Dynamic-error analysis of digital and combined analog-digital computer systems

by

ELMER G. GILBERT

The University of Michigan

FOREWORD

Except for very minor changes, the following paper is identical with a chapter of the notes used in connection with the one-week intensive course on hybrid computation, given July 1965 at the University of Michigan. Although the approach taken is in some respects novel, and there are several previously unpublished results in Section 5, the intent of the presentation is tutorial. There are four main areas of contribution: an introduction to the essentials of the z-transform; a review of basic integration formulas for digital computer solution of differential equations, and a z-transform analysis of their performance; the development of methods for the dynamic analysis of mixed-data systems (i.e., systems wherein both discrete- and continuous-time signals are present); and the application of these methods to the analysis of some typical hybrid computer loops. It is hoped that more extensive use of analytical methods such as those described will lead to better technical decisions in the selection of computing techniques for the solution of dynamical problems.

1 INTRODUCTION

Until recent years the analog computer has been the almost exclusive tool for the simulation of complex dynamical systems. The greatly increased speed of digital computers and demands for high accuracy in such fields as space technology have changed this situation; all-digital and combined analog-digital simulations are today not uncommon. In comparing the technological and economic aspects of various methods for simulation or solution of differential equations, there are two primary areas of consideration—speed and accuracy. With a given analog computer, the speed and accuracy of the parallel operating components are fixed. This means that it is possible to make some trade-off between the speed and accuracy of the overall computation, but there is a definite upper limit on accuracy. With a digital computer a more varied trade-off between speed and accuracy is feasible (*e.g.*, by the use of multiple-precision arithmetic), but there is an upper limit on speed having to do with the complexity of the problem being solved and the size, flexibility, and speed of the computer. If the digital computer has an adequate word size, the primary source of errors is the approximate representation of continuous-time functions by a sequence of samples. Methods for studying this error source, whether it occurs in a digital or combined analog-digital system, are the principal subject of this paper.

In all-digital systems, the dynamic errors are the result of truncation errors due to the integration method used. There are a variety of methods for treating these errors which can be found in books on numerical analysis. [3,7] The approach taken here, which is quite different, is to make use of the z-transform. This approach has two advantages: it gives a systems theory interpretation of the dynamic errors, and it may be conveniently extended to the analysis of combined analog-digital (hybrid) systems. Since the z-transform is effectively limited to linear systems, nonlinear systems must be treated by consideration of "small motions" through the use of linearized equations of motion; but this limitation is characteristic of most methods of error analysis. The goal of this paper is not to provide an exhaustive treatment of digital and analog-digital systems, but rather to acquaint the reader with general techniques for their analysis. It is hoped that the examples which are used as a vehicle for developing these techniques will also provide insight into the limitations, as well as the (better known) advantages, of the digital computer in dynamical computations.

The plan of the paper is as follows. In Section 2 we give a brief introduction to the theory of the z-transform. Section 3 applies the z-transform to the analysis of standard methods for integrating differential equations. In combined analog-digital systems the presence of data expressed in both sequence and time-function form significantly complicates the analysis. Section 4 is a resume of the needed theory, while Section 5 gives several examples of its application to the analysis of hybrid computer loops.

Much of the material in Sections 3 and 5 appeared in an earlier report by R. M. Howe.[4]

2 INTRODUCTION TO THE z-TRANSFORM

As will be seen in subsequent sections, the z-transform is an effective mathematical tool for analyzing systems which involve sequences of data points. In this section a brief introduction to the theory and application of the z-transform is given. For a fuller and more precise presentation see reference 1, or standard texts on the subject.

Data sequences which will be considered are of the form $f_0, f_1, \ldots, f_n, \ldots$. The index n frequently denotes equally spaced points in time, nT, where T is the sampling period. The values f_n may be sample values of a time function, or simply data points produced by a digital computer program. For compactness of notation the entire sequence $f_0, f_1, \ldots,$ will be denoted by $\{f_n\}$.

The z-transform of the data sequence $\{f_n\}$ is defined to be the following function of the complex variable z:

$$\mathcal{Z}\{f_n\} \overset{\Delta}{=} F^*(z) \overset{\Delta}{=} \sum_{n=0}^{\infty} f_n z^{-n} \tag{2.1}$$

If reasonable conditions are imposed on the data sequence $\{f_n\}$ and $|z|$ is sufficiently large, the series converges and $F^*(z)$ is analytic. When the series does converge (the sequence $\{f_n\}$ is *z-transformable*) knowledge of $F^*(z)$ implies knowledge of $\{f_n\}$. This fact is expressed in the *inversion integral*

$$f_n = \frac{1}{2\pi j} \oint_C z^{n-1} F^*(z)\, dz, \tag{2.2}$$

where C is a contour in the complex z-plane which must contain all the singularities of $F^*(z)$. Let us demonstrate the correspondence between $F^*(z)$ and $\{f_n\}$ by means of some examples.

Example 2.1: $f_n = a^n$. This is the data sequence equivalent of an exponential time function, *i.e.*, $\{f_n\}$ may be obtained by sampling an exponential time function. From the definition (2.1)

$$F^*(z) = \sum_{n=0}^{\infty} a^n z^{-n} = \sum_{n=0}^{\infty} (az^{-1})^n = \frac{1}{1 - az^{-1}} = \frac{z}{z - a}, \tag{2.3}$$

where $(1-x)^{-1} = 1 + x + x^2 + \cdots$ has been used to close the series. Notice that $F^*(z)$ is analytic for $|z| > a$.

Example 2.2: $f_n = \cos nb$.

Let $\cos nb = (1/2)\, e^{jnb} + (1/2)\, e^{-jnb}$.

Then

$$\begin{aligned} F^*(z) &= \frac{1}{2}\sum_{n=0}^{\infty} e^{jnb} z^{-n} + \frac{1}{2}\sum_{n=0}^{\infty} e^{-jnb} z^{-n} \\ &= \frac{1}{2}\sum_{n=0}^{\infty} (e^{jb} z^{-1})^n + \frac{1}{2}\sum_{n=0}^{\infty} (e^{-jb} z^{-1})^n \\ &= \frac{1}{2}\frac{z}{z - e^{jb}} + \frac{1}{2}\frac{z}{z - e^{-jb}} \\ &= \frac{z^2 - z\cos b}{z^2 - 2z\cos b + 1} \end{aligned} \tag{2.4}$$

By extending the techniques illustrated in these examples it is possible to derive many other transform pairs. Table 2.1 displays a few of the more common pairs.

Table 2.1 – Table of z-transforms

No.	$\{f_n\}$	$F^*(z)$
1	1	$\frac{z}{z-1}$
2	n	$\frac{z}{(z-1)^2}$
3	n^2	$\frac{z(z+1)}{(z-1)^3}$
4	n^3	$\frac{3z(z+1)}{(z-1)^4} + \frac{z(z+2)}{(z-1)^3}$
5	a^n	$\frac{z}{z-a}$
6	$(n+1)a^{n+1}$	$\frac{z^2 a}{(z-a)^2}$
7	$n\,a^n$	$\frac{z\,a}{(z-a)^2}$
8	$a^n \sin n\,b$	$\frac{za \sin b}{z^2 - z\,2a\cos b + a^2}$
9	$a^n \cos n\,b$	$\frac{z(z - a\cos\ b)}{z^2 - z\,2a\cos b + a^2}$

Often it is necessary to determine $\{f_n\}$ from $F^*(z)$. This can be done by evaluating the inversion integral (perhaps via the method of residues), but it is generally preferable to use a *power series expansion* or a *partial-fraction expansion*. Let us demonstrate these techniques by means of an example:

Example 2.3: Obtain the data sequence corresponding to

$$F^*(z) = \frac{z}{z^2 + 3z + 2} \tag{2.5}$$

The power series expansion method involves long division of $F^*(z)$ expressed as a rational function of z^{-1}:

$$1 + 3z^{-1} + 2z^{-2} \;\overline{)\; z^{-1} - 3z^{-2} + 7z^{-3} + \ldots}$$

Thus

$$F^*(z) = 0{\cdot}z^{-0} + 1{\cdot}z^{-1} - 3{\cdot}z^{-2} + 7{\cdot}z^{-3} + \ldots \tag{2.6}$$

and

$$f_0 = 0,\ f_1 = 1,\ f_2 = -3,\ f_3 = 7, \ldots \tag{2.7}$$

A partial-fraction expansion of $z^{-1}F^*(z)$ permits the use of table 2.1 in evaluating $\{f_n\}$. First,

$$z^{-1}F^*(z) = \frac{1}{z^2+3z+2} = \frac{1}{(z+1)(z+2)} = \frac{1}{z+1} + \frac{-1}{z+2} \tag{2.8}$$

Thus

$$F^*(z) = \frac{z}{z+1} - \frac{z}{z+2}. \tag{2.9}$$

Since the z-transform has the property of linearity ($\mathcal{Z}\{k_g g_n + k_h h_n\} = k_g \mathcal{Z}\{g_n\} + k_h \mathcal{Z}\{h_n\}$ the elements of $\{f_n\}$ can be obtained by summing the elements of the data sequences corresponding to $\frac{z}{z+1}$ and $\frac{-z}{z+2}$.

Therefore from line 5 of Table 2.1

$$f_n = (-1)^n - (-2)^n, n \geq 0 \tag{2.10}$$

The partial-fraction expansion method shows that the *form* of the data sequence corresponding to $F^*(z)$ is characterized by the location of the singularities of $F^*(z)$ in the complex plane. To be more specific, suppose that $F^*(z)$ is rational and that the singularities of $F^*(z)$ are simple poles at $z = z_1, z_2, \ldots z_N$. Then

$$F^*(z) = \frac{N(z)}{(z-z_1)(z-z_2)\ldots(z-z_N)} = z\left[\frac{a_1}{z-z_1} + \frac{a_2}{z-z_2} + \ldots + \frac{a_N}{z-z_N}\right] \tag{2.11}$$

where $N(z)$ is a polynomial in z and $a_1, a_2 \ldots a_N$ are dependent on $N(z)$. Assuming that $z_i \neq 0$, $i = 1, \ldots, N$, it follows that

$$f_n = a_1(z_1)^n + a_2(z_2)^n + \ldots + a_N(z_N)^n. \tag{2.12}$$

Thus the z_i determine the form, though not the amplitude, of the additive terms contributing to $\{f_n\}$.

To better appreciate the behavior of $\{z_1{}^n\}$ let us visualize $z_1{}^n$ as a sample taken from the complex exponential time function $e^{s_1 t} = e^{\sigma_1 t}e^{j\omega_1 t}$ where σ_1 determines the damping and ω_1 determines the frequency. If T is the sample period $(z_1)^n = e^{s_1 nT}$ (see figure 2.1 for case where $\omega_1 = 0$ and $\sigma_1 < 0$). Therefore

$$z_1 = e^{s_1 T}, |z_1| = e^{\sigma_1 T}, \angle z_1 = \omega_1 T \tag{2.13}$$

Thus for $|z_1| < 1$, $\sigma_1 < 0$ and the data sequence converges to zero as $n \to \infty$. Conversely, $|z_1| > 1$ implies a divergent data sequence. If $\angle z_1 = 0$, the sequence is nonoscillatory.

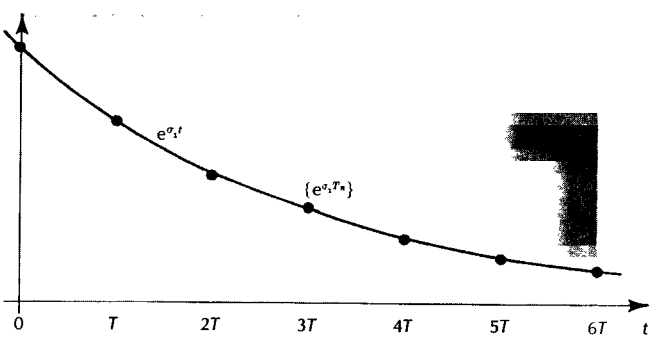
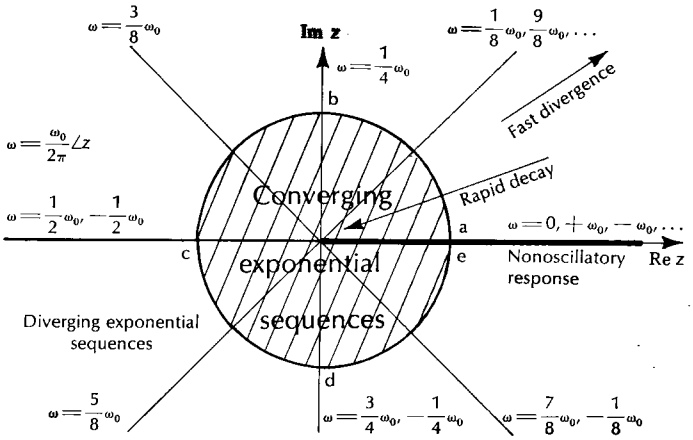
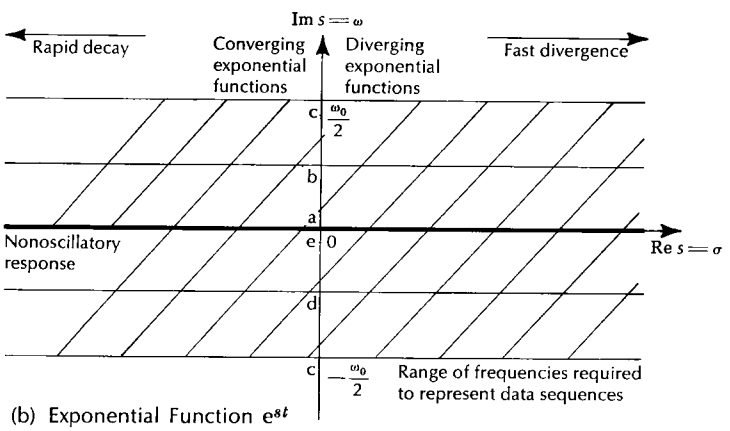


Figure 2.1 – Generation of data sequence by sampling of exponential time function

For $\angle z_1 \neq 0$, $\frac{\angle z_1}{T} = \frac{\angle z_1}{2\pi}\omega_0 = \omega_1$ represents the frequency of the complex time function from which the samples are taken ($\omega_0 = 2\pi/T$ is the sampling frequency in radians/second). Notice that frequencies in the range $-\omega_0/2 < \omega_1 \leq \omega_0/2$ are sufficient to represent any z_1. These results are summarized in figure 2.2.



(a) Exponential Data Sequence $\{z^n\}$



(b) Exponential Function e^{st}

Figure 2.2 – Complex z- and s-planes, showing regions for various classes of exponential response

In order to apply the z-transform to the problem of determining system response, it is necessary to consider certain shift properties of the z-transform. Let $\{f_{n-k}\} = \{g_n\}$, $k =$ positive integer, be the data sequence obtained by *delaying* the elements of $\{f_n\}$ by k steps (adopt the convention $f_n = 0$, $n < 0$):

$$g_0 = 0, g_1 = 0, \ldots, g_{k-1} = 0, \quad g_k = f_0, g_{k+1} = f_1, \ldots \tag{2.14}$$

Clearly

$$G^*(z) = f_0 z^{-k} + f_1 z^{-k-1} + \ldots = z^{-k}\sum_{n=0}^{\infty} f_n z^{-n} = z^{-k}F^*(z). \tag{2.15}$$

Thus we have obtained the

Delay Property: $\mathcal{Z}\{f_{n-k}\} = z^{-k}F^*(z)$, $k =$ positive integer.

Consider now the advanced data sequence $\{f_{n+k}\} = \{g_n\}$, $k =$ positive integer, that is

$$g_0 = f_k,\ g_1 = f_{k+1},\ \ldots \tag{2.16}$$

We can write

$$\begin{aligned} C^*(z) &= f_k + f_{k+1}z^{-1} + f_{k+2}z^{-2} + \ldots \\ &= z^k[f_kz^{-k} + f_{k+1}z^{-k-1} + f_{k+2}z^{-k-2} + \ldots] \\ &\quad + z^k[f_0 + f_1z^{-1} + \ldots + f_{k-1}z^{-k+1}] \\ &\quad - z^k[f_0 + f_1z^{-1} + \ldots + f_{k-1}z^{-k+1}] \\ &= z^k \sum_{n=0}^{\infty} f_nz^{-n} - [z^kf_0 + z^{k-1}f_1 + \ldots + zf_{k-1}]. \end{aligned} \tag{2.17}$$

This gives the

Advance property:

$$\mathcal{Z}\{f_{n+k}\} = z^kF^*(z) - [z^kf_0 - z^{k-1}f_1 + \ldots + zf_{k-1}],$$
$$k = \text{positive integer}$$

Let us use this last property to obtain the solution of a linear difference equation.

Example 2.4: Solution of the forced, second-order, linear difference equation

$$c_{n+2} + \frac{5}{6}c_{n+1} + \frac{1}{6}c_n = r_n,$$

$c_0,\ c_1 =$ specified initial values. (2.18)

Since this equation holds for any $n \geq 0$, we have

$$\{c_{n+2}\} + \frac{5}{6}\{c_{n+1}\} + \frac{1}{6}\{c_n\} = \{r_n\}. \tag{2.19}$$

Here $\{r_n\}$ should be interpreted as the input or forcing sequence and $\{c_n\}$ as the output or response sequence. Taking the z-transform of (2.19) by utilizing the advance property yields

$$\begin{aligned} z^2C^*(z) - [z^2c_0 + zc_1] + \frac{5}{6}zC^*(z) - \frac{5}{6}zc_0 \\ + \frac{1}{6}C^*(z) = R^*(z). \end{aligned} \tag{2.20}$$

This may be solved for $C^*(z)$,

$$\begin{aligned} C^*(z) = \frac{z^2 + \frac{5}{6}z}{\left(z+\frac{1}{2}\right)\left(z+\frac{1}{3}\right)}c_0 + \frac{z}{\left(z+\frac{1}{2}\right)\left(z+\frac{1}{3}\right)}c_1 \\ + \frac{1}{\left(z+\frac{1}{2}\right)\left(z+\frac{1}{3}\right)}R^*(z). \end{aligned} \tag{2.21}$$

The first two terms on the right side of (2.21) are initial condition or transient terms; the last term represents the forced part of the response. By using a partial-fraction expansion the reader may confirm that for a "step input," $r_n = 1$, $n \geq 0$, the response is

$$\begin{aligned} c_n = &\left[-2\left(-\frac{1}{2}\right)^n + 3\left(-\frac{1}{3}\right)^n\right]c_0 \\ &+ \left[-6\left(-\frac{1}{2}\right)^n + 6\left(-\frac{1}{3}\right)^n\right]c_1 \\ &+ \left[4\left(-\frac{1}{2}\right)^n - \frac{9}{2}\left(-\frac{1}{3}\right)^n + \frac{1}{2}\right]. \end{aligned} \tag{2.22}$$

If in Example 2.4 the system is initially at rest ($c_0 = c_1 = 0$) the z-transform of the response can be written

$$C^*(z) = H^*(z)R^*(z) \tag{2.23}$$

where $H^*(z) = (z + 1/2)^{-1}\ (z + 1/3)^{-1}$ can be interpreted as the *transfer function* of the system.

Let us pursue this interpretation further. Consider a system which processes an input data sequence $\{r_n\}$ by the formula

$$c_n = \sum_{m=0}^{\infty} h_mr_{n-m} \tag{2.24}$$

to produce an output data sequence $\{c_n\}$. This type of system which occurs commonly, we will call a *digital system*. Following well-established terminology, we say that $\{c_n\}$ is the *convolution* of the sequences $\{r_n\}$ and $\{h_n\}$. From

$$\begin{aligned} C^*(z) &= \sum_{n=0}^{\infty} c_nz^{-n} = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} h_mr_{n-m}z^{-n} \\ &= \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} h_mr_{n-m}z^{-n+m}z^{-m}, \end{aligned} \tag{2.25}$$

the substitution $k = n - m$, and $r_k = 0$ for $k < 0$, it is seen that

$$\begin{aligned} C^*(z) &= \sum_{k=-m}^{\infty}\sum_{m=0}^{\infty} h_mz^{-m}r_kz^{-k} \\ &= \left(\sum_{k=0}^{\infty} r_kz^{-k}\right)\left(\sum_{m=0}^{\infty} h_mz^{-m}\right) \end{aligned} \tag{2.26}$$

Letting

$$H^*(z) = \mathcal{Z}\{h_n\} \tag{2.27}$$

equation (2.26) can be written as equation (2.23). Thus for $c_0,\ c_1 = 0$ the system described by the difference equation (2.18) is a digital system whose output can be expressed by the convolution formula (2.24). The simplicity of (2.23) as compared with (2.24) is one of the reasons why the z-transform is an effective tool for analyzing systems with sampled data.

Many properties of digital-system response can be determined from $H^*(z)$. For example, a digital system is *stable* if $H^*(z)$ is analytic for $|z| \geq 1$. This seems reasonable in view of the results of figure 2.2. Also, by utilizing (2.24) it is easy to show that $r_n = e^{j\omega nT}$, $-\infty < n < +\infty$, gives $c_n = H^*(e^{j\omega T})e^{j\omega nT}$. Thus the magnitude and angle of $H^*(e^{j\omega T})$ represent the "gain" and "phase shift" of the digital system for an input which is obtained by sampling a sinusoidal time function with frequency ω.

3 APPLICATION OF THE z-TRANSFORM TO THE ANALYSIS OF INTEGRATION METHODS

When differential equations are solved on the digital computer, they must first be "approximated" by difference equations. The z-transform provides a method of investigating the dynamic errors introduced by this approximation. The differential equations considered here are linear, and of first and second order. By laborious calculations or the use of vector notation, results similar to those described may be obtained for linear systems of higher order. In any case, most complex dynamical systems include a number of linear first- and second-order loops. Thus the results of this section give considerable insight into the nature of the dynamic errors. Computer round-off errors, which may be a serious problem with high-order integration formulas such as Runge-Kutta, are a nonlinear effect and are not investigated.

3.1 *The Euler method, solution of a first-order differential equation*

The Euler method is the simplest and most naive approach for constructing the approximating difference equation. In it the solution of the differential equation is extended over an integration period by making the approximation that the rate-of-change of the solution is constant. Thus if the differential equation to be solved is

$$\dot{x} = f(x, r(t)), \quad x(0) \tag{3.1}$$

its approximating difference equation is

$$x_{n+1} = x_n + Tf(x_n, r_n) \tag{3.2}$$

where T is the (fixed) interval or step size and $r_n = r(nT)$. Taking $x_0 = x(0)$, successive application of equation (3.2) generates a sequence of points $\{x_n\}$, which, if T is sufficiently small, should approximate $\{\bar{x}(nT)\}$, where $\bar{x}(t)$ is the desired solution of (3.1). Let us now explore the nature of the approximation for the linear case where equations (3.1) and (3.2) become

$$\dot{x} = ax + br(t), \quad x(0) = x_0, \tag{3.3}$$

$$x_{n+1} = x_n + Tax_n + Tbr_n. \tag{3.4}$$

Viewing (3.4) as the sequence relationship $\{x_{n+1}\} = (1 + Ta)\{x_n\} + Tb\{r_n\}$, we obtain from the advance property of the z-transform

$$zX^*(z) - zx_0 = (1 + Ta)X^*(z) + TbR^*(z) \tag{3.5}$$

Thus

$$X^*(z) = \frac{zx_0}{z - 1 - Ta} + \frac{Tb}{z - 1 - Ta} R^*(z). \tag{3.6}$$

First we examine the unforced solution of (3.4), $R^*(z)=0$. Then reference to table 2.1 yields from (3.6)

$$x_n(1 + aT)^n x_0 \tag{3.7}$$

If this equation is written as

$$x_n = e^{\sigma T n} x_0 \tag{3.8}$$

σ represents the effective exponential constant obtained in the solution of (3.4). Since the unforced solution of (3.3) is $\bar{x}(t) = e^{at}x_0$, σ should ideally be equal to a. From (3.7) and (3.8), $e^{\sigma T} = 1 + aT$. Thus

$$\sigma = \frac{1}{T} \ln (1 + aT). \tag{3.9}$$

Using the power series $\ln(1+y) = y - (1/2)y^2 + (1/3)y^3 + \ldots$ it follows that for $aT << 1$

$$\sigma \cong a - \frac{1}{2} a^2 T = a\left(1 - \frac{1}{2} aT\right). \tag{3.10}$$

Therefore the attained exponential constant is too small by a fractional error $(1/2)aT$.

Now consider the response due to input forcing. A variety of different test inputs can be used to evaluate the solution error, but the sinusoidal input $r(t) = e^{j\omega t}$ gives the greatest insight. With $x_0 = 0$ equation (3.6) takes the form of equation (2.23) where the transfer function is

$$H^*(z) = \frac{Tb}{z - 1 - Ta} \tag{3.11}$$

For (3.3) the transfer function for sinusoidal inputs is $\bar{H}(j\omega) = b(j\omega - a)^{-1}$. Thus the magnitude and angle of

$$\bar{H}^{-1}(j\omega)H^*(e^{j\omega T}) = \frac{j\omega - a}{b} \frac{Tb}{e^{j\omega T} - 1 - Ta} \tag{3.12}$$

represent the gain and phase error introduced by (3.4). Equation (3.12) may be evaluated at any frequency ω. For example, $\bar{H}^{-1}(j0)H^*(e^{j0}) = 1$, and thus there is no error for the constant input $r(t) = e^{j0} = 1$. Another good test frequency is $\omega = a$, where $\bar{H}(j\omega)$ gives a phase shift of $-45°$ and the gain is down from $\bar{H}(j0)$ by $2^{-1/2}$. Using a power series in aT for e^{jaT}, we obtain

$$\bar{H}(ja)H^*(e^{jaT}) = 1 - \frac{(1 + j)}{4} \times \left(aT + j\frac{1}{3} a^2T^2 - \frac{1}{12} a^3T^3 + \ldots\right). \tag{3.13}$$

Thus the error is the order of aT and (3.4) is a good representation of (3.3) only if $aT << 1$. When $a = 0$ in (3.12), it is easy to show that $\bar{H}^{-1}(j\omega)H^*(e^{j\omega T}) = 1 + j(1/2)\omega T - (1/6)\omega^2T^2 + \cdots$. Thus when (3.3) corresponds to an integrator ($a = 0$), the error is small when $\omega T << 1$.

3.2 *The Euler method, solution of a second-order differential equation*

Second-order differential equations are introduced by considering the solution of the simultaneous first-order differential equations

$$\begin{aligned} \dot{x} &= f(x, y, r(t)), & x(0) &= x_0 \\ \dot{y} &= g(x, y, r(t)), & y(0) &= y_0. \end{aligned} \tag{3.14}$$

Following the pattern developed in Section 3.1, the Euler method yields

$$\begin{aligned} x_{n+1} &= x_n + Tf(x_n, y_n, r_n) \\ y_{n+1} &= y_n + Tg(x_n, y_n, r_n) \end{aligned} \tag{3.15}$$

The special problem to be analyzed here is the second-order system $\ddot{x} + 2\zeta\dot{x} + x = r(t)$, whose undamped natural frequency is one radian/second and whose damping ratio is ζ. By introducing $\dot{x} = y$ we obtain

$$\begin{aligned} \dot{x} &= y, & x(0) &= x_0 \\ \dot{y} &= -2\zeta y - x + r(t), & y(0) &= \dot{x}(0) = y_0, \end{aligned} \tag{3.16}$$

which is in the format of (3.14). Thus the Euler integration method gives

$$\begin{aligned} x_{n+1} &= x_n + Ty_n \\ y_{n+1} &= y_n - 2\zeta Ty_n - Tx_n + Tr_n. \end{aligned} \tag{3.17}$$

By using the z-transform on (3.17) we obtain

$$zX^*(z) - zx_0 = X^*(z) + TY^*(z)$$

$$zY^*(z) - zy_0 = (1 - 2\zeta T)Y^*(z) - TX^*(z) + TR^*(z)\,. \quad (3.18)$$

Solving these equations for $X^*(z)$ gives

$$X^*(z) = \frac{(z - 1 + 2\zeta T)zx_0 + zy_0T}{z^2 - 2(1 - \zeta T)z + 1 - 2\zeta T + T^2}$$

$$+ \frac{T^2}{z^2 - 2(1 - \zeta T)z + 1 - 2\zeta T + T^2} R^*(z) \quad (3.19)$$

To obtain the unforced solution we set $R^*(z) = 0$ and compare with lines 8 and 9 of table 2.1:

$$x_n = a^n\left[x_0 \cos nb + \frac{x_0\zeta + y_0}{\sqrt{1 - \zeta^2}} \sin nb \right], \quad (3.20)$$

where

$$a = \sqrt{1 - 2\zeta T + T^2} \quad \text{and} \quad b = \tan^{-1}\frac{T\sqrt{1 - \zeta^2}}{1 - \zeta T}.$$

This equation can be written as

$$x_n = e^{\sigma Tn}\left[x_0 \cos \omega Tn + \frac{x_0\zeta + y_0}{\sqrt{1 - \zeta^2}} \sin \omega Tn \right] \quad (3.21)$$

where σ and ω determine the damping and frequency of the solution. By inspecting the solution $\bar{x}(t)$ of (3.16), it can be shown that $\bar{x}(nT) = x_n$ if $\sigma = -\zeta$ and $\omega = \sqrt{1 - \zeta^2}$. Actually, comparison of (3.20) and (3.21) shows

$$\sigma = \frac{1}{2T} \ln (1 - 2\zeta T + T^2) = -\zeta$$

$$+ T(1 - 2\zeta^2) + T^2( ) + \cdots$$

$$\omega = \frac{1}{T} \tan^{-1}\frac{T\sqrt{1 - \zeta^2}}{1 - \zeta T}$$

$$= \sqrt{1 - \zeta^2}\left(1 + \zeta T - \frac{1}{3}T^2 + \frac{4}{3}\zeta^2T^2 + \cdots\right) \quad (3.22)$$

Thus the effective error in the damping ratio varies approximately as the first power of $T (T << 1)$, being too low for $\zeta < 1/\sqrt{2}$ and too high for $\zeta > 1/\sqrt{2}$. If $\zeta = 0$, the response diverges when it should remain bounded. For nonzero ζ, the attained frequency is fractionally high by ζT $(T << 1)$.

An analysis of the error for sinusoidal forcing can be carried out as in Section 3.1. For example $\bar{H}^{-1}(j0)H^*(e^{j0}) = 1$, and again there is no error for a constant input. For $0 < \omega << 1/T$, it is easy to show that the error is the order of T.

3.3 *The Heun method, solution of a first-order differential equation*

The Heun method achieves better accuracy than the Euler method by working with data at both ends of the integration interval. It converts the differential equation (3.1) to the difference equation

$$x_{n+1} = x_n + \frac{T}{2} f(x_n, r_n) + \frac{T}{2} f(x_n + Tf(x_n, r_n), r_{n+1}) \quad (3.23)$$

Thus for (3.3) we obtain

$$x_{n+1} = x_n + Tax_n + \frac{1}{2}T^2ax_n + \frac{1}{2}Tbr_n +$$

$$\frac{1}{2}Tbr_{n+1} + \frac{1}{2}T^2bar_n \quad (3.24)$$

and hence from the z-transform

$$X^*(z) = \frac{zx_0}{z - 1 - Ta - \frac{1}{2}T^2a^2} +$$

$$\frac{\frac{1}{2}Tb(z + 1 + Ta)}{z - 1 - Ta - \frac{1}{2}T^2a^2} R^*(z). \quad (3.25)$$

Consider now the solution of (3.24) for $r_n = 0$, $n \geq 0$. From table 2.1 and (3.25)

$$x_n = \left(1 + Ta + \frac{1}{2}T^2a^2\right)^n x_0 \quad (3.26)$$

Comparing (3.26) and (3.8) as in Section 3.1 yields

$$\sigma = \frac{1}{T} \ln\left(1 + Ta + \frac{1}{2}T^2a^2\right)$$

$$= a\left(1 - \frac{1}{6}T^2a^2 + \cdot\cdot\cdot\right). \quad (3.27)$$

Thus the attained exponential constant σ is too small by a fractional error $(1/6)\,T^2a^2 (Ta << 1)$, a great improvement over the $(1/2)Ta$ obtained with the Euler method.

Writing the last term in (3.25) as $H^*(z)R^*(z)$, it is seen that

$$H^*(z) = \frac{1}{2}Tb\left(z + 1 + Ta\right)\left(z - 1 - Ta - \frac{1}{2}T^2a^2\right)^{-1}.$$

Thus the error for sinusoidal forcing may be determined from $H^*(e^{j\omega T})$. As before $\bar{H}^{-1}(j0)H^*(e^{j0}) = 1$. Using a power series expansion for $e^{j\omega T}$, it can be shown that

$$\bar{H}^{-1}(ja)H^*(e^{jaT}) = 1 - \frac{2 + 7j}{12}(aT)^2 + \cdots \quad (3.28)$$

Comparing this with (3.13), it is seen that the error term is the order of $(aT)^2$ rather than aT. Thus the error for sinusoidal forcing is much less than that obtained with the Euler method. If in (3.3) $a = 0$, it is easy to show that $\bar{H}^{-1}(j\omega)H^*(e^{j\omega T}) = 1 - (1/12)\,\omega^2T^2 + \cdots$ where the omitted terms are of order $(T\omega)^4$ and higher. Thus when (3.3) is an integrator, a small sinusoidal response error will result for $(\omega T)^2 << 1$.

3.4 *The Heun method, solution of a second-order differential equation*

The application of the Heun method to (3.14) yields

$$x_{n+1} = x_n + \frac{T}{2} f(x_n, y_n, r_n) + \frac{T}{2} \times f(x_n + Tf(x_n, y_n, r_n), y_n + Tg(x_n, y_n, r_n), r_{n+1})$$

$$y_{n+1} = y_n + \frac{T}{2} g(x_n, y_n, r_n) + \frac{T}{2} \times g(x_n + Tf(x_n, y_n, r_n), y_n + Tg(x_n, y_n, r_n), r_{n+1}) \tag{3.29}$$

After some manipulation, the application of (3.29) to (3.16) gives

$$x_{n+1} = \left(1 - \frac{T^2}{2}\right) x_n + (T - T^2\zeta)\, y_n + \frac{1}{2} T^2 r_n$$

$$y_{n+1} = (-T + T^2\zeta)\, x_n + \left(1 - 2T\zeta + 2T^2\zeta^2 - \frac{1}{2}T^2\right) y_n + \frac{1}{2} T(r_n + r_{n+1}) - \frac{1}{2} T^2\zeta r_n \tag{3.30}$$

To simplify further developments it will be assumed that $\zeta = 0$. Then it follows from the z-transform, (3.30), and the elimination of $Y^*(z)$ that

$$X^*(z) = \frac{\left(z - 1 + \frac{1}{2}T^2\right) z x_0 + T z y_0}{\left(z - 1 + \frac{1}{2}T^2\right)^2 + T^2} + \frac{T^2\left(z + \frac{1}{4}T^2\right)}{\left(z - 1 + \frac{1}{2}T^2\right)^2 + T^2} R^*(z)\,. \tag{3.31}$$

Again we examine the unforced case, $R^*(z) = 0$. By comparing the first term in (3.31) with lines 8 and 9 of Table 2.1, it is seen that

$$x_n = a^n [x_0 \cos nb + y_0 \sin nb], \tag{3.32}$$

where

$$a = \sqrt{1 + \frac{1}{4}T^4} \quad \text{and} \quad b = \tan^{-1} \frac{T}{1 - \frac{1}{2}T}.$$

Ideally x_n should be given by (3.21) (with $\zeta = 0$), where $\sigma = 0$ and $\omega = 1$. Actually, comparison of (3.21) and (3.32) gives

$$\sigma = \frac{1}{2T} \ln\left(1 + \frac{1}{4}T^4\right) = \frac{1}{8}T^3 - \frac{1}{64}T^7 + \cdots$$

$$\omega = \frac{1}{T} \tan^{-1} \frac{T}{1 - \frac{1}{2}T^2} = 1 + \frac{1}{6}T^2 + \cdots\,. \tag{3.33}$$

Thus a very notable improvement over the Euler method is noted (see equations (3.22)).

3.5 *The Runge-Kutta method, solution of a first-order differential equation*

The Euler and Heun methods discussed in the previous sections lead to difference equations which are examples of a family of formulas called Runge-Kutta formulas. [3,7] As these formulas become more complex, they tend to produce more accurate results. In this section we examine one of the more commonly used methods which takes the name of the family of formulas.

With the Runge-Kutta method the difference equation corresponding to (3.1) becomes

$$x_{n+1} = x_n + T\phi(x_n, r_n, r_{n+1/2}, r_{n+1}) \tag{3.34}$$

The function ϕ (which is called the *increment function*) is sufficiently complex that it is best defined in terms of the following set of formulas:

$$\phi = \frac{1}{6}[k_a + 2k_b + 2k_c + k_d]$$

$$k_a = f(x_n, r_n)$$

$$k_b = f\left(x_n + \frac{1}{2}Tk_a, r_{n+1/2}\right)$$

$$k_c = f\left(x_n + \frac{1}{2}Tk_b, r_{n+1/2}\right)$$

$$k_d = f(x_n + Tk_c, r_{n+1})\,. \tag{3.35}$$

If (3.34) and (3.35) are applied to (3.3), after some manipulation it follows that

$$x_{n+1} = \left[1 + Ta + \frac{1}{2}(Ta)^2 + \frac{1}{6}(Ta)^3 + \frac{1}{24}(Ta)^4\right] x_n + \frac{1}{6}T\left[1 + Ta + \frac{1}{2}(Ta)^2 + \frac{1}{4}(Ta)^3\right] br_n + \frac{1}{6}T\left[4 + 2Ta + \frac{1}{2}(Ta)^2\right] br_{n+1/2} + \frac{1}{6}Tbr_{n+1} \tag{3.36}$$

Since the basic sampling period on the input is $(1/2)T$ (half the sampling period for the response), it should be the basic period if we wish to obtain the transfer function by the z-transform. That is, we should replace n by $2m$, x_n by $\tilde{x}_m$, x_{n+1} by $\tilde{x}_{m+2}$, r_n by $\tilde{r}_m = r(m(1/2)T)$, $r_{n+1/2}$ by $\tilde{r}_{m+1}$, and r_{n+1} by $\tilde{r}_{m+2}$. Then $\tilde{x}_m$, evaluated for $m =$ even integers, gives x_n. In order to avoid this complication, we will consider (3.36) only for the unforced case. Then z-transform may be applied to (3.36) in the customary fashion to yield

$$X^*(z) = \frac{z x_0}{z - 1 - Ta - \frac{1}{2}(Ta)^2 - \frac{1}{6}(Ta)^3 - \frac{1}{24}(Ta)^4} \tag{3.37}$$

The above result and reference to table 2.1 shows that

$$x_n = \left[1 + (Ta) + \frac{1}{2}(Ta)^2 + \frac{1}{6}(Ta)^3 + \frac{1}{24}(Ta)^4\right]^n x_0\,. \tag{3.38}$$

Following the approach taken in previous sections

$$\sigma=\frac{1}{T}\ln\left[1+(Ta)+\frac{1}{2}(Ta)^2+\frac{1}{6}(Ta)^3+\frac{1}{24}(Ta)^4\right], \tag{3.39}$$

which can be written

$$\sigma=\frac{1}{T}\ln\left[e^{aT}-\frac{1}{5!}(Ta)^5+\cdots\right]=a+\frac{1}{T}$$
$$\times\ln\left[1-\frac{1}{5!}(Ta)^5e^{-aT}+\cdots\right]=a\left[1-\frac{1}{5!}(Ta)^4\cdots\right]. \tag{3.40}$$

Thus the fractional error in the attained exponential constant is only $(1/120)\,(Ta)^4$.

The steps in evaluating formulas (3.35) may require considerable computer time. It would be much quicker to solve directly the equation $x_{n+1}=\alpha x_n+\beta_0 r_n+\beta_1 r_{n+1/2}+\beta_2 r_{n+1}$, where α, β_0, β_1, and β_2 are obtained from (3.36). Of course, this simplification is feasible only if (3.1) is linear.

3.6 *Multistep methods*

The methods of Euler, Heun, and Runge-Kutta are examples of one-step methods, where the solution at index $n+1$ can be determined from the solution at index n. For multistep methods the solution at index n and still earlier indices is required. This section examines briefly multistep methods of the *linear* type.

For the first-order differential equation (3.1), the linear multistep method yields

$$x_{n+k}+\alpha_1 x_{n+k-1}+\cdots+\alpha_k x_n$$
$$=T(\beta_0 f_{n+k}+\beta_1 f_{n+k-1}+\cdots+\beta_k f_n), \tag{3.41}$$

where

$$f_n=f(x_n, r_n) \tag{3.42}$$

Notice that for the *first-order* differential equation this *k-step formula* is a *kth order* difference equation. Given $x_0, x_1, \ldots x_{k-1}$ equation (3.41) can be solved for x_k, and then successively for $x_{k+1}, \cdots$. If the formula is *closed* ($\beta_0 \neq 0$) and $f(x, r)$ is nonlinear, (3.41) may present some difficulty in that x_{n+k} is contained in f_{n+1}, and therefore its determination from $x_{n+k-1}, \ldots, x_n$ involves the solution of a nonlinear equation. Various predictor-corrector schemes are available for solving this problem. The k-step formula also has a "start-up problem" in that one must know more than x_0 to begin. Usually $x_1, \ldots x_{k-1}$ are supplied from x_0 by $k-1$ iterations of a one-step method such as Runge-Kutta.

For the linear differential equation (3.3), the k-step formula yields

$$x_{n+k}+\alpha_1 x_{n+k-1}+\cdots+\alpha_k x_n=Ta$$
$$\times(\beta_0 x_{n+k}+\cdots+\beta_k x_n)+Tb(\beta_0 r_{n+k}+\cdots+\beta_k r_n) \tag{3.43}$$

In order to avoid the complexity of treating initial condition terms, let us assume hereafter that $x_0, x_1, \ldots, x_{k-1}$ and $r_0, r_1, \ldots r_{k-1}$ are all zero. Then the z-transform of (3.43) gives

$$(z^k+\alpha_1 z^{k-1}+\cdots+\alpha_k)X^*(z)=T$$
$$\times(\beta_0 z^k+\cdots+\beta_k)[aX^*(z)+bR^*(z)], \tag{3.44}$$

which by introducing

$$M(z)=\frac{z^k+\alpha_1 z^{k-1}+\cdots+\alpha_k}{\beta_0 z^k+\cdots+\beta_k} \tag{3.45}$$

can be written as

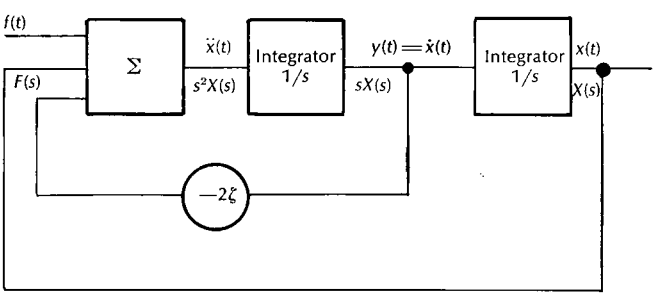
$$X^*(z)=\frac{b}{T^{-1}M(z)-a}R^*(z)=H^*(z)R^*(z) \tag{3.46}$$

It is interesting to note that $H^*(z)$ can be written in terms of the transfer function, $\bar{H}(s)=b(s-a)^{-1}$, corresponding to (3.3) by replacing s by $T^{-1}M(z)$, *i.e.*, $H^*(z)=\bar{H}(T^{-1}M(z))$. This remarkably simple result turns out to be true for linear systems of all orders! Thus for the second-order system (3.16)

$$H^*(z)=\bar{H}(T^{-1}M(z))=\frac{1}{T^{-2}M^2(z)+2\zeta T^{-1}M(z)+1} \tag{3.47}$$

Figure 3.1 shows a block diagram which indicates the correspondence between $\bar{H}(s)$ and $H^*(z)$ for the second-order system.

(a) Block diagram representation of differential equation



(b) Block diagram representation of corresponding difference equation

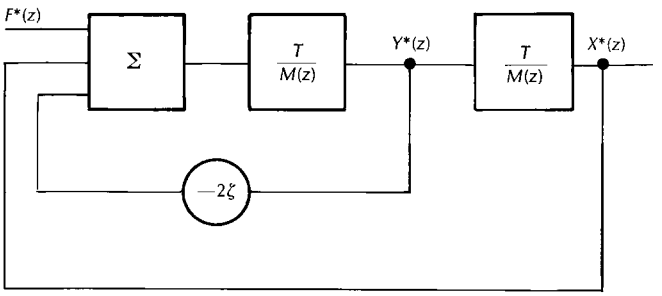


Figure 3.1 – Correspondence between s and $T^{-1}M(z)$ for second-order linear system

Let us now consider some specific realizations of (3.43). For $k=1$, $\alpha_1=-1$, $\beta_0=0$, $\beta_1=1$, we have the Euler method. The reader may verify (3.11) and the last term of (3.19) by the substitution rule given in the preceding paragraph. It turns out that similar substitution rules do not exist for the more elaborate single-step methods such as the Heun and Runge-Kutta methods.

For $k=2$, $\alpha_1=0$, $\alpha_2=-1$, $\beta_0=0$, $\beta_1=2$, $\beta_2=0$, we have the two-step Nystrom method. Substituting

$$T^{-1}M(z)=\frac{z^2-1}{2Tz} \tag{3.48}$$

for s in $\overline{H}(s)=b(s-a)^{-1}$, we obtain $H^*(z)$ corresponding to the differential equation (3.3):

$$H^*(z)=\frac{2Tbz}{z^2-1-2Taz}=\frac{2Tbz}{(z-z_1)(z-z_2)} \tag{3.49}$$

where

$$z_1=Ta+\sqrt{1+(Ta)^2}=1+Ta+\frac{1}{2}(Ta)^2-\frac{1}{8}(Ta)^4+\cdot$$

$$z_2=Ta-\sqrt{1+(Ta)^2}=-1+Ta-\frac{1}{2}(Ta)^2+\frac{1}{8}(Ta)^4+ \tag{3.50}$$

Recalling equations (2.11), (2.12), and (2.13) and letting $X^*(z)=H^*(z)R^*(z)$, it is seen that one response term of $\{x_n\}$ will have the form $a_1(z_1)^n=a_1(e^{s_1T})^n$. Since

$$s_1=\sigma_1=\frac{1}{T}\ln z_1=a\left(1-\frac{1}{6}(Ta)^2+\frac{3}{40}(Ta)^4+\cdots\right), \tag{3.51}$$

it is seen that $a_1(z_1)^n$ corresponds closely to samples taken from the expected transient response term in the solution of (3.3), ce^{-at}. Thus the attained exponential constant σ_1 approximates a within a fractional error of $1/6(Ta)^2$. This is the same result obtained by Heun's method. The response term $a_2(z_2)^n$ is undesired and appears because (3.43) is a second-order difference equation acting as an "approximation" of a first-order system. If $a<0$, that is the transient response of (3.3) is damped, $z_2\cong -1+Ta$ has magnitude greater than one, and $(z_2)^n\to\infty$ as $n\to\infty$. In this case the method is said to be *weakly unstable* (weak because $|z_2|\to 1$ as $T\to 0$) and will for sufficiently large n give inaccurate results. If a_2 is very small (as it often is)this may not be a problem.

The sinusoidal response errors for the Nystrom method may be evaluated as in previous sections, with the following results:

$$\overline{H}^{-1}(j0)H^*(e^{j0})=1,$$

$$\overline{H}^{-1}(ja)H^*(e^{jaT})=1+\frac{1-j}{2}a^2T^2+\cdots. \tag{3.52}$$

Thus the error here is comparable with that obtained for Heun's method, (3.28).

Similar results are obtained when the Nystrom method is applied to systems of higher order. If a linear differential equation has a characteristic root $\overline{\lambda}$, *i.e.*, $\overline{H}(s)$ has a pole at $s=\overline{\lambda}$, the substitution rule implies $H^*(z)$ has two poles at $\overline{z}_1$ and $\overline{z}_2$ where $(\overline{z}_i^2-1)[2\overline{z}_iT]^{-1}=\overline{\lambda}$ or

$$\overline{z}_1=1+T\overline{\lambda}+\frac{1}{2}(T\overline{\lambda})^2-\frac{1}{8}(T\overline{\lambda})^4+\cdots$$

$$\overline{z}_2=-1+T\overline{\lambda}-\frac{1}{2}(T\overline{\lambda})^2+\frac{1}{8}(T\overline{\lambda})^4+\cdots. \tag{3.53}$$

The pole at z_1 is the desired one in that

$$\frac{1}{T}\ln\overline{z}_1=\overline{\lambda}\left[1-\frac{1}{6}(T\overline{\lambda})^2+\frac{3}{40}(T\lambda)^4+\cdots\right] \tag{3.54}$$

approximates $\overline{\lambda}$. For example, the differential equation $\ddot{x}+x=r(t)$ leads to $\overline{H}(s)=(s^2+1)^{-1}$ which has poles at $\pm j$. Thus $H^*(z)$ has four poles, the desired two being at $\pm j[1+(1/6)T^2+(3/40)T^4+\cdots]$. These values indicate that the approximate system has the desired damping of zero, but that the natural frequency is $1+(1/6)T^2+(3/40)T^4+\cdots$ rather than 1. Note that the undesired poles of $H^*(z)$ are at $-1\pm jT+\cdots$ and (for $T<<1$) are *outside* the unit circle in the z-plane. Thus for this system the Nystrom method is weakly unstable.

Finally consider the two-step Simpson-Milne method where $k=2$, $\alpha_1=0$, $\alpha_2=-1$, $\beta_0=1/3$, $\beta_1=4/3$, $\beta_2=1/3$.
Thus

$$M(z)=\frac{3(z^2-1)}{z^2+4z+1}, \tag{3.55}$$

and it is a simple exercise to derive the results corresponding to those obtained for the Nystrom method. In particular, corresponding to (3.50) and (3.51), we have

$$z_1=1+Ta+\frac{1}{2}(Ta)^2+\frac{1}{6}(Ta)^3+\frac{1}{24}(Ta)^4+\cdots$$

$$z_2=-1+Ta-\frac{1}{2}(Ta)^2+\cdots \tag{3.56}$$

$$s_1=a\left[1-\frac{13}{72}(Ta)^4+\cdots\right] \tag{3.57}$$

Thus the fractional error in the attained exponential constant is improved from $(1/6)(Ta)^2$ to $(13/72)(Ta)^4$ (for $Ta<<1$). Also for $a<0$, $|z_2|>1$, and the method is weakly unstable. For the second-order system, the attained damping is again zero, while the natural frequency is $1-13/72\,T^4+\cdots$.

3.7 *Derivation of difference equations by quadrature formulas*

For any linear differential equation, it is possible to express the solution in closed form. By applying quadrature integration formulas to such solution expressions, a family of difference equations may be derived which are different from those discussed in previous sections. To keep the presentation reasonably brief we will consider only (3.3),

although it is possible to derive similar results for time-varying *linear* differential equations of any order.

Equation (3.3) has the closed-form solution

$$\bar{x}(t) = e^{at}\left[x_0 + \int_0^t e^{-a\sigma} br(\sigma)d\sigma\right] \tag{3.58}$$

Let us obtain an approximation to $\bar{x}(T)$ by using a quadrature formula for the integral on the right side of (3.58). A simple quadrature integration formula can be obtained by letting $r(t)$ be approximated by

$$r_0 = r(0) \quad \text{for } 0 \leq t < T.$$

Then

$$\bar{x}(T) \cong x_1 = e^{aT}\left[x_0 + \int_0^T e^{-a\sigma}\, d\sigma b r_0\right], \tag{3.59}$$

which when integrated gives

$$x_1 = e^{aT}x_0 + a^{-1}\left[e^{aT} - 1\right]br_0. \tag{3.60}$$

This suggests the difference equation,

$$x_{n+1} = e^{aT}x_n + a^{-1}\left[e^{aT} - 1\right]br_n, \tag{3.61}$$

as a method for obtaining an approximate solution of (3.3). Let us examine (3.61) by methods of the previous sections.

The z-transform of (3.61) gives

$$zX^*(z) - zx_0 = e^{aT}X^*(z) + a^{-1}\left[e^{aT} - 1\right]bR^*(z) \tag{3.62}$$

Thus

$$X^*(z) = \frac{zx_0}{z - e^{aT}} + \frac{a^{-1}\left[e^{aT} - 1\right]b}{z - e^{aT}}R^*(z) \tag{3.63}$$

For the unforced case, $R^*(z) = 0$, table 2.1 yields

$$x_n = e^{aTn}x_0 \tag{3.64}$$

Therefore $x_n = \bar{x}(nT)$ and there is *no solution error*!

Before becoming too elated about the desirability of (3.61), let us consider the response with sinusoidal forcing. Noting that $H^*(z) = a^{-1}\left[e^{aT} - 1\right]b(z - e^{aT})^{-1}$ it is easy to show that

$$\bar{H}^{-1}(j0)H^*(e^{j0}) = 1$$

$$\bar{H}^{-1}(ja)H^*(e^{jaT}) = 1 + \frac{1}{2}aT + \cdots \tag{3.65}$$

Thus the error is comparable to that attained with the Euler method. From the manner in which (3.61) was derived, it is clear that there are inputs for which $x_n = \bar{x}(nT)$. For instance, suppose $r(t)$ is the unit step at $t = 0$, *i.e.*, $r(t) = 1$ for $t \geq 0$, and $r(t) = 0$ for $t < 0$. Then the approximation made for $r(\sigma)$ in (3.58) becomes exact. However, if the step occurs at some other time, such as $t = .1T$, the error may be quite large.

We can obtain an improvement over (3.61) by using a better quadrature formula. For $0 \leq t < T$, let $r(\sigma)$ in (3.58) be approximated by $r_0 + T^{-1}(r_1 - r_0)\sigma$. Then

$$\bar{x}(T) \cong x_1 = e^{aT}\left[x_0 + \int_0^T e^{-a\sigma}d\sigma br_0 + T^{-1}\int_0^T e^{-a\sigma}\sigma d\sigma b(r_1 - r_0)\right]. \tag{3.66}$$

Once the indicated integrations are performed:

$$x_1 = e^{aT}x_0 + \left[a^{-1}e^{aT} + T^{-1}a^{-2}(1 - e^{aT})\right]br_0 + \left[-a^{-1} - T^{-1}a^{-2}(1 - e^{aT})\right]br_1 \tag{3.67}$$

Thus we use the difference equation which is obtained from (3.67) by augmenting the indices by n. For $R^*(z) = 0$ we again have (3.64), and there is no error for the unforced case.

With sinusoidal forcing it can be shown that

$$\bar{H}^{-1}(j0)H^*(e^{j0}) = 1$$

$$\bar{H}^{-1}(ja)H^*(e^{ja}) = 1 - \frac{3 + 2j}{6}(aT)^2 + \cdots. \tag{3.68}$$

Therefore the errors are comparable to those obtained with Heun's method and the Nystrom method. For any input $r(t)$ which is piecewise linear with jumps, and jumps in slope only at $t = nT$, $n = 0, 1, \ldots$, there will be no solution error. In practical applications, however, there is no reason to believe that the input will have this form.

Higher-order difference equations of the above type can be derived. For instance, by approximating $r(\sigma)$ by a cubic, a difference equation which compares favorably with the Runge-Kutta equation (3.38) may be derived. However, in this case the fact that the unforced solution has no error is not of great importance because the Runge-Kutta method produces an unforced solution of very high accuracy.

3.8 *Summary and remarks*

Many of the results obtained in previous sections are summarized along with a few more in table 3.1. The values shown are approximate in that additive terms of higher powers in T are omitted.

Let us compare the integration methods with respect to their performance on the period of the undamped second-order differential equation. To keep the error in period below 1 part in 10^4 with the Runge-Kutta method, it is necessary to have $T^4 < 120 \times 10^{-4}$ or $T < .33$. This corresponds to $2\pi/.33 \cong 19$ points per cycle. For the methods of Heun, Nystrom, and Simpson-Milne the corresponding numbers are, respectively: 256, 256, and 41.

It should be pointed out that there are many other factors besides those above which enter into the choice of a method. For example, the complex operations required at each step in the Runge-Kutta method may lead to large solution errors because of computer round-off errors. Thus from a practical point of view, the Simpson-Milne formula may be more satisfactory (provided the integration period is sufficiently short so that the instability of the method is not a factor).

Another important factor which is frequently overlooked is the required "smoothness" of the differential equations being solved. For example, the Runge-Kutta formula for "approximating" the first-order differential equation (3.1) is derived under the assumption that

$$\frac{\partial^4}{\partial x^4}f(x, r(t)) \quad \text{and} \quad \frac{\partial^4}{\partial t^4}f(x, r(t))$$

are continuous in x and t. [3,7] In many practical problems, such as an on-off control system, $f(x, r(t))$ may be *discontinuous* in t and even the Euler method is subject to doubt.

To explore this difficulty further, consider the solution of (3.3) where $b = -a$ and $r(t)$ is the unit step function at $t = 0$. The actual solution is

$$\begin{aligned} \overline{x}(t) &= 0, \qquad t < 0 \\ &= 1 - e^{at}, \quad t \geq 0. \end{aligned} \tag{3.69}$$

The corresponding solutions obtained from the difference equation representations of (3.3) are:

Euler: $$x_n = 1 - [1 + aT]^n \tag{3.70}$$

Heun: $$x_n \cong 1 - \left(1 + \frac{1}{2}Ta\right) \times \left[1 + aT + \frac{1}{2}(aT)^2\right]^n \tag{3.71}$$

Runge-Kutta: $$x_n \cong 1 - \left(1 + \frac{1}{6}Ta\right)\left[1 + aT + \frac{1}{2}(aT)^2 + \frac{1}{6}(aT)^3 + \frac{1}{24}(aT)^4\right]^n \tag{3.72}$$

Nystrom: $$x_n \cong 1 - \left(1 + \frac{1}{2}Ta\right)[z_1]^n + \frac{1}{2}Ta\,[z_2]^n, \tag{3.73}$$

where for the coefficients of $[\ ]^n$ additive terms of order $(Ta)^2$ and higher have been omitted. Thus for small n the errors produced by *all methods* are the same order of magnitude. For large n the Nystrom method is particularly bad for $a < 0$ because the term $1/2\,Ta[z_2]^n = 1/2\,Ta[-1 + aT + \cdots]^n$, which is initially large, grows rapidly. The above results may become even worse if the step is not applied at $t = 0$, since the sampling of $r(t)$ means that the position of the jump is uncertain within a timing error of up to $T(T/2$ in the case of Runge-Kutta).

Perhaps a less artificial method for evaluating the errors present with rapidly varying inputs is to use sinusoidal response. For the first-order system and $\omega = a$ it has been observed [see (3.13), (3.28), (3.52)] that errors are smaller for the better integration methods. Table 3.1 shows some additional results for the first-order system, where $\omega = (1/4)\omega_0, (1/2)\omega_0$, and $\omega_0 = 2\pi/T$ is the sampling frequency associated with the integration-formula step size. It has been assumed that $(aT) << 1$, and terms of higher order in aT than those given have been omitted. In every case the values $\overline{H}^{-1}(j\omega)H^*(e^{j\omega T}) - 1$ shown depart considerably from the desired value of zero. As might be suspected from the stability result of the previous paragraph, the Nystrom method gives particularly bad results.

Problems of the type described above can sometimes be detected during the numerical integration by evaluating residual terms generated by the integration formulas. In such cases the step size T may be adjusted automatically to bring the errors down to an acceptable level.

Table 3-1 – Approximate results for various integration methods

Integration Method	First-Order Equation (3.3)	Undamped Second-Order Equation (3.16)		First-Order Equation (3.3) $\overline{H}^{-1}(j\omega)H^*(e^{j\omega T}) - 1$		
	Fractional Error in a	Error in Damping	Error in Frequency	$\omega = a$	$\omega = \frac{\omega_0}{4}$	$\omega = \frac{\omega_0}{2}$
Euler	$-\frac{1}{2}aT$	T	$-\frac{1}{3}T^2$	$(-.25 - j.25)Ta$	$.11 - j1.11$	$-1 - j1.57$
Heun	$-\frac{1}{6}(aT)^2$	$\frac{1}{8}T^3$	$\frac{1}{6}T^2$	$(-.17 - j.58)(Ta)^2$	$-.21$	$-1.$
Runge-Kutta	$-\frac{1}{120}(aT)^4$	$-\frac{1}{144}T^5$	$-\frac{1}{120}T^4$		$-.37$	-2.05
Nystrom	$-\frac{1}{6}(aT)^2$	0	$\frac{1}{6}T^2$	$(.5 - j.5)(Ta)^2$	$-j\frac{3.14}{Ta}$	$.57$
Simpson-Milne	$-\frac{13}{72}(aT)^4$	0	$-\frac{13}{72}T^4$		$-j\frac{3.14}{Ta}$	$.05$

4 METHODS FOR THE ANALYSIS OF MIXED-DATA SYSTEMS

The dynamic analysis of hybrid systems is inherently more difficult than the analysis of digital systems, which have been the subject of the previous sections. This added difficulty occurs because of the joint presence of data sequences and time functions (continuous-data signals). In this section we extend the z-transform to the treatment of general systems with mixed data, and in the next section we apply the theory of this section to the dynamic analysis of several hybrid systems. As in Section 2, the treatment here is abbreviated, and for additional detail the reader should see reference 1, which takes the point of view developed below, or other texts on the theory of sampled-data systems.

First of all let us review the methods for expressing the response of continuous-data systems. If the system is linear, time-invariant, and initially at rest, it is possible to write

$$c(t) = \int_0^\infty h(t-\sigma)r(\sigma)d\sigma\,, \tag{4.1}$$

where $r(t)$ is the input, $c(t)$ is the output, and $h(t)$ is the impulse response. Alternatively, by introducing the Laplace transform of the three time functions, *e.g.*,

$$C(s) = \mathfrak{L}\,[c(t)] = \int_0^\infty e^{-st}c(t)\,dt, \tag{4.2}$$

the convolution integral (4.2) may be replaced by

$$C(s) = H(s)R(s) \tag{4.3}$$

Because of the inherent simplicity of the frequency domain characterization, we will favor it in our subsequent work.

Now consider the simple mixed-data system shown in Figure 4.1. The input is a sequence of the data points $\{r_n\}$ with the point r_n being generated at time nT. The digital-to analog converter "reconstructs" from these data points a time function $r_e(t)$. Figure 4.2b shows the zero-order hold or zero-order extrapolation process where $r_e(t)$ is held at the value r_n for $nT \leq t < nT + T$. Thus in terms of the function $h_e(t)$, the *extrapolation generating function,*

$$r_e(t) = \sum_{k=0}^{\infty} r_k h_e(t-kT) \tag{4.4}$$

See figure 4.2a. More elaborate data reconstruction processes involve more elaborate generating functions. By using $r_e(t)$ as an input to the dynamic system (impulse response $h_s(t)$), it is possible to determine the continuous-data response of the mixed-data system. Let us seek a more effective approach by working in the frequency domain with the z-transform of $\{r_n\}$.

As the first step in this direction, we express the response $c(t)$ in terms of the *data-point response* of the mixed-data system. *The data-point response $h(t)$ is the response $c(t)$ to the unit input: $r_0 = 1$, $r_n = 0$ for $n > 0$.* It is clear from (4.4) that $h(t)$ is the response of the dynamic system to $h_e(t)$, that is

$$h(t) = \int_0^\infty h_s(t-\sigma)h_e(\sigma)d\sigma\,. \tag{4.5}$$

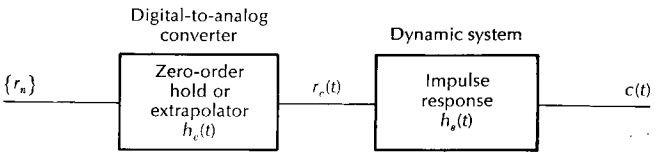
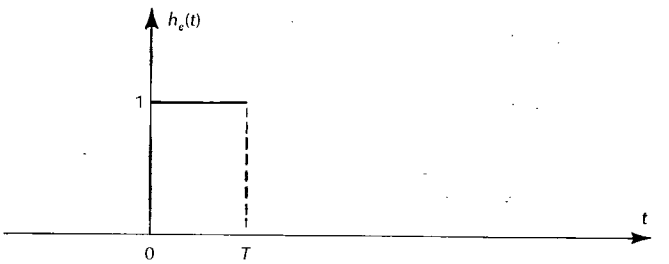
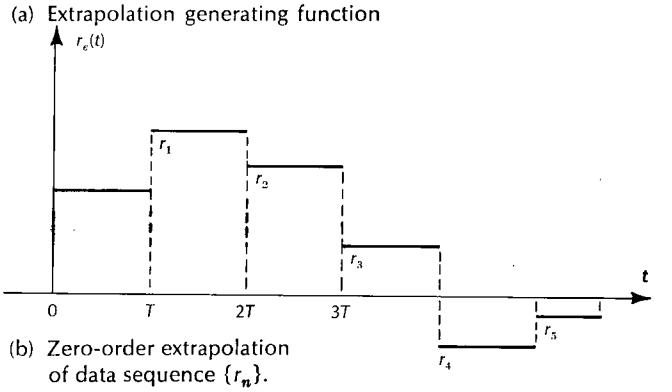


Figure 4.1 – Mixed-data system



(a) Extrapolation generating function



(b) Zero-order extrapolation of data sequence $\{r_n\}$.

Figure 4.2

Furthermore the linearity and time-invariance of the dynamic system imply that

$$c(t) = \sum_{k=0}^{\infty} h(t-kT)r_k. \tag{4.6}$$

This formula is the basis of all the results which follow.

The Laplace transform of equation (4.6) is

$$C(s) = \sum_{k=0}^{\infty} \mathfrak{L}\,[h(t-kT)]\,r_k \tag{4.7}$$

Making use of the translation theorem of the Laplace transform ($\mathfrak{L}\,[f(t-\tau)] = F(s)\,e^{-s\tau}$) and the definition of the z-transform (equation (2.1)) we obtain

$$C(s) = \sum_{k=0}^{\infty} H(s)\,e^{-skT}r_k = H(s)\sum_{k=0}^{\infty} (e^{sT})^{-k}r_k$$

$$C(s) = H(s)R^*(e^{sT}) \tag{4.8}$$

where

$$H(s) = \mathfrak{L}\,[h(t)]\,, \tag{4.9}$$

Note that $H(s) = H_e(s)H_s(s)$ follows from (4.5) in the same way that (4.3) follows from (4.1). Thus in the frequency domain a very simple expression for system response holds (compare with (4.6)). For apparent reasons $H(s)$ is called the *data-point transfer function* of the mixed-data system. Although we will find (4.8) most useful in setting up response equations for complex interconnections of systems, it is not very useful for actually evaluating $c(t)$. This is because $C(s)$ is generally a mixture of functions rational in s and rational in e^{sT}, and tables of inverse Laplace transforms for such functions are not available.

While it is difficult to obtain $c(t)$ for all $t \geq 0$, it is not difficult to obtain $c(t)$ for $t = nT$, $n = 0, 1, \ldots$. Employing equation (4.6),

$$c_n = c(nT) = \sum_{k=0}^{\infty} h(nT - kT) r_k \tag{4.10}$$

Defining $h_m = h(mT)$, (4.10) can be written as

$$c_n = \sum_{k=0}^{\infty} h_{n-k} r_k \tag{4.11}$$

Thus the results of Section 2 imply that

$$\mathcal{Z}\{c_n\} = C^*(z) = H^*(z)R^*(z)\,, \tag{4.12}$$

where

$$H^*(z) = \mathcal{Z}\{h(nT)\}\,. \tag{4.13}$$

From $C^*(z)$, c_n may be evaluated using the procedures described in Section 2, e.g., the inverse z-transform, the power series expansion in z^{-1}, and tables.

If the response $c(t)$ is sufficiently "smooth" so that it is "well represented" by $\{c_n\}$, (4.12) serves as an adequate description for the dynamic response of the mixed-data system. By comparing (4.12) and (2.23), we see that we may interpret the mixed-data system as a digital system. Thus, for example, if $r_n = e^{j\omega nT}$, $-\infty < n < \infty$,

$$c_n = H^*(e^{j\omega T})\, e^{j\omega nT} \tag{4.14}$$

and $H^*(e^{j\omega T})$ determines the "gain" and "phase shift" at frequency ω. Our remaining problem is to obtain $H^*(z)$ when the sampled-data system has a more complex configuration than that shown in figure 4.1.

The complex systems which we wish to consider consist of the interconnection of the basic elements shown in figure 4.3. In addition to the response expressions shown in figure 4.3, we need the following notation,

$$\mathcal{Z}[f(t)] = \mathcal{Z}[F(s)] = \mathcal{Z}\{f(nT)\} \tag{4.15}$$

Thus the z-transform of a time function is the z-transform of the data sequence generated by taking samples of the time function at $t = nT$. Correspondingly, the z-transform of a Laplace transform $F(s)$ is the z-transform of the time function $f(t)$ corresponding to the Laplace transform $F(s)$. If $f(t)$ is piecewise continuous with a jump at $t = nT$ for some integer n, particular care must be taken in the definition of $\mathcal{Z}[f(t)]$. See reference 1. Using this notation, we will now derive the following important result:

$$\mathcal{Z}[X(s)Y^*(e^{sT})] = \mathcal{Z}[X(s)]Y^*(z) = X^*(z)Y^*(z)\,. \tag{4.16}$$

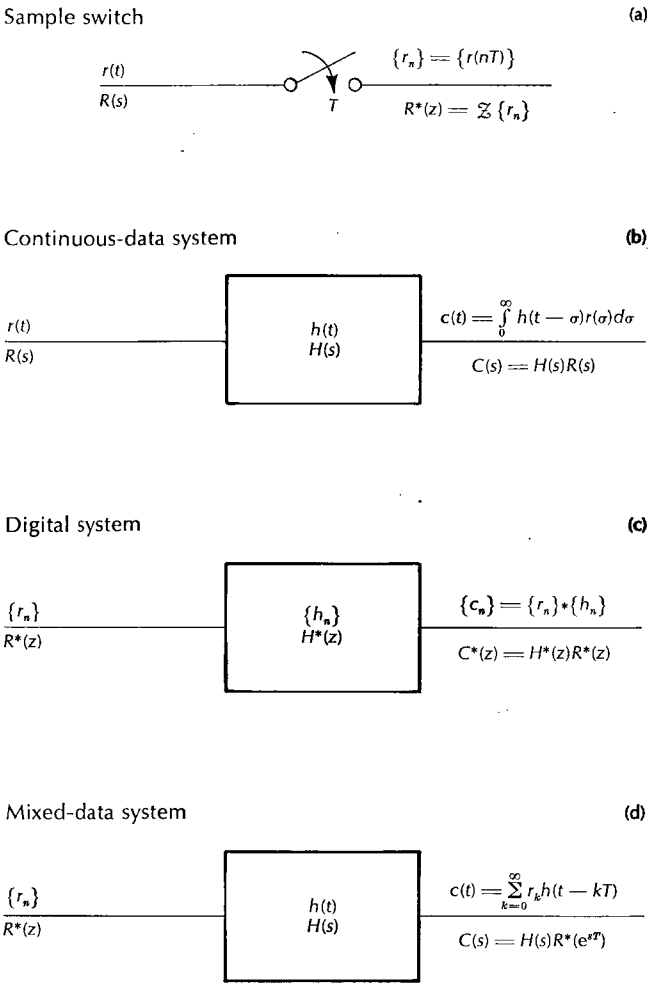


Figure 4.3 – Basic elements of sampled-data systems

First note that by the definition of $Y^*(z)$

$$X(s)Y^*(e^{sT}) = \sum_{k=0}^{\infty} X(s)(e^{sT})^{-k} y_k = \sum_{k=0}^{\infty} X(s)e^{-skT} y_k\,. \tag{4.17}$$

From the translation theorem of the Laplace transform, it follows that the time function

$$v(t) = \mathcal{L}^{-1}[X(s)Y^*(e^{sT})] = \sum_{k=0}^{\infty} x(t - kT) y_k\,. \tag{4.18}$$

Using this, $\mathcal{Z}[X(s)] = X^*(z) = \mathcal{Z}\{x(nT)\}$, and $v(nT) = \sum_{k=0}^{\infty} x(nT - kT)y_k$, we have by the convolution property of the z-transform

$$\mathcal{Z}[X(s)Y^*(e^{sT})] = \mathcal{Z}\{v(nT)\} = X^*(z)Y^*(z)\,, \tag{4.19}$$

Q.E.D.

In reference 1, Section 15, formula (4.16), and the results of figure 4.3 are used to derive transfer function expressions for a variety of complex systems. We limit our derivations to the hybrid systems of the next section.

5. ANALYSIS OF HYBRID COMPUTER SYSTEMS

In this section we will illustrate, by means of several examples the application of the methods of Section 4 to the analysis of hybrid computer systems. As in Section 3 we must assume that the systems treated can be represented satisfactorily by a linear model.

Let us consider a hybrid system where differential equations are integrated by means of analog elements, but where some of the required generation of nonlinear functions is implemented by table storage in a digital computer. To be more specific, let us direct our attention to a typical computer loop, which might occur as part of an elaborate computer simulation. The differential equation being solved by this loop has the form

$$\ddot{x} + F(x, y(t), z(t), \cdots) = f(t) \tag{5.1}$$

where $\ddot{x}$ is an acceleration; $F(x, y(t), z(t), \cdots)$ is a nonlinear force term, dependent on x and other computer variables such as $y(t)$ and $z(t)$; and $f(t)$ is an external forcing function. Figure 5.1 shows the computer block diagram necessary for implementing (5.1) when F is generated by a digital computer.

The required computer operations proceed as follows. At time $t = nT$ the variables $x(t)$, $y(t)$, $z(t)$ are sampled. Then they are given a digital representation by the conversion system and are stored in the computer memory. After a table look-up (which may involve interpolation), the digital computer places the value $F(x_n, y_n, z_n, \ldots)$ in the output register. Let us suppose that all these steps are carried out before $t = nT + T$. Then at $t = nT + T$ the value $F(x_n, y_n, z_n, \ldots)$ is transferred to the register of the digital-to-analog converter, where it produces a constant analog value $F(x_n, y_n, z_n, \ldots)$ (within the quantizing error) during the period $nT + T \leq t \leq nT + 2T$. Different computer arrangements are possible (e.g., longer delay, more elaborate data hold), but this is what will be used for the subsequent analysis.

In order to simplify things still further, let us make the following assumptions: a) $y(t), z(t), \ldots$ are varying very slowly so that they may be assumed constant with values $\hat{y}$ and $\hat{z}$; b) for the range of $x(t)$ considered $F(x, \hat{y}, \hat{z}, \ldots) \cong a^2x$. While these assumptions may not always hold, they will at least allow us to get some feeling about the dynamics of the process. Figure 5.2 shows the simplified system using the notation of figure 4.3.

Let us now apply the theory of Section 4 to the determination of a response expression. Working in the frequency domain, we see from figure 4.3b

$$X(s) = G(s)E(s) = G(s)[F(s) - C_e(s)] \tag{5.2}$$

The zero-order hold (D-to-A converter), having a data sequence for an input and function of time for an output, is a mixed-data system. Since a unit data point ($c_0 = 1$; $c_n = 0$, $n > 0$) at its input produces a response

$$\begin{aligned} h(t) &= 1 \ , \quad 0 \leq t < T \\ &= 0 \ , \quad t \geq T \end{aligned} \tag{5.3}$$

(see figure 4.2), the data point transfer function is $\mathfrak{L}[h(t)] = H(s) = (1 - e^{-sT})s^{-1}$. In any case figure 4.3d shows that

$$C_e(s) = H(s)C^*(e^{sT}). \tag{5.4}$$

Because $D^*(z)$ represents the function generation which has a computer lag of one sample period, the delay property of the z-transform gives $D^*(z) = a^2z^{-1}$, *i.e.*,

$$C^*(z) = a^2z^{-1}X^*(z) = D^*(z)X^*(z). \tag{5.5}$$

Substituting (5.4) and (5.5) into (5.2)

$$X(s) = G(s)[F(s) - H(s)D^*(e^{sT})X^*(e^{sT})]. \tag{5.6}$$

We would like to solve (5.6) for either X or X^*, but cannot because X and X^* are both contained in the equation. To get around this difficulty, we take the z-transform of (5.6), using (4.16) on the last term $[X(s) \sim G(s)H(s), Y^*(e^{sT}) \sim D^*(e^{sT})X^*(e^{sT})]$, and obtain

$$X^*(z) = \overline{GF}^*(z) - \overline{GH}^*(z)D^*(z)X^*(z), \tag{5.7}$$

where $\overline{GF}^*(z) = \mathfrak{Z}[G(s)F(s)]$ and $\overline{GH}^*(z) = \mathfrak{Z}[G(s)H(s)]$. Finally,

$$X^*(z) = \frac{1}{1 + D^*(z)\overline{GH}^*(z)}\,\overline{GF}^*(z) = Y^*(z)\overline{GF}^*(z). \tag{5.8}$$

Once we have determined $[1 + D^*(z)\overline{GH}^*(z)]^{-1} = Y^*(z)$ we can use (5.8) to determine $X^*(z)$ and hence $\{x(nT)\}$ for a variety of inputs. For instance, if $f(t)$ is the unit step at time $t = 0$, $F(s) = 1/s$, and $G(s)F(s) = 1/s^3$ corresponds to $t^2/2$, which is the double integral of the unit step. Thus

$$\overline{GF}^*(z) = \mathfrak{Z}\left[\frac{1}{s^3}\right] = \mathfrak{Z}\left[\frac{t^2}{2}\right] = \mathfrak{Z}\left\{\frac{n^2T^2}{2}\right\}$$
$$= \frac{T^2}{2}(z-1)^{-3}z(z+1),$$

a result which follows from Table 2.1 or still more simply from a table of z-transforms of Laplace transforms (reference 1, table 13.1). Alternatively, if $f(t) = e^{j\omega t}$, we find

$$x_n = [Y(e^{j\omega T})(-\omega^{-2})]e^{j\omega Tn} \tag{5.9}$$

Thus the gain and phase shift of the hybrid system may be evaluated. Notice that in our development above we have omitted all initial-condition terms, that is, we have assumed the system to be initially at rest. To do otherwise would introduce additional complexity. It is also possible to obtain an expression for $C(s)$, but it is too complex to work with conveniently.

Let us now determine $\overline{GH}^*(z)$ and $Y^*(z)$. From the above

$$\overline{GH}^*(z) = \mathfrak{Z}[H(s)G(s)] = \mathfrak{Z}[s^{-3}(1 - e^{-sT})]$$
$$= \mathfrak{Z}[s^{-3}] - \mathfrak{Z}[s^{-3}e^{-sT}] = (1 - z^{-1})\,\mathfrak{Z}[s^{-3}], \tag{5.10}$$

where we have observed from the translation property of the Laplace transform that the data sequence corresponding to $s^{-3}e^{-sT}$ can be obtained by delaying the data sequence corresponding to s^{-3} by one sample period. We have seen in the preceding paragraph that $\mathfrak{Z}[s^{-3}] = T^2/2 \times (z-1)^{-3}z(z+1)$. Thus

$$\overline{GH}^*(z) = (T^2/2)(z+1)(z-1)^{-2} \tag{5.11}$$

Finally,

$$Y^*(z) = \frac{1}{1 + a^2z^{-1}\overline{GH}^*(z)}$$
$$= \frac{z(z-1)^2}{z^3 - 2z^2 + \left(1 + \frac{1}{2}a^2T^2\right)z + \frac{1}{2}a^2T^2}. \tag{5.12}$$

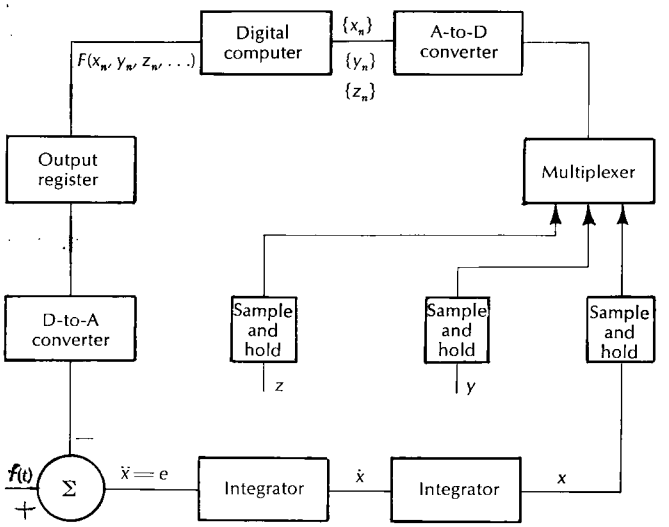


Figure 5.1 – Block diagram of hybrid system

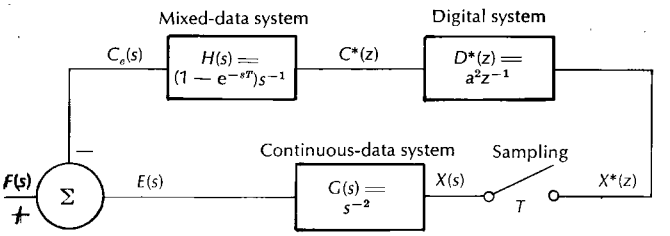


Figure 5.2 – Simplified representation of figure 5.1 showing mathematical operations and notation

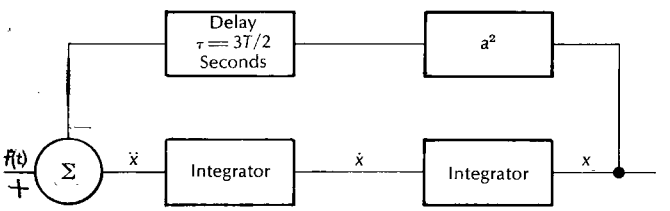


Figure 5.3 – Approximate representation of system in figure 5.2

It has been demonstrated in our analysis of integration methods that the poles of the system transfer function are a good indication of system response because they determine the form of the response terms originating from the system properties. There are three poles of $Y^*(z)$ at z_1, z_2, and z_3. Let us determine z_1, z_2, and z_3. For $aT = 0$ it is clear that the characteristic equation $z^3 - 2z^2 + [1 + (1/2)\,a^2T^2]z + (1/2)a^2T^2 = 0$ has a root $z_1 = 0$. Thus for $aT \neq 0$ but $0 < aT << 1$, $z_1 \cong 0$. By letting $z_1 = c_1(aT) + c_2(aT)^2 + \cdots$ and choosing $c_1, c_2, \cdots$ to make z_1 be a solution of the characteristic equation, it can be shown that

$$z_1 = -\frac{1}{2}(aT)^2 + \frac{3}{4}(aT)^4 + \cdots \qquad (5.13)$$

By factoring the characteristic equation, we have

$$(z - z_1)\left[z^2 + (z_1 - 2)\,z + 1 + \frac{1}{2}(aT)^2 - 2z_1 + z_1^2\right] = 0. \qquad (5.14)$$

By using the quadratic formula on the quadratic in the brackets and substituting for z_1 from (5.13), it follows that

$$z_2, z_3 = 1 + \frac{1}{4}(aT)^2 - \frac{3}{8}(aT)^4 + \cdots \pm j\,aT\left(1 + \frac{9}{32}(aT)^2 + \cdots\right). \qquad (5.15)$$

Since $|z_1| << 1$, the response term corresponding to z_1 decays very quickly and does not play a significant role in system response. If the hybrid system is to give high solution accuracy, the remaining roots, z_2 and z_3, should be properly related to the solution properties of (5.1). In particular, with $F = a^2x$ the system (5.1) is linear and has transient response terms of the form $e^{\pm jat}$, *i.e.*, the response is purely oscillatory with frequency a. This means that if we write $z_2, z_3 = e^{\sigma T}e^{\pm j\omega T}$, we should have $\sigma = 0$ and $\omega = a$. Actually, from

$$|z_2|^2 = |z_3|^2 = \left[1 + \frac{1}{4}(aT)^2 - \frac{3}{8}(aT)^4 + \cdots\right]^2 + (aT)^2\left[1 + \frac{9}{32}(aT)^2 + \cdots\right]^2 = 1 + \frac{3}{2}(aT)^2 - \frac{5}{4}(aT)^4 + \cdots \qquad (5.16)$$

and

$$\angle z_2 = -\angle z_3 = \tan^{-1}\frac{aT - \frac{9}{32}(aT)^3 + \cdots}{1 + \frac{1}{4}(aT)^2 + \cdots} = aT + \left(-\frac{17}{32} - \frac{1}{3}\right)(aT)^3 + \cdots, \qquad (5.17)$$

we see that

$$\sigma = \frac{1}{2T}\ln|z_2|^2 = \frac{3}{4}a^2T + \cdots \qquad (5.18)$$

and

$$\omega = \frac{1}{T}\angle z_1 = a[1 - .864(aT)^2 + \cdots]. \qquad (5.19)$$

Thus we have errors in both the frequency and damping. The damping error corresponds to an attained damping ratio $\zeta \cong -(3/4)\,aT$, and, being to the first power in aT, is the worse of the two. The values of σ and ω are perhaps the best single indicators of overall computational accuracy for the system in figure 5.1. A complete analysis of the type undertaken in section 3 would include determination of initial condition response and sinusoidal response.

Let us see if we can obtain a further understanding of the source of the above errors. Consider the approximate representation of the system shown in figure 5.3. Here we have replaced the sampling, function-generation, and data-hold operations with a transport lag (transfer function $e^{-s\tau}$) and gain a^2. The delay time τ should include the delay T in the digital computer, and some measure of

the delay in the zero-order hold. From figure 4.2 it would seem that $T/2$ is the most reasonable choice for this delay. Thus we take $\tau = 3T/2$.

From figure 5.3, we see that the system equation,

$$\ddot{x}(t) + a^2 x[t - (3T/2)] = f(t),$$

is a *difference differential* equation. Since there is some difficulty in solving (5.4), let us obtain an ordinary differential equation which approximates it. If T is not too large and $x(t)$ is sufficiently smooth

$$x\left(t - \frac{3}{2}T\right) \cong x(t) - \frac{3}{2}T\dot{x}(t) + \frac{9}{8}T^2\ddot{x}(t)\,. \quad (5.21)$$

Making this approximation in (5.20), we obtain

$$\left(1 + \frac{9}{8}a^2T^2\right)\ddot{x} - \frac{3}{2}a^2T\dot{x} + a^2x = f(t) \quad (5.22)$$

The characteristic equation,

$$\left(1 + \frac{9}{8}a^2T^2\right)\lambda^2 - \frac{3}{2}a^2T\lambda + a^2 = 0\,,$$

for this system has roots, λ_1, $\lambda_2 = \sigma \pm j\omega$, where (by the quadratic formula)

$$\sigma = \frac{3}{4}a^2T + \cdots \quad (5.23)$$

$$\omega = a\left[1 - \frac{27}{32}(aT)^2 + \cdots\right] = a[1 - .844\,(aT)^2 + \cdots]. \quad (5.24)$$

Equations (5.18) and (5.23) agree exactly (except for terms of order $(aT)^3$ which have been neglected), while (5.19) and (5.24) agree very closely (the fractional error in ω is $-.844(aT)^2$ instead of $-.864(aT)^2$).

Since the above analysis seems to indicate that the principal error source is a time delay of $3T/2$, the system in figure 5.4 suggests itself. See references 5 and 6. Here

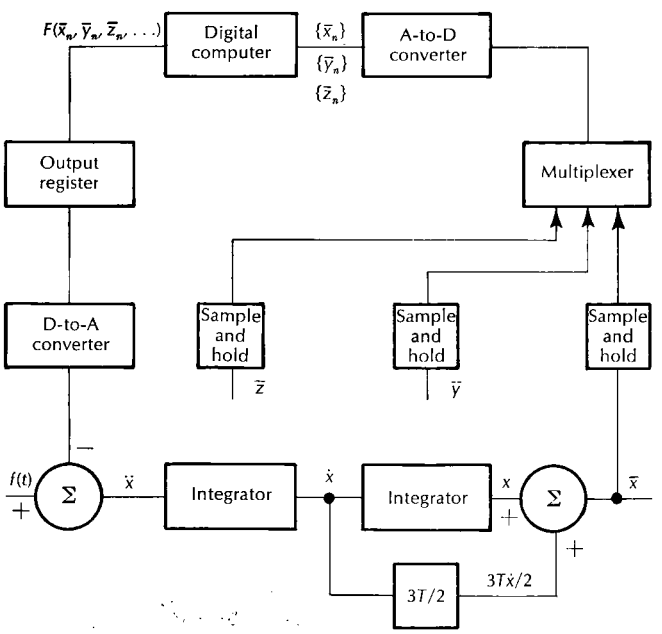


Figure 5.4 – Block diagram of hybrid system with delay compensation

we have the same system as shown in figure 5.1, but now the signal $x(t) + 3T\dot{x}(t)/2 = \bar{x}(t)$ is passed on to the conversion system. Since $\dot{x}(t)$ is continuous (it is an integrator output), $\bar{x}(t)$ is a reasonable approximation to $x(t + 3T/2)$ and hence the scheme should tend to cancel the delay in the function generation system. If $y(t)$ and $z(t)$ were rapidly varying, it would make sense to replace them by $\bar{y}(t) = y(t) + 3T\dot{y}(t)/2$ and $\bar{z}(t) = z(t) + 3Tz(t)/2$ as shown.

Let us now analyze this new system. The approximation technique used in equations (5.21) through (5.24) is now not productive and the sampled-data system model must be used. If we replace X by $\bar{X}$ and G by

$$G = s^{-2}\left(\frac{3}{2}Ts + 1\right) = s^{-2} + \frac{3}{2}Ts^{-1} \quad (5.24)$$

in figure 5.2, it will represent the system in figure 5.4. Thus we may write

$$\bar{X}^*(z) = Y^*(z)\overline{GF}^*(z) \quad (5.25)$$

where $Y^*(z)$ has the same form as before. Let us determine where the poles of $Y^*(z)$ are now located.

First we note that

$$\begin{aligned}\overline{GH}^*(z) &= \mathcal{Z}\,[H(s)G(s)] = \mathcal{Z}\,(1 - e^{sT})\left(s^{-3} + \frac{3}{2}Ts^{-2}\right)\\ &= (1 - z^{-1})\left(\mathcal{Z}\,[s^{-3}] + \frac{3}{2}T\,\mathcal{Z}\,[s^{-2}]\right)\\ &= (1 - z^{-1})\left[\frac{T^2}{2}\frac{(z+1)z}{(z-1)^3} + \frac{3T^2}{2}\frac{z}{(z-1)^2}\right]\\ &= T^2\frac{2z-1}{(z-1)^2}\end{aligned} \quad (5.26)$$

where we have used $\mathcal{Z}\,[s^{-3}]$ as before and $\mathcal{Z}\,[s^{-2}] = \mathcal{Z}\,[t] = \mathcal{Z}\,\{nT\} = Tz(z-1)^2$ from Table 2.1. Thus

$$Y^*(z) = \frac{1}{1 + a^2z^{-1}\overline{GH}^*(z)} = \frac{z(z-1)^2}{z^3 - 2z^2 + (1 + 2a^2T^2)z - a^2T^2}, \quad (5.27)$$

and the characteristic equation is $z^3 - 2z^2 + (1 + 2a^2T^2) \times z - a^2T^2 = 0$. As before, the root $z_1 \cong 0$, and we can obtain a power series for z_1, which in this case has the form

$$z_1 = (aT)^2 + \cdots, \quad (5.28)$$

where the omitted terms are of order $(aT)^6$ and higher. Factoring the characteristic equation gives

$$(z - z_1)[z^2 + (z_1 - 2)z + 1 + 2a^2T^2 - z_1 + z_1^2] = 0\,, \quad (5.29)$$

from which the quadratic formula and (5.28) imply

$$z_2, z_3 = 1 - \frac{1}{2}(aT)^2 + \cdots \pm jaT\left(1 + \frac{3}{8}(aT)^2 + \cdots\right). \quad (5.30)$$

By repeating the steps which lead to (5.18) and (5.19), it follows that in this case

$$\sigma = \frac{1}{2} a^4 T^3 + \cdots, \tag{5.31}$$

$$\omega = a\left(1 + \frac{13}{24}(aT)^2 + \cdots\right). \tag{5.32}$$

Thus the fractional error in frequency is $(13/24)(aT)^2 = .542(aT)^2$ rather than the $-.864(aT)^2$ obtained with the original system. Although the sign of error is different, its magnitude has not been significantly affected. However, (5.31) shows that the attained damping ratio $\zeta \cong -(aT)^3/2$, which is a very significant improvement over the earlier $\zeta \cong -3aT/4$. For example, a damping-ratio accuracy of .001 requires a computer period $T = .001\,(4/3)\,a^{-1}$ (*i.e.*, $(3/4) \cdot 1000 \cdot 2\pi \cong 4710$ calculations/cycle) if delay compensation is not used, whereas with compensation $T = (.002)^{1/3} a^{-1}$ (*i.e.*, $(500)^{1/3}\, 2\pi \cong 50$ calculations/cycle) will suffice.

The above analyses do not disagree with most of the experimental work in references 5 and 6. However, figure 16 in reference 6, which shows the experimental error in natural frequency for the compensated system, is at variance with equation (5.32). Of the basic results (5.18), (5.19), (5.31), and (5.32) only (5.18) is given (without full derivation) in references 5 and 6.

REFERENCES

1 E G GILBERT
Notes on sampled-data systems

2 University of Michigan 1965
Summer intensive course on hybrid computation

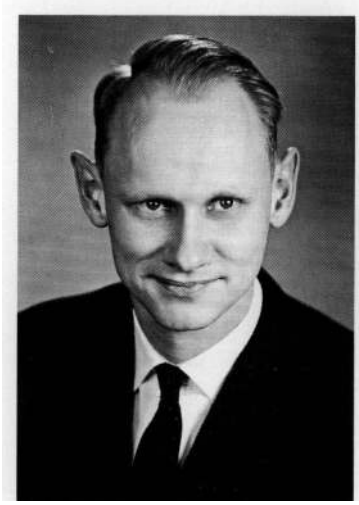
3 P HENRICI
Discrete variable methods in ordinary differential equations
Wiley 1962

4 R M HOWE
Notes on error analysis of combined analog-digital computer systems
Unpublished

5 T MIURA J IWATA
Effects of digital execution time in a hybrid computer
AFIPS Conference Proceedings vol 24 1963 pp 251-266

6 T MIURA J IWATA
Study on time-shared analog computation techniques
SIMULATION vol 5 no 5 1965 pp 318-327

7 J TODD (Editor)
Survey of numerical analysis
McGraw Hill 1962 (see chapters 2 and 9)



ELMER G. GILBERT received the BS and MS degrees in electrical engineering and the PhD degree in instrumentation engineering from the University of Michigan, where he is currently a professor in the Department of Aerospace Engineering.

His interests in computation and simulation go back to 1953 when he worked with his brother Edward O. Gilbert and Robert M. Howe on analog computer systems and applications. Since 1955 he has taught in the information and control engineering (formerly instrumentation engineering) program at the University of Michigan. His current interests and past work include the mathematical theory of control, mathematical programming as applied to optimal control, computer devices, and hybrid computation with emphasis on optimization techniques, processing of random signals, and error analysis.

He is one of the founders of Applied Dynamics, Inc. and maintains a regular consulting relation with this firm.

Professor Gilbert is a member of IEEE and serves on several committees of the Automatic Control Group. He is also a member of AAAS, SIAM, and several honorary societies.