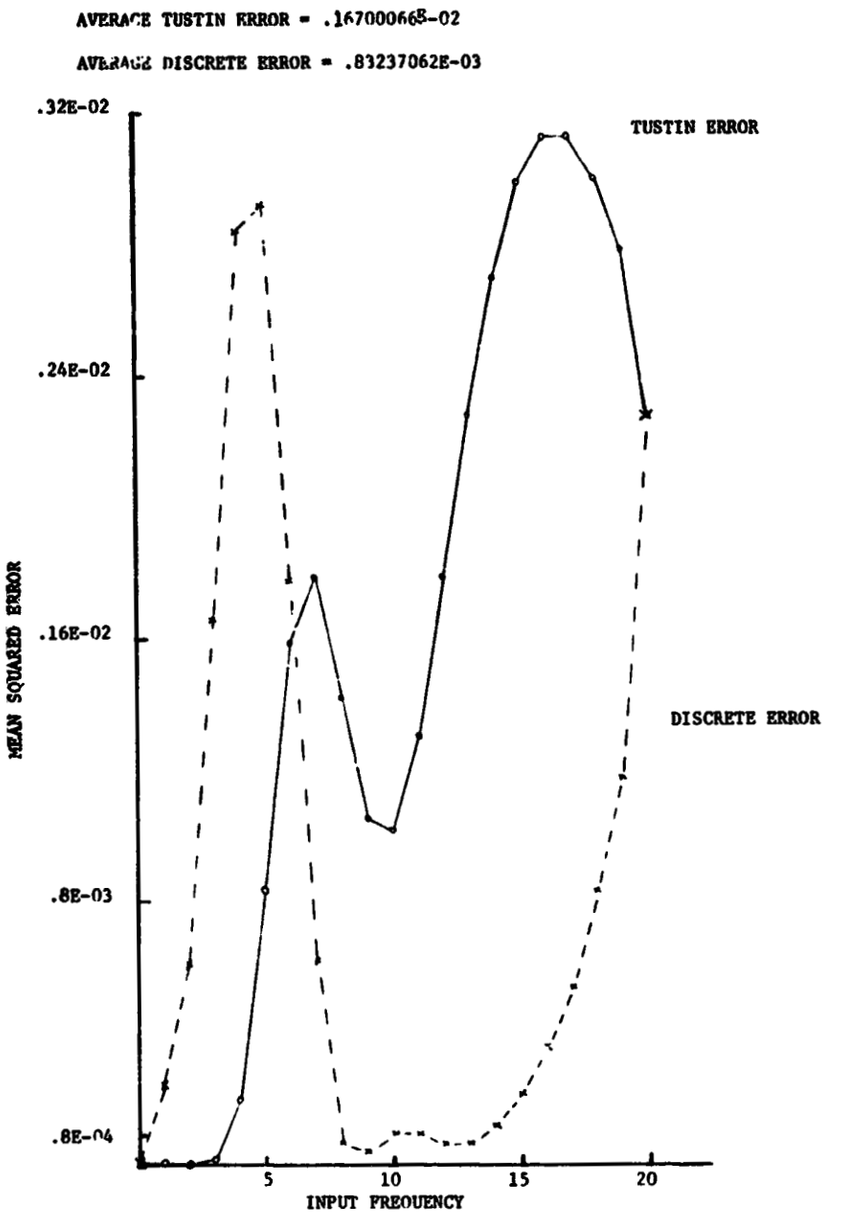


Figure 3.7 Mean Square Error Plot for Discrete Approximation and Tustin's Approximation

MEAN SQUARE ERROR

RAD/SEC	TUSTIN	DISCRETE
1	.16553969 E-05	.24048390 E-03
2	.28388786 E-05	.16118844 E-03
3	.22027003 E-04	.16690678 E-02
4	.22509128 E-03	.2850k575 E-02
5	.84790011 E-03	.29299688 E-02
6	.15901896 E-02	.17630448 E-02
7	.17861110 E-02	.57113966 E-03
8	.14286007 E-02	.78087308 E-04
9	.10556364 E-02	.46068448 E-04
10	.10220778 E-02	.94572395 E-04
11	.13234237 E-02	.95276045 E-04
12	.17965967 E-02	.74684822 E-04
13	.22863912 E-02	.76157714 E-04
14	.26955963 E-02	.11930201 E-03
15	.29781393 E-02	.20891043 E-03
16	.31201926 E-02	.34898954 E-03
17	.31259800 E-02	.54962620 E-03
18	.30092672 E-02	.82658841 E-03
19	.27900444 E-02	.11962247 E-02
20	.22923737 E-02	.22923735 E-02

Averaging over the 20 sinusoidal inputs yields the following data:

Tustin Error = .16700066 E-02

Discrete Error = .83237062 E-03

For the step input the errors are:

Tustin Error = .49399356 E-03

Discrete Error = .99913922 E-03

For the sinusoidal inputs the average mean squared error for the Discrete approximation is one half that of Tustin's approximation. The greatest improvements over Tustin's method lie in the frequency range 8-17 rad/sec. In this range the magnitude curve of the DISCRETE system fits that of the continuous system very closely, with a maximum difference of .34 dB. The phase curves also have a good fit over most of this frequency range, with less than 10-degree error from 8 to 15 rad/sec. and 18-degree error at 17 rad/sec.

Over this same frequency range, the Tustin frequency domain magnitude error varies from 2.1 to 15.7 dB, referenced to the exact magnitude function. The phase shift error in this frequency range averages 16 degrees, with a maximum of 32 degrees at 17 rad/sec. and is 25 degrees at 15 rad/sec.

Conclusion

A procedure has been presented by which a discrete time system can be designed to have frequency domain characteristics similar to that of a continuous time system. For the magnitude curve a close fit can be obtained over the frequency range 0 to $\omega_s/2$. For the phase curve the closeness of fit depends on the phase characteristics of the continuous system. The design procedure can be carried out either graphically or by computer. The location of all critical frequencies in H(p) and, thus, the form of H(z) could be optimized, based on a number of different performance criteria. An improvement in time domain performance in the middle and upper frequencies was obtained, compared with Tustin's method, with a degradation of performance in the low frequencies.

It appears that frequency domain methods, and the particular procedure described here, are valid design approaches. The characteristics of the continuous time system being modeled and the input frequency ranges of interest will determine which approach is best.

IV. SUBSTITUTIONAL METHODS

An investigation into the suitability of various substitutional formulas and techniques for real time simulation is being conducted. The study is divided into two phases, the first being concerned with linear systems and the second with nonlinear systems. Results from the first phase were reported in the Semi-Annual Report of September 1975 and will not be repeated herein. Progress to date in the second phase of this work is the subject of the remainder of this section.

Several methods for digital simulation of linear systems have been investigated, and results show that the IBM method is probably the best in terms of error and computation time. Present efforts are being directed at the study of nonlinear systems.

Along with the substitutional methods, other methods, such as IBM, Optimum Discrete Approximation and Discrete Compensation, are being considered. There are no other methods currently available from literature and publications.

Comprehensive studies of simulations of nonlinear systems, such as error analysis, selection of simulation increment, etc. is currently beyond the state-of-the-art. This is because nonlinear systems are difficult to classify. Comparisons of these digital simulation methods, therefore, are largely experimental. Several systems with different degrees of complexity and nonlinearities will be studied before any conclusions are reached.

A simple and "slightly nonlinear" system was investigated, and the results showed little effect by the nonlinearity on the overall system response obtained with different methods. Another system is currently under investigation and will be simulated as soon as the design is completed.

The following paragraphs suggest the methodology being used to develop the necessary computer software for analyzing various procedures.

1. IBM method: The objective is to derive transfer functions of Fig. 4.1 such that the poles of the individual transfer functions ($G_1(z)$

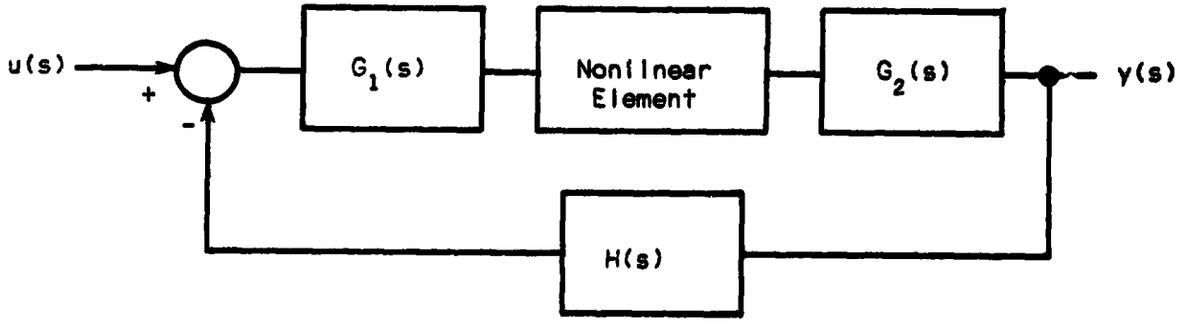


Figure 4.1 A General Nonlinear Feedback System

and $G_2(z)$), as well as the poles of the overall closed loop transfer function, are correct. This is done by matching the closed-loop eigenvalues and the static gains of the continuous and discrete system models.

Poles of $G_1(z)$ and $G_2(z)$ are in the same location as $G_1(s)$ and $G_2(s)$ in the s-plane; therefore, the transient response will be correct. This characteristic is not adequately preserved in other substitutional methods (Tustin, Boxer-Thaler, Madwed, etc.). Note that $G_1(z)$ and $G_2(z)$ are <u>not</u> the Z-transforms of $G_1(s)$ and $G_2(s)$ but are transfer functions to simulate the continuous response.

The design procedure is as follows:

(a) Replace the continuous transfer functions by the discrete transfer function, using the Z-transform. A single-period delay is inserted in the feedback paths to insure realizability. This delay will be compensated for in the final model.

(b) For each transfer function the static gains are matched between the discrete and continuous blocks. The final value theorem is applied to find the gain necessary to equate the static gains between the two transfer functions.

(c) Each nonlinear element is replaced by a nominal gain, and the closed loop eigenvalues of the continuous and the discrete systems are made equal by inserting a gain in the forward loop. This is the most tedious task for a complex system of order higher than 4 and with a multiple input/output. Fortunately, this mechanism can be implemented on a digital computer, as proposed in an IBM report (<u>Numerical Techniques for Real-Time Digital Flight Simulation</u>). This program will be discussed in more detail in a later part of this section.

(d) Finally, the steady-state gains of the over-all closed loop system are matched. The discrete system is also matched to either a specific input or its approximation. An input transfer function will be attached in front of the discrete system. There are two ways of approximating the input:

one by a stair-step function (zero order hold); and one by straight line segments (first order hold).

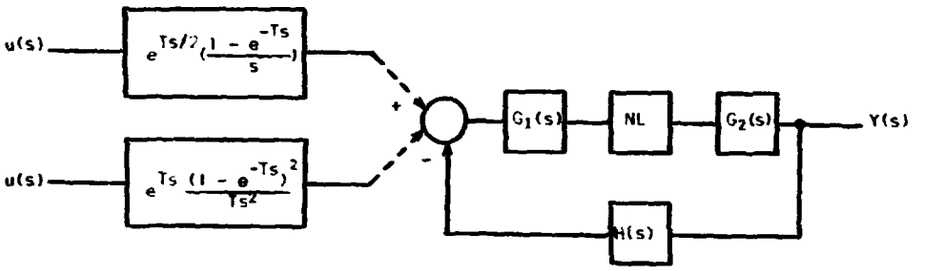
Input approximation by a stair-step function: To find the input transfer function, equate the Z-transform of the over-all continuous system with the transfer function of the over-all discrete system. Since the zero-order hold will introduce a half-period lag, the stair-step approximation must be designed with a half-period advance so as to obtain the correct simulation.

Input approximation by straight-line segments: The same procedure is applied as above, but no compensation for the time shift is necessary, A summary in block diagram form is shown in Fig. 4.2.

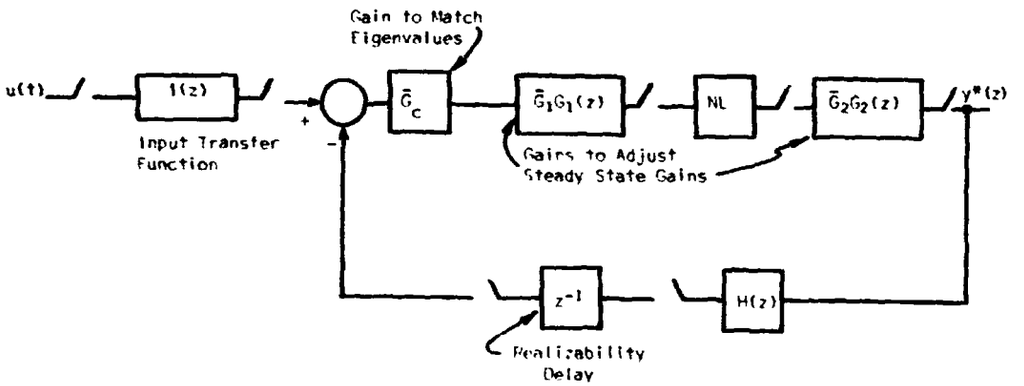
Programming details for matching eigenvalues in the IBM method are as follows. As mentioned earlier, matching eigenvalues of complex continuous and discrete systems is a formidable task which cannot be performed by hand. It has been proposed that computer programs be used to plot root loci of the systems and, thereby, determine the gain required to match their pole locations.

There are two possible approaches for using a digital computer to compute root loci. One is to determine the characteristic equation as a function of loop gain and solve it for various values of loop gain. In a multiple loop complex system this is not always possible. Therefore, it is desirable to have the computer derive the characteristic equation. Another approach is to apply Evan's rule directly with the aid of the computer. This method, however, does not fully take advantage of the high speed computer.

IBM has developed a method based on Evan's rule and used the computer efficiently at the same time. The program developed by IBM allows the user to find root loci of systems (discrete or continuous) directly from the block diagram. In the very near future a similar program which can be used on our computer will be developed so that more complex systems can be simulated.



(a) Continuous System to be Simulated



(b) Final Form Using IBM Method for Simulation

Figure 4.2 Block Diagram of Design Method

The principle of the program is as follows. The characteristic equation of the system shown in Fig. 4.3 is:

$$KG(s)H(s) + 1 = 0 \tag{4.1}$$

or

$$G(s)H(s) = -1/K \tag{4.2}$$

Equation (4.1) is satisfied only if G(s)H(s) is real, since K is real. Furthermore, if a complex number $S = \sigma + j\omega$ is substituted into G(s)H(s), the result will be another complex number $q = u + iv$. Therefore, the solution of Eq. (4.1) is all s that will give $v = 0$ and $K = -1/U$. It can also be shown that for any path crossing a locus, u will change sign. Therefore, the left-half plane of s is scanned until a change in sign of v is observed, then the point lying on the locus can be found by iteration. The same principle can be applied for a discrete system to find its root locus.

It has been reported that the developers of the IBM method successfully simulated in real-time complex, six-degree-of-freedom space vehicles and aircraft with considerable improvements over other methods. For example, for the same degree of accuracy, this method reduces the computation time by 90 to 95% over the Runge-Kutta method in several cases.

2. Optimum Discrete Approximation:

The objective of this design technique is to approximate each element of the transfer function so that the summation of squared error is minimized. This technique eliminates the need for inserting a period delay in the feedback path, since the realizability has been taken into account during the design process of the discrete transfer function. Each discrete transfer function is considered to consist of a tandem connection of two elements, $F_1(z)$ and $F_2(z)$. $F_2(z)$ is called the fixed portion, and it can take the form

$$F_2(z) = Z^{-p} \tag{4.3}$$

where $p = 0, 1$. Whether or not $F_2(z) = 1$ or $F_2(z) = Z^{-1}$ depends on the system under consideration. Transfer functions where the degree of the numerator is lower than the degree of the denominator are called closed-

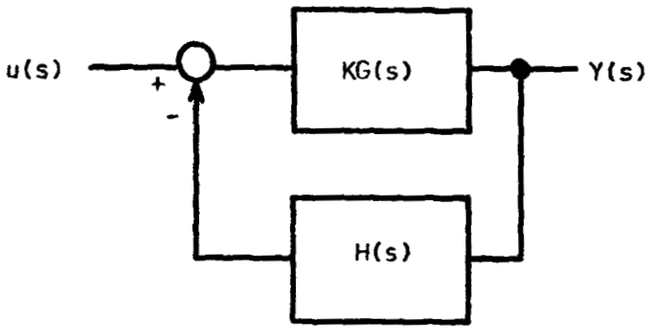


Figure 4.3 General System Diagram

loop realizable to ensure the realizability of the whole system. Therefore, in designing the discrete transfer functions we initially select $F_2(z) = 1$. If none of the discrete transfer functions turn out to be closed-loop realizable, then we pick the simplest one to be redesigned with $F_2(z) = Z^{-1}$. This will greatly simplify the work involved.

Sage and Burt have shown that the discrete "optimum" transfer function is

$$F_1opt(Z) = \frac{\left\{ \dfrac{F_2(Z^{-1})U(Z^{-1})A(Z)}{[F_2(Z)F_2(Z^{-1})U(Z)U(Z^{-1})]_-} \right\}_{p^+}}{[F_2(Z)F_2(Z^{-1})U(Z)U(Z^{-1})]_+} \tag{4.4}$$

with

$$A(z) = Z[G(s)U(s)] \tag{4.5}$$

where $U(t)$ is a test signal and where $[\cdot]_+$ denotes all poles and zeros within the unit circle, $[\cdot]_-$ denotes all poles and zeros outside the unit circle, and $\{\cdot\}_{p+}$ is the portion with poles within the unit circle. The procedure can be summarized as follows:

1. Select a test signal, $U(s)$, usually either a unit ramp or a unit step.

2. Design each discrete transfer function, using the $F_1opt(z)$ equation. Let $F_2(z) = 1$ in each case; and, if $F_1opt(z)$ is not closed-loop realizable, then redesign it with $F_2(z) = Z^{-1}$. Then

$$G(z) = F_1opt(z)F_2(z). \tag{4.6}$$

The procedure above does not take into account the nonlinear characteristic of the system. It will work adequately for a slight nonlinearity. If the system is decidedly nonlinear, the approximation error can be reduced if we make use of the fact that the system is actually nonlinear in determining the "optimum" discrete approximation. We can expand the method above, using gain parameters before and after the nonlinearity. These gain parameters will be adjusted during the simulation process.

3. Discrete Compensation:

In substitutional methods it is necessary to insert a period delay in the feedback path to ensure realizability. This may shift the pole locations in some systems. The IBM and Optimum Discrete Approximation methods compensate this delay in the design process. In the Discrete Compensation method the closed-loop transfer function is discretized, using any discrete integrator, but improvements are made through a compensator to adjust the eigenvalues of the closed-loop system.

For example, take a simple system, such as shown in Fig. 4.1 and let

$$G_1(z) = G_1(s)\Big|_{\frac{1}{s} = \text{some integrator}} \tag{4.7}$$

$$G_2(z) = G_2(s)\Big|_{\frac{1}{s} = \text{some integrator}} \tag{4.8}$$

$$H(z) = H(s)\Big|_{\frac{1}{s} = \text{some integrator}} \tag{4.9}$$

We then discretize the whole system as

$$G_A(z) = \frac{\bar{G}_3 G_1(s) G_2(s)}{1 + H(s)\bar{G}_3 G_1(s) G_2(s)}\Bigg|_{\frac{1}{s} = \text{some integrator}} \tag{4.10}$$

where $\bar{G}_3$ is the nominal gain of the nonlinear element. Insert a compensator after the input of the discrete system and compare it with $G_A(z)$ to obtain the coefficients of the compensator, whose form is

$$D(z) = \frac{b_0 + b_1 z + \quad + b_m z^m}{a_0 + a_1 z + \quad + a_n z^n} \qquad (m < n) \tag{4.11}$$

$$G_A(z) = D(z)\,\frac{G_1(z) G_2(z) \bar{G}_3}{1 + \bar{G}_3 G_1(z) G_2(z)} \tag{4.12}$$

The final form is shown in Fig. 4.4.

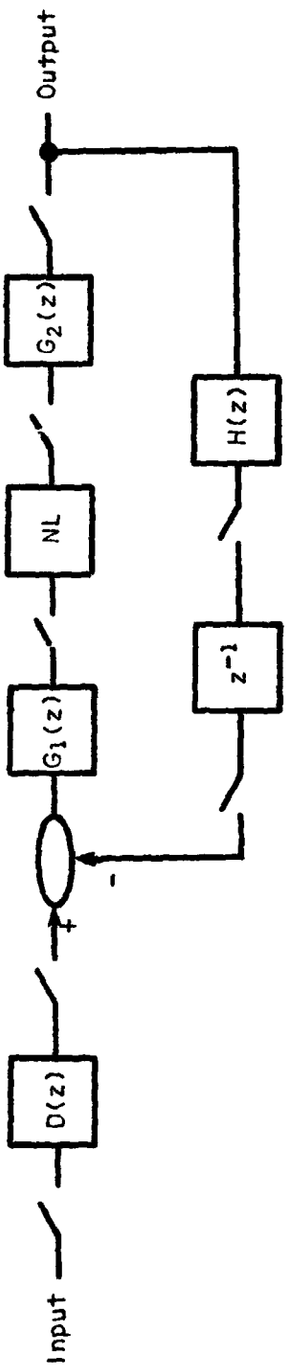


Figure 4.4 Final Form of Discrete Compensation Method

If further accuracy is required (at the expense of more computation time), the coefficients of D(z) can be adjusted, depending on the instantaneous gain of the nonlinear element; that is, instead of using $\bar{G}_3$ as a nominal gain to obtain fixed coefficients of D(z), we can make the coefficients of D(z) a function of $\bar{G}_3$, where $\bar{G}_3$ can be determined by direct measurement or by interpolation.

Summary

In the three methods considered above the Discrete Compensation Method is the best in terms of design efforts. Its drawback is excess computation time, especially with the coefficient adjustor. Once the computer program for the IBM method is completed, the IBM method may be the logical choice, even though the Optimum Discrete Approximation may be more accurate for decidedly nonlinear systems. However, no final conclusions can be drawn at this time.

Future Tasks

- Developing a computer program to aid the design of the IBM method.
- Further search and studies of Sage's method, using calculus of variations and other techniques for a nonlinear system.
- Simulations of various nonlinear systems with special attention given to the criteria mentioned above, observing possible effects of a certain nonlinearity on a certain method.

V. USE OF PADÉ APPROXIMANTS TO THE MATRIX EXPONENTIAL FOR COMPUTER SOLUTIONS OF STATE EQUATIONS

Introduction

In system theory, there is a large class of problems which may be phrased in terms of linear time-invariant differential equations, and which lend themselves to straightforward solutions. A second class, at the opposite end of the spectrum, consists of problems characterized by behavior which includes large time variation and strong non-linearities. There is a middle ground, however, consisting of a class of problems in which the time-dependent parameters vary relatively slowly, with weak non-linearities. In these, it may be necessary (or merely desirable) to include the effects of time- and state-dependent variables, and at the same time undesirable to utilize algorithms used normally on systems of large computational complexity. The results of the theory of linear time-invariant systems may be used, with proper modification, to approximate this class of systems very closely.

The matrix differential equation

$$\dot{x} = Ax + Bu \tag{5.1}$$

where A and B are constants, has the well known solution

$$x(t + T) = G(T)x(t) + H(T)u(t),\ t = kT,\ k = 0,.,2,\ \ldots\ , \tag{5.2}$$

where T is taken to satisfy the Nyquist criterion on u,

$$G(T) = \exp(AT) = I + AT + (A^2T^2/2!) + \ldots\ , \tag{5.3}$$

and

$$H(T) = \int_0^T \exp(A\tau)d\tau B = [(IT + (AT^2/2!) + (A^2T^3/3!) + \ldots)]B. \tag{5.4}$$

The last term may be reduced to

$$H(T) = A^{-1}[(\exp(AT) - I)]B, \tag{5.5}$$

if the inverse of A exists. The above may be derived directly from the Taylor series of x(t + T).

A more general statement is true where A = A(x,t) and is a slowly varying function of both x and t. For the class of problems where the system time-dependence and non-linearities are small, we may still use the matrix exponentail to approximate the system's behavior [5.1]. That is,

$$x(t + T) = \exp(A(x(T),t)T)x(t) + \int_0^T \exp(A(x(T),t)\tau)d\tau Bu(t). \qquad (5.6)$$

Allowing A to vary with x and t, however, suggests consideration of efficient techniques for computation of exp(AT). In some cases, simple summation of the terms of the power series representation will be sufficient since A may vary so slowly that only occasional updatings of exp(AT) are necessary. In other cases, it may be necessary to recompute exp(AT) at each sample point to achieve the desired accuracy. The Padé approximants to exp(AT) which will be considered offer, in some types of problems, significant advantages in computational speed and accuracy over the power series representation.

Definition of Padé Approximants

A Padé approximant to a scalar power series F(z) is a ratio of two polynomials P(z) and Q(z), of order p and q respectively, abbreviated (p,q). Its significant feature is that the power series expansion of (p,q) is identical with that of F(z) up to and including the coefficient of z^N, where N = p + q is the order of the approximant. There are N + 1 Padé approximants of order N, and they are unique [5.2]. For example, the three approximants for N = 2 and F(z) = exp(z) are:

$$\begin{aligned} (2,\ 0) &= 1 + z + z^2/2 \\ (1,\ 1) &= (1 + z/2)/(1 - z/2); \quad |z| < 2 \\ (0,\ 2) &= 1/(1 - z + z^2/2) \qquad |z| < 1.414 \end{aligned} \qquad (5.7)$$

Similar statements may be made of Padé approximants to a matrix power series; for example, the (1, 1) approximant to exp(AT) would be:

$$(1,\ 1) = (I - AT/2)^{-}(I + AT/2) = (I + AT/2)(I - AT/2)^{-} \qquad (5.8)$$

Of the general class of Pade approximants, some are more effective computationally than others. In general it requires no more computation to calculate a (P, P) Padé approximant than a (P-M,P) or a (P, P-M), where M < P. Because the (P, P) approximant is accurate through more terms, it is the most beneficial form to use. In this paper we shall compare (1) the truncated series (N, 0), and, (2) those in which p = q, i.e., (1, 1), (2, 2), (3, 3), and so forth. The scalar approximants (2, 2) and (3, 3) for exp(z) are:

$$(2, 2) = (1 + z/2 + z^2/12)/(1 - z/2 + z^2/12); \quad |z| < 3.464 \qquad (5.9)$$

$$(3, 3) = (1 + z/2 + z^2/10 + z^3/120)/(1 - z/2 + z^2/10 - z^3/120); \quad |z| < 4.644 \qquad (5.10)$$

It is shown in Appendix A that the error involved in using a Padé approximant to the matrix series exp(AT) is a linear function of the error in approximating the scalar series $\exp(\lambda_m T)$ where λ_m is the magnitude of the maximum eigenvalue of A. This allows discussion of accuracy in terms of the scalar series $\exp(\lambda_m T)$, with results which carry over directly to the matrix series approximation.

Accuracy vs. Required Computation

The two classes of approximations for the exponential were compared directly, using three criteria: (1) the numer of matrix operations required for the approximation to exp(AT), (2) the percent error in terms of the scalar approximation of $\exp(\lambda_m T)$, and (3) the allowable range of $\lambda_m T$ for which reasonable results could be achieved. It will be shown that the inclusion of higher-ordered terms in the (p, p) type of approximant reduces error, extends the range of $\lambda_m T$, and is in general more efficient than the corresponding (p + p, 0) truncated series.

A matrix operation may be defined as a multiplication or an inverse; each requires about n^3 multiplications and divisions, where n is the order of the matrix [5.3]. This is an effective standard, since the matrix manipulations dominate the calculation time. By this standard, the (N, 0) series requires N-1 matrix operations, while the (p, p) series requires

p + 1 operations. For example, the (2, 2) and the (4, 0) both require three matrix operations; however the (3, 3) uses only four while the (6, 0) requires five operations. For systems of large order, this may be significant.

A computer was used to calculate % error in the various forms of approximations to exp ($\lambda_m T$). Some of the more important features are demonstrated in Figs. 5.1 and 5.2. For example, in Fig. 5.1, for $\lambda_m T = .5$, the (4, 0) truncation yields about .04% error. The (2, 2) gives approximately .004% error, or about one order of magnitude improvement. For large values of $\lambda_m T$ the results are even more dramatic. For $\lambda_m T = 2.0$, a (9, 0) approximation gives .17% error compared to the .15% error of the (3, 3) approximation, which also affords a savings of four matrix operations. This alone may result in a considerable reduction in computer time. Fig. 5.2 shows the maximum allowable values of $|\lambda_m T|$ for given amounts of error. The (p, p) type of approximant is seen to have consistently greater range than its corresponding truncated series.

Previously, it has been necessary to choose $|\lambda_m T|$ to be relatively small in order to achieve the required accuracy with a short series approximation. Using the Padé (p, p) approximants, there is much more room for choice. Larger system eigenvalues may be included in the problem statement, or a longer sample time T chosen (subject to the Nyquist rate on u(t)), and the desired accuracy may still be achieved, or bettered, within the context of a real-time computer simulation.

Summary

In summary then, it has been shown that the (p, p) Padé approximant to the matrix exponential exp(AT) is a useful tool in control system simulation and operation. It has the advantages of greater accuracy, larger range, more flexibility, and in some cases greater computational efficiency than the truncated series approximation.

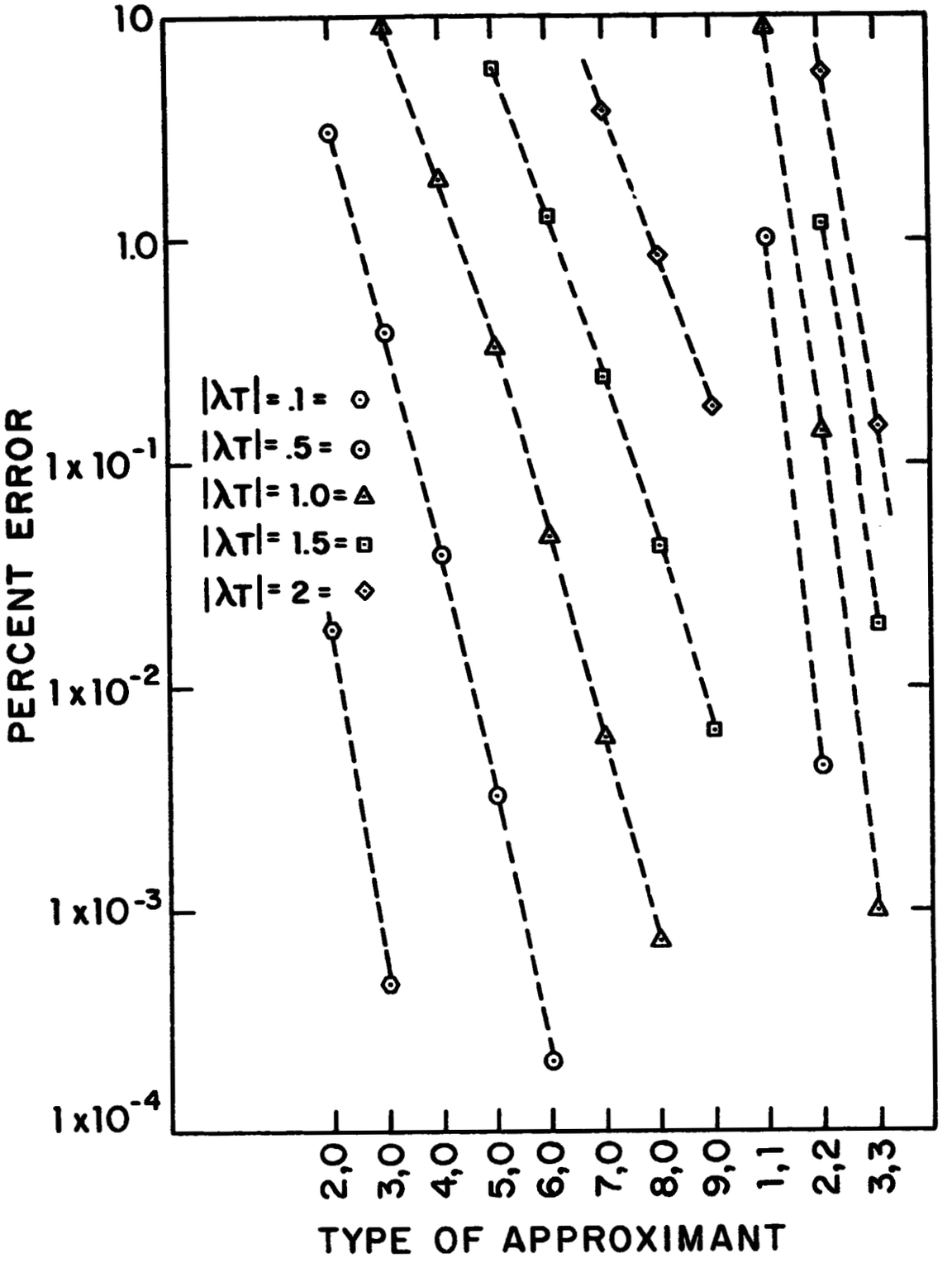


Figure 5.1 Percent Error as a Function of Order of Approximant for Both the (N,0) and the (p,p) Types of Approximants

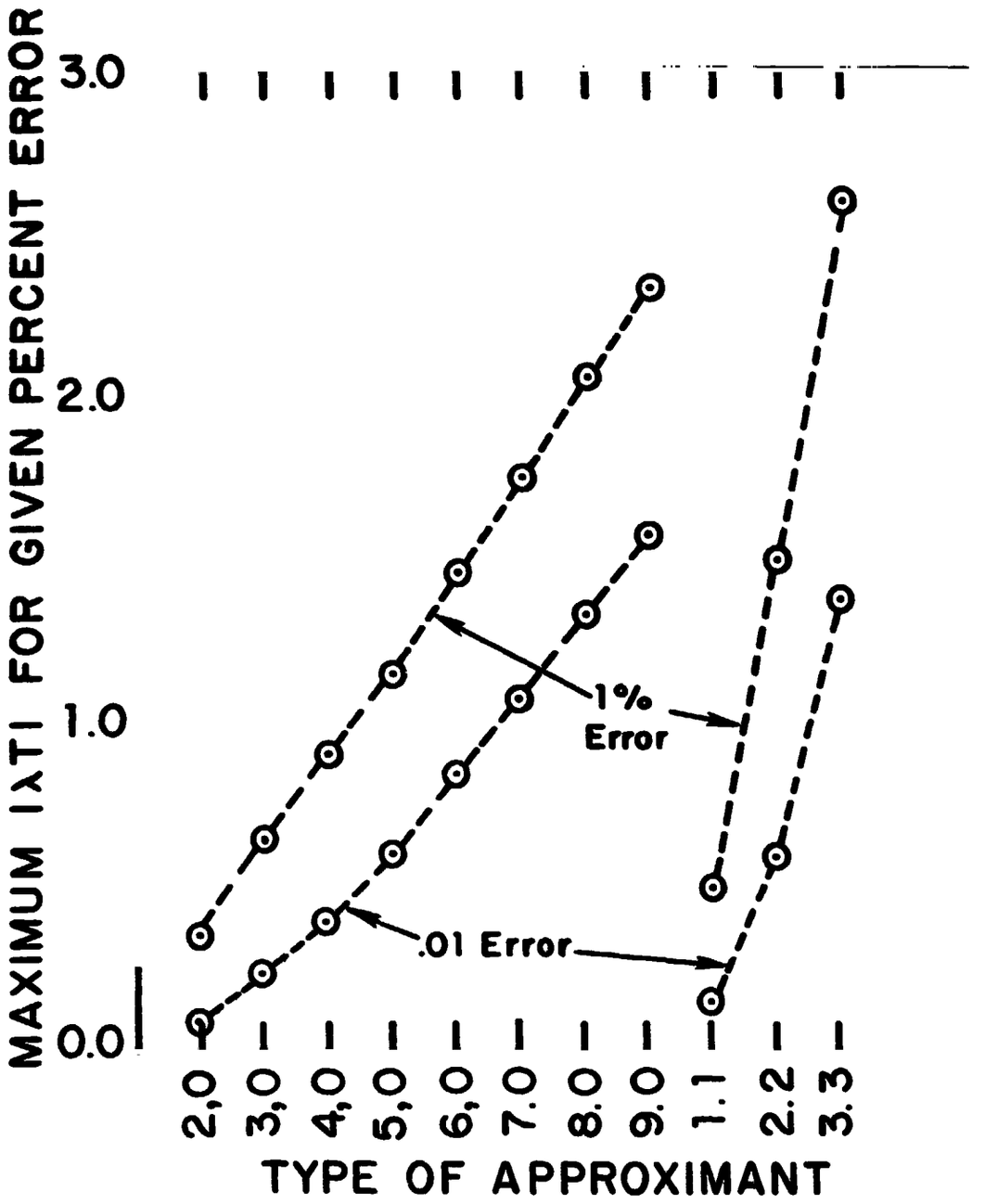


Figure 5.2 Maximum Allowable Values of $|\lambda T|$ as a Function of Type Approximant for a Given Error Bound

References

5.1 Barker, Bowles, and Williams, "Development and Application of a Local Linearization Algorithm for the Integration of Quaternion Rate Equations in Real Time Flight Simulation Problems," NASA TN 0-7347, Dec. 1973, p. 9.

5.2 Storer, Passive Network Synthesis, McGraw-Hill, 1957, p. 272.

5.3 Isaacson and Keller, Analysis of Numerical Methods, Wiley, 1966, p. 36.

5.4 Bellman, R., Introduction to Matrix Analysis, New York, McGraw-Hill, 1960, pp. 199-200.

Appendix A

Relation Between Matrix Series and Scalar Series

To obtain a measure of the error introduced by using approximations for the matrix exponential, one can transform to the normal coordinates, i.e.

$$\lambda = S^{-1}AS$$

and

$$\hat{x} = S^{-1}x \ .$$

It will be assumed that A has distinct eigenvalues. This is not overly restrictive since any matrix with repeated eigenvalues may be approximated arbitrarily closely by one with distinct eigenvalues [5.4].

The matrix exponential operating on a vector may be written as

$$\begin{aligned}
\exp(AT)x(t) &= SS^{-1}\exp(AT)SS^{-1}x(t) \\
&= SS^{-1}(I + AT + \tfrac{1}{2}A^2T^2 + \text{---})SS^{-1}x(t) \\
&= S(S^{-1}IS + S^{-1}AST + \tfrac{1}{2}S^{-1}ASS^{-1}AST^2 + \text{---})S^{-1}x(t) \\
&= S(I + \lambda T + \tfrac{1}{2}\lambda^2T^2 + \text{---})\hat{x}(t)
\end{aligned}$$

Likewise one can write the (1, 1) Padé approximants as

$$\begin{aligned}
(I - \tfrac{1}{2}AT)^{-1}(I + \tfrac{1}{2}AT)x(t) &= SS^{-1}(I - \tfrac{1}{2}AT)^{-1}(I + \tfrac{1}{2}AT)SS^{-1}x(t) \\
&= SS^{-1}(I + \tfrac{1}{2}AT + \tfrac{1}{4}A^2T^2 + \text{---})SS^{-1}(I + \tfrac{1}{2}AT)SS^{-1}x(t) \\
&= S(S^{-1}IS + \tfrac{1}{2}S^{-1}AST + \tfrac{1}{4}S^{-1}ASS^{-1}AST^2 + \text{---})(S^{-1}IS + \tfrac{1}{2}S^{-1}AST)S^{-1}x(t) \\
&= S(I + \tfrac{1}{2}\lambda T + \tfrac{1}{4}\lambda^2T^2 + \text{---})(I + \tfrac{1}{2}\lambda T)\hat{x}(t)
\end{aligned}$$

The other ordered Padé approximants follow in a like manner. In each case the approximant as well as the exact expression for exp(AT) is a diagonal matrix in the normal coordinates. Also each diagonal element is a function of one and only one eigenvalue. This fact makes it possible to compare different approximations to exp(AT) in the normal coordinates on an element by element basis. Also, although we have not proved it mathematically, our numerical results indicate that the maximum error occurs in that element corresponding to the eigenvalue of maximum value. Thus one can test for the accuracy of matrix polynomial representations of exp(AT) by testing the corresponding scalar polynomials and utilizing the eigenvalue of the A matrix having greatest magnitude.

Appendix B

Padé Approximants for H(T)

One further use for the Padé approximation may be considered here. It was shown above that if A^{-1} exists, then

$$H(T) = A^{-1}[\exp(AT) - I]B.$$

The A matrix may be singular, however, and no convenient closed-form expression can be used for H(T). For example, a zero eigenvalue in A will impose this restriction. However, H(T) may be written

$$H(T) = T(I + AT/2 + A^2T^2/3! + ---)B.$$

This series written in scalar form is

$$1 + w/2 + w^2/3! + ---$$

where $w = \lambda_m T$. This series has the scalar Padé approximants:

$$(1, 1) = (1 + w/6)/(1 - w/3); \quad |w| < 3$$

$$(2, 2) = (1 + w/10 + w^2/60)/(1 - 2w/5 + w^2/20); \quad |w| < 4.472$$

$$(3, 3) = (1 + w/14 + w^2/42 + w^3/840)/(1 - 3w/7 + w^2/14 - w^3/210); \quad |w| < 5.649$$

Use of these approximants parallels the discussion above for the matrix exponential.

VI. ANALYTICAL INTEGRATION OF STATE EQUATIONS, USING AN INTERPOLATED INPUT

One approach to solving differential equations discretely has been to use the state equations, along with polynomial approximations to the input signal. The coefficients of the polynomial are determined by the value of the input signal at the sample times. Once the polynomial is chosen, the state equations can be integrated analytically. One important feature of this approach is that the system is modeled exactly. The approximation is in sampling the input. The various polynomial fits to the input serve as interpolators.

Using a first order polynomial to represent the input, one finds that

$$X_N = e^{AT}X_{N-1} + \int_{(N-1)T}^{NT} e^{A(NT-t)}BU(t)dt \qquad (6.1)$$

can be approximated by

$$X_N = e^{AT}X_{N-1} + [-A^{-1} + \frac{1}{T}A^{-2}(e^{AT} - I)]BU_N + [A^{-1}e^{AT} - \frac{1}{T}A^{-2}(e^{AT} - I)]BU_{N-1} \qquad (6.2)$$

Expressions for the zeroth order polynomial fit and the 2nd order polynomial fit have been determined and are reported along with more details on the method in our Semi-Annual Report of September, 1975, on this same project, University Report No. EE-4041-101-75.

Because our efforts have been directed in other areas, we have not yet obtained test results on the accuracy of this technique.