The Analysis of Sampled-Data Systems

J. R. RAGAZZINI
MEMBER AIEE

L. A. ZADEH
ASSOCIATE MEMBER AIEE

THERE is an important class of feedback control systems known as sampled-data systems or sampling servomechanisms in which the data at one or more points consist of trains of pulses or sequences of numbers. Such systems may have a variety of forms, a common example of which is shown in Figure 1. In the case illustrated, the sampling is performed on the control error by a so-called sampler which is indicated as a mechanical switch which closes momentarily every T seconds. The data at the output of such a switch consist of a train of equally spaced pulses of short duration whose envelope is the control error function. In some practical systems, the separation between successive pulses is controlled by some characteristic of the input signal and consequently is not constant. Such systems will not be considered in this paper.

In a typical sampled-data system such as that shown in Figure 1, the sampler is followed by a smoothing circuit, commonly referred to as hold or clamp circuit, whose function is to reproduce approximately the form of the original error function by an interpolation or extrapolation of the pulse train. Following the hold circuit, there are the usual components of the feedback loop, shown in Figure 1 as H and G, comprising amplifiers, shaping networks, and the controlled member.

It is apparent that the insertion of a sampler into an otherwise continuous control system in general should result in an inferior performance due to a loss of information in the control data. Yet, sampled-data systems have certain engineering advantages which make them preferable in some applications to continuous-data systems. The most important of these advantages is the fact that error sampling devices can be made extremely sensitive at the expense of bandwidth. An example of such a device is the electromechanical galvanometer and chopper bar. In this device a very sensitive though sluggish galvanometer is used to detect error and its position is sampled periodically by means of a chopper bar. The latter permits an auxiliary source of power to rotate a sizable potentiometer to a position determined by the clamped galvanometer needle. The process is carried out at uniform intervals and a sampled and clamped output is obtained for use in the continuous part of the control system. Bandwidth is lost through the sluggishness of the unloaded galvanometer, but the power gain is enormous. Similar devices for measurement of pressure errors, flow, or other phenomena can be devised along the same general lines.

In addition, there are some systems in which the data-collecting or transmission means are intermittent. Radars and multichannel time-division communication links are examples of this type of device. Such devices may be treated, in general, as sampled-data systems provided the duration of sampling is small by comparison with the settling time of the system.

Despite the increasing use of sampled-data devices in the fields of communication and control, the volume of published material on such devices is still rather limited.[1-10] The several different methods which have been developed for the analysis of sampled-data systems are closely related to the well-known mathematical techniques of solution of difference equations. It is the purpose of this paper to unify and extend the methods described in the literature and to investigate certain basic aspects of sampled-data systems.

Input-Output Relations

A central problem in the analysis of sampled-data systems is that of establishing a mathematical relation between the input and output of a specified system. This problem has received considerable attention in the literature of sampled-data systems, with the result that several different types of input-output relations have been developed, notably by Shannon,[1] Hurewicz,[2] and Linvill.[3] Shannon's relation involves the Fourier transforms of the sampled input and output; Hurewicz's relation is based on the use of so-called generating functions, which in this paper are referred to as z-transforms and which are, in fact, a disguised form of the Laplace transforms; while Linvill's relation involves directly the Laplace transforms of the input and output. The principal difference between Shannon's and Hurewicz's relations on the one hand, and that of Linvill on the other, is the fact that the former yield only the values of the output at the sampling instants, while the latter provides the expression for the output at all times, though at the cost of greater labor.

The analysis presented in this section has a dual objective: to achieve a unification of the approaches used by Shannon, Hurewicz, and Linvill; and to formulate the input-output relations for the basic types of sampled-data systems. In the next section, the problem of establishing a relation between the input and output will be approached from a significantly different point of view. Specifically, a sampled-data system will be treated as a time-variant system and its behavior will be characterized by a system function which involves both frequency and time.

The basic component of sampled-data systems is the sampler, whose output has the form of a train of narrow pulses occurring at the sampling instants 0, $\pm T$, $\pm 2T$, . . . , where T is the sampling interval; see Figure 2. The frequency $\omega_0 = 2\pi/T$ is called the sampling frequency.

For purposes of mathematical convenience, it is expedient to treat the output pulses as impulses whose areas are equal to the values of the sampled time function at the respective sampling instants. (This is permissible provided the pulse duration is small compared with the settling time of the system and the gain of the amplifier following the sampler is multiplied by a factor equal to the time duration of the sampling pulse.) Thus, if the input and output of the sampler are denoted by $r(t)$ and $r^*(t)$ respectively, the

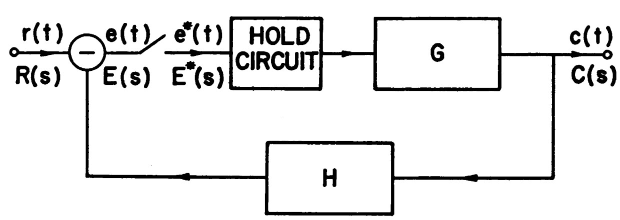


Figure 1. Typical sampled-data control system

Paper 52-161, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Technical Program Committee for presentation at the AIEE Summer General Meeting, Minneapolis, Minn., June 23–27, 1952. Manuscript submitted March 21, 1952; made available for printing April 16, 1952.

J. R. RAGAZZINI and L. A. ZADEH are both with the Department of Electrical Engineering, Columbia University, New York, N. Y.

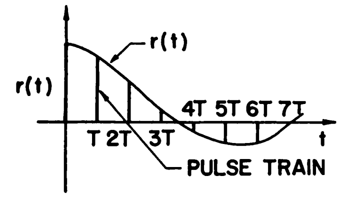


Figure 2. Pulse train at output of sampler

relation between them reads

$$r^*(t)=r(t)\delta_T(t) \tag{1}$$

where $\delta_T(t)$ represents a train of unit impulses (delta functions)

$$\delta_T(t)=\sum_{n=-\infty}^{\infty}\delta(t-nT) \tag{2}$$

Equation 1 may be written equivalently as

$$r^*(t)=\sum_{n=0}^{\infty}r(nT)\delta(t-nT) \tag{3}$$

where the negative values of n are absent by virtue of the assumption that $r(t)$ vanishes for negative values of t.

Equation 3 furnishes an explicit expression for the output of the sampler. It is more convenient, however, to deal with the Laplace transform of $r^*(t)$, which is denoted by $R^*(s)$

$$R^*(s)=\mathcal{L}\{r(t)\delta_T(t)\} \tag{4}$$

One expression for $R^*(s)$ can be obtained at once by transforming both sides of equation 3; this yields

$$R^*(s)=\sum_{n=0}^{\infty}r(nT)\epsilon^{-nTs} \tag{5}$$

An alternative expression for $R^*(s)$ can be obtained by expressing $\delta_T(t)$ in the form of a complex Fourier series

$$\delta_T(t)=\frac{1}{T}\sum_{n=-\infty}^{\infty}\epsilon^{jn\omega_0 t} \tag{6}$$

and substituting this expression in equation 1. Then, transforming the resulting series term by term there results

$$R^*(s)=\frac{1}{T}\sum_{n=-\infty}^{\infty}R(s+jn\omega_0) \tag{7}$$

where $R(s)$ is the Laplace transform of $r(t)$. It is of interest to note that the equivalence between the two expressions for $R^*(s)$

$$\frac{1}{T}\sum_{n=-\infty}^{\infty}R(s+jn\omega_0)=\sum_{n=0}^{\infty}r(nT)\epsilon^{-nTs} \tag{8}$$

was discovered more than a century ago by Poisson, and that equation 8 is essentially equivalent to the Poisson summation rule.

An inspection of either of the alternative expressions for $R^*(s)$ indicates that when s in $R^*(s)$ is replaced by $s+jm\omega_0$, where m is any integer, the resulting expression is identical with $R^*(s)$. This implies that $R^*(s)$ is a periodic function of s with period $j\omega_0$; thus

$$R^*(s+jm\omega_0)=R^*(s) \tag{9}$$

where m = any integer.

The infinite series expression for $R^*(s)$ given by equation 5 readily can be put into closed form whenever $r(t)$ is a linear combination of products of polynomials and exponential functions. For instance, when $r(t)=\epsilon^{-at}$, the right-hand member of equation 5 is a geometric series which upon summation yields

$$R^*(s)=\frac{1}{1-\epsilon^{-at}\epsilon^{-Ts}} \tag{10}$$

When expressed in the form given by equation 5, the transform of the pulsed output, $R^*(s)$, is a function of ϵ^{sT}. This suggests that an auxiliary variable $z=\epsilon^{sT}$ be introduced and that $R^*(s)$ be written in terms of this variable. When this is done, the function $R^*(s)$, expressed as a function of z, is called the z-transform of $r(t)$. For notational convenience it is denoted by $R^*(z)$ although strictly speaking it should be written as $R^*(1/T \log z)$. With this convention, the transform given in equation 10, for example, reads

$$R^*(z)=\frac{1}{1-\epsilon^{-aT}z^{-1}} \tag{11}$$

In what follows, the symbols $R^*(s)$ and $R^*(z)$ will be used interchangeably since they represent the same quantity, namely the Laplace transform of $r(t)\delta_T(t)$. Needless to say, in cases where $r(t)$ has the form of a sequence of impulses (or numbers equal to the areas of respective impulses), the z-transform $R^*(s)$ is simply the Laplace transform of $r(t)$, and not of $r(t)\delta_T(t)$.

It will be noted that the z-transform as defined is closely related to the generating function used by Hurewicz. However, the z-transform is a more natural concept since it stems directly from the Laplace transform of the sampled time function. It is of historical interest to note that generating functions, which, as pointed out, are essentially equivalent to z-transforms, were introduced by Laplace[11] and were extensively used by him in connection with the solution of difference equations.

In practice, $R^*(z)$ is generally a rational function of z, and its inversion, that is the determination of a function $r(t)$ of which $R^*(z)$ is the z-transform, is most rapidly carried out by using a table of z-transforms such as the one compiled in Table I. (More extensive tables of closely related types of transforms may be found in references 12 and 13.) One use of this table is in finding the z-transform corresponding to the Laplace transform of a given function. Despite its brevity, Table I is adequate for most practical purposes in view of the fact that both the Laplace transforms and the z-transforms can be expanded into partial fractions each term of which can be inverted individually.

It is important to note that the inverse of a z-transform is not unique. Thus, if $F^*(z)$ is an entry in the table and $f(t)$ is its correspondent, then any function of time which coincides with $f(t)$ at the sampling instants 0, T, $2T$, $3T$, . . . , has the same z-transform as $f(t)$. To put it another way, if $G^*(z)$ is the z-transform of some function $g(t)$, then the inverse of $G^*(z)$, as found from the table, is not, in general, identical with $g(t)$, although it coincides with $g(t)$ at the sampling instants. Thus, from the z-transform of a function one can find only the values of the function at the sampling instants. In this connection, it should be noted that the value of a time function at the nth sampling instant is equal to the coefficient of z^{-n} in the power series expansion of its z-transform (regarded as a function of z^{-1}). In cases requiring numerical computations, this property of z-transforms affords an alternative, and frequently convenient, means of calculating the values of corresponding time functions at the sampling instants.

It will be helpful to summarize at this

Table I. Abbreviated Table of Laplace and z-Transforms

	Laplace Transform $F(s)$	Time Function $f(t)$	z-Transform $F^*(z)$
(1)..	1	$\delta(t)$	z^{-0}
(2)..	ϵ^{-nTs}	$\delta(t-nT)$	z^{-n}
(3)..	$\frac{1}{s}$	1	$\frac{1}{1-z^{-1}}$
(4)..	$\frac{1}{s^2}$	t	$\frac{Tz^{-1}}{(1-z^{-1})^2}$
(5)..	$\frac{1}{s+a}$	ϵ^{-at}	$\frac{1}{(1-\epsilon^{-aT}z^{-1})}$
(6)..	$\frac{a}{s(s+a)}$	$(1-\epsilon^{-at})$	$\frac{z^{-1}(1-\epsilon^{-aT})}{(1-z^{-1})(1-\epsilon^{-aT}z^{-1})}$
(7)..	$\frac{a}{s^2+a^2}$	$\sin at$	$\frac{\sin aTz^{-1}}{1-(2\cos aT)z^{-1}+z^{-2}}$
(8)..	$F(s+a)$	$\epsilon^{-at}f(t)$	$F^*(\epsilon^{-aT}z)$
(9)..	$\epsilon^{-sT}F(s)$	$f(t-T)$	$z^{-1}F^*(z)$
(10)..	$\epsilon^{as}F(s)$	$f(t+a)$	$z^{a/T}F^*(z)$
(11)..	$\frac{1}{s-\frac{1}{T}\ln a}$	$a^{t/T}$	$\frac{z}{z-a}$
(12)..	$\frac{s}{s^2+a^2}$	$\cos at$	$\frac{1-\cos aTz^{-1}}{1-(2\cos aT)z^{-1}+z^{-2}}$

stage the basic points of the foregoing discussion:

1. A sampler transforms a function $r(t)$ into a train of impulses, $r^*(t)=r(t)\delta_T(t)$, where $\delta_T(t)$ represents a train of unit impulses with period T.

2. The Laplace transform of $r^*(t)$, $R^*(s)$, is expressible in two different but equivalent forms given by equations 5 and 7.

3. $R^*(s)$ is a periodic function of s with period $j\omega_0$, where ω_0 is the sampling frequency.

4. The z-transform of $r(t)$, $R^*(z)$, is equal to $R^*(s)$ with ϵ^{sT} in $R^*(s)$ replaced by z; that is

$$R^*(z)=\mathcal{L}\{r(t)\delta_T(t)\} \tag{12}$$

with ϵ^{sT} replaced by z.

As a preliminary to the consideration of sampled-data feedback systems, it will be helpful to establish one basic property of z-transforms. The property in question concerns the relation between the z-transforms of the output and input of the system illustrated in Figure 3. Denoting the input by $r(t)$, the output by $c(t)$, and the transfer function by $G(s)$, this relation reads

$$C^*(z)=G^*(z)R^*(z) \tag{13}$$

where $C^*(z)$ and $R^*(z)$ are the z-transforms of $c(t)$ and $r(t)$ respectively, and $G^*(z)$ is referred to as the starred transfer function. The importance of this relation derives from its similarity to the familiar relation $C(s)=G(s)R(s)$, which would obtain in the absence of samplers. This similarity makes it possible to treat z-transforms and starred transfer functions in much the same manner as the conventional Laplace transforms and transfer functions.

The proof of equation 13 is quite simple. From inspection of Figure 3, it is evident that the Laplace transform of $c(t)$ is given by

$$C(s)=G(s)R^*(s) \tag{14}$$

and correspondingly, by applying equation 7

$$C^*(s)=\frac{1}{T}\sum_{n=-\infty}^{\infty} G(s+jn\omega_0)R^*(s+jn\omega_0) \tag{15}$$

Because of the periodicity of $R^*(s)$, the following identity is noted

$$R^*(s+jn\omega_0)\equiv R^*(s) \tag{16}$$

Consequently equation 15 reduces to

$$C^*(s)=R^*(s)\left[\frac{1}{T}\sum_{n=-\infty}^{\infty} G(s+jn\omega_0)\right] \tag{17}$$

Denoting the bracketed term by $G^*(s)$

$$G^*(s)=\frac{1}{T}\sum_{n=-\infty}^{\infty} G(s+jn\omega_0) \tag{18}$$

equation 17 assumes the following form

$$C^*(s)=G^*(s)R^*(s) \tag{19}$$

which is equivalent to equation 13. It is seen that the starred transfer function $G^*(s)$, which is related to the transfer function $G(s)$ by equation 18, may be regarded as the ratio of the z-transforms of the output and input of the sampled-data system under consideration.

The mathematical essence of the foregoing discussion is the fact that a relation of the form $C(s)=G(s)R^*(s)$ implies $C^*(s)=G^*(s)R^*(s)$. This fact *per se* is very useful in the analysis of sampled-data systems involving one or more feedback loops. In the sequel, the process of passing from equation 14 to equation 19 will be referred to as the z-transformation of both sides of equation 14. The tacit understanding exists, of course, that the quantities actually subjected to the z-transformation are the time functions corresponding to the two members of equation 14.

Among the properties of the starred transfer function $G^*(s)$ there are two that are of particular importance. First, suppose that an input of the form $r(t)=\epsilon^{st}$ is applied to the system shown in Figure 3. This input is transformed by the sampler into an impulse train $r^*(t)$ which in view of equations 1 and 6 may be written as

$$r^*(t)=\frac{1}{T}\sum_{n=-\infty}^{\infty} \epsilon^{(s+jn\omega_0)t} \tag{20}$$

Operating on this expression with the transfer function $G(s)$ gives

$$c(t)=\frac{1}{T}\sum_{n=-\infty}^{\infty} G(s+jn\omega_0)\epsilon^{(s+jn\omega_0)t} \tag{21}$$

Sampling $c(t)$ and taking note of equation 18 yields after minor simplifications the expression for the sampled response of the system to $r(t)=\epsilon^{rt}$; that is

$$c^*(t)=[G^*(s)\epsilon^{st}]\delta_T(t) \tag{22}$$

where $\delta_T(t)$ denotes a train of unit impulses. It is seen that the response has the form of a train of impulses whose envelope is the bracketed term in equation 22. More specifically, this means that $G^*(s)\epsilon^{st}$ represents the envelope of the response of the system to an input of the form ϵ^{st}. Consequently, it may be concluded that the starred transfer function $G^*(s)$ relates the input $r(t)$ and the envelope of the sampled output $c^*(t)$ in the same manner as the transfer function $G(s)$ relates the input and output of N.

Another important property of $G^*(s)$ concerns the impulsive response of N which is denoted by $g(t)$. Since $G(s)$ is the Laplace transform of $g(t)$, it follows at once from equations 7 and 18 that $G^*(s)$ is the z-transform of $g(t)$, that is

$$G^*(s)=\mathcal{L}\{g(t)\delta_T(t)\} \tag{23}$$

Consequently, $G^*(s)$ may be expressed in terms of the values of $g(t)$ at the sampling instants $t_n=nT$ or, alternatively, in terms of the system function $G(s)$ via equation 18. Needless to say, $G^*(s)$ may be obtained directly from either $g(t)$ or $G(s)$ by the use of a table of z-transforms such as Table I.

It frequently happens that the system N consists of a tandem combination of two or more systems. In particular, if N consists of two networks N_1 and N_2 with respective transfer functions $G_1(s)$ and $G_2(s)$, then the transfer function of N is given by the usual relation

$$G(s)=G_1(s)G_2(s) \tag{24}$$

and correspondingly the associated starred transfer function is given by

$$G^*(s)=\frac{1}{T}\sum_{n=-\infty}^{\infty} G_1(s+jn\omega_0)G_2(s+jn\omega_0) \tag{25}$$

which in abbreviated form will be written

$$G^*(s)=G_1G_2^*(s) \tag{26}$$

The important point noted here is that the starred transfer function $G^*(s)$ of two cascaded linear systems N_1 and N_2 which are not separated by a sampler is not the product of the respective starred transfer functions $G_1(s)$ and $G_2(s)$ but rather a new transfer function given by equation 25. On the other hand, if N_1 and N_2 are separated by a sampler, as shown in Figure 4, then from equation 19 it follows at once that the z-transform of the output of N_2 is $G_1^*(s)G_2^*(s)R^*(s)$. Hence, in this case the over-all starred transfer function is

$$G^*(s)=G_1^*(s)G_2^*(s) \tag{27}$$

Consequently, it may be concluded that the over-all starred transfer function of two or more networks cascaded through samplers is equal to the product of the starred transfer functions of the individual networks.

The expressions for the z-transforms of the output of more complex structures such as those encountered in feedback

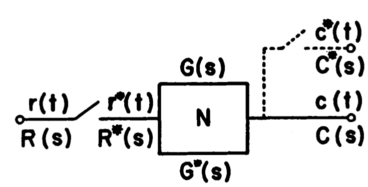
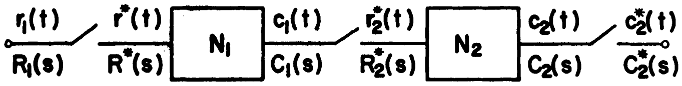


Figure 3. Pulsed linear system showing important variables and their transforms



$$C_2^*(s)=R_1^*(s)\,G_1^*(s)\,G_2^*(s)$$

Figure 4. Cascaded linear systems separated by sampler

systems in general can be derived in a similar manner. Such expressions for several basic types of sampled-data systems are given in Table II. In this table, the first column gives the basic structure, the second gives the expression for the Laplace transform of the output, and the third gives the z-transform of the output.

It will suffice to go through the derivation of $C(s)$ and $C^*(s)$ for a typical system, say number 6 in the table. On inspection of the block diagram, it is seen that the expression for the Laplace transform of the error is

$$E(s)=R(s)-H(s)C^*(s) \tag{28}$$

where $C^*(s)$ is the Laplace transform of the input to the feedback circuit or, equivalently, the z-transform of the output of the system.

The Laplace transform of the output $C(s)$ is related to $E(s)$ by

$$C(s)=G(s)E(s) \tag{29}$$

Combining this relation with equation 28 gives

$$C(s)=G(s)R(s)-G(s)H(s)C^*(s) \tag{30}$$

Applying the z-transformation (see equation 19 and following) to both sides of this equation results in

$$C^*(s)=GR^*(s)-GH^*(s)C^*(s) \tag{31}$$

which upon solving for $C^*(s)$ yields the z-transform of the output

$$C^*(s)=\frac{GR^*(s)}{1+GH^*(s)} \tag{32}$$

as given in Table II. Finally, substituting this expression into equation 30 gives the Laplace transform of the output

$$C(s)=G(s)\left[R(s)-\frac{H(s)GR^*(s)}{1+GH^*(s)}\right] \tag{33}$$

which is the expression for $C(s)$ listed in the table.

As an illustration of the use of the expressions given in Table II, a typical problem involving the fourth structure shown in the table will be considered. Suppose that $H(s)=1$ and $G(s)=K/(s+a)$, where K and a are constants, and that it is desired to determine the values of the response of the system to a unit step input at the sampling instants. To this end, the z-transform of the output must be found. From Table II, this is

$$C^*(z)=\frac{G^*(z)R^*(z)}{1+GH^*(z)} \tag{34}$$

Referring to Table I, the z-transform of the unit step is

$$R^*(z)=\frac{1}{1-z^{-1}} \tag{35}$$

From the same table, the z-transform associated with $G(s)$ is

$$G^*(z)=\frac{K}{1-\epsilon^{-aT}z^{-1}} \tag{36}$$

Substituting these expressions in equation 34, there results after minor simplifications

$$C^*(z)=\frac{K}{(1-z^{-1})(1+K-\epsilon^{-aT}z^{-1})} \tag{37}$$

Expanding this into partial fractions

$$C^*(z)=\frac{K}{(1+K-\epsilon^{-aT})(1-z^{-1})}-\frac{K\epsilon^{-aT}}{(1+K-\epsilon^{-aT})(1+K-\epsilon^{-aT}z^{-1})} \tag{38}$$

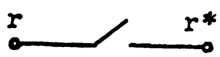
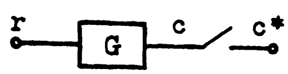
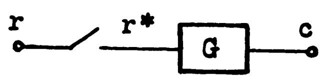
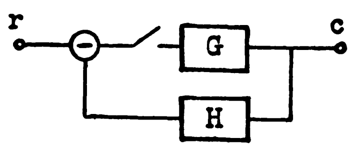
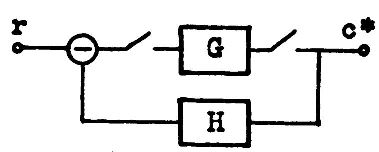
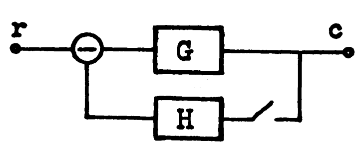
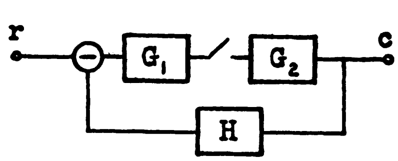
finding the respective inverse transforms from Table I and combining the results yields

$$c(t)=\frac{K}{(1+K-\epsilon^{-aT})}\left\{1-\left(\frac{\epsilon^{-aT}}{1+K}\right)^{\frac{t+T}{T}}\right\} \tag{39}$$

This function coincides with the actual output at the sampling instants and hence at the nth instant, $t_n=nT$, the value of the response to a unit step is

$$c(nT)=\frac{K}{(1+K-\epsilon^{-aT})}\left\{1-\left(\frac{\epsilon^{-aT}}{1+K}\right)^{n+1}\right\} \tag{40}$$

Table II. Output Transforms for Basic Sampled-Data Systems

	System	Laplace Transform of Output $C(s)$	z-Transform of Output $C^*(z)$
(1)		$R^*(s)$	$R^*(z)$
(2)		$GR^*(s)$	$GR^*(z)$
(3)		$G(s)R^*(s)$	$G^*(z)R^*(z)$
(4)		$\frac{G(s)R^*(s)}{1+HG^*(s)}$	$\frac{G^*(z)R^*(z)}{1+HG^*(z)}$
(5)		$\frac{G^*(s)R^*(s)}{1+H^*(s)G^*(s)}$	$\frac{G^*(z)R^*(z)}{1+H^*(z)G^*(z)}$
(6)		$G(s)\left[R(s)-\frac{H(s)RG^*(s)}{1+HG^*(s)}\right]$	$\frac{RG^*(z)}{1+HG^*(z)}$
(7)		$\frac{G_2(s)RG_1^*(s)}{1+HG_1G_2^*(s)}$	$\frac{G_2^*(z)RG_1^*(z)}{1+HG_1G_2^*(z)}$

A sequence of ordinates obtained by evaluating $c(nT)$ for successive values of n yields a graph which can be used to assess the transient performance of the system. It will be noted that the system is stable for all K and a such that $\epsilon^{-aT}<1+K$. A brief discussion of the question of stability will be given in a subsequent section.

Variable Network Approach

By employing the techniques discussed in the preceding section one can obtain, in most practical cases, an explicit expression for the z-transform $C^*(s)$ and, if need be, the Laplace transform $C(s)$ of the output of a specified sampled-data system. The former can be used to find the values of the output at the sampling instants. The latter may be used, in principle, to determine the output at all times by calculating the inverse Laplace transform of $C(s)$. In practice, however, the inversion of $C(s)$ is difficult because $C(s)$ is a rational function in both s and ϵ^{sT} and no tables of inverse transforms for such functions are available.

An alternative approach which works quite well in those cases where an approximate expression for the continuous output—and not just its values at the sampling instants—is desired, is based on treating a sampling system as a periodically varying linear network. This approach is developed in the sequel, following a brief introductory discussion of the frequency analysis technique of handling time-variant systems.[14]

In using the frequency analysis technique, a linear time-variant system N is characterized by its system function $K(s;t)$, which is defined by the statement that $K(s;t)\epsilon^{st}$ represents the response of N to an exponential input ϵ^{st}. If the system function of N is known, then the response of N to an arbitrary input $r(t)$ can be obtained by superposition. More specifically, the output is given by

$$c(t)=\mathcal{L}^{-1}\{K(s;t)R(s)\} \tag{41}$$

where $\mathcal{L}^{-1}$ represents the inverse Laplace transformation and $R(s)$ is the Laplace transform of $r(t)$. The variable t in $K(s;t)$ should be treated as if it were a parameter. This implies that in evaluating the inverse Laplace transform of $K(s;t)R(s)$, one may use a standard table of Laplace transforms and treat t in $K(s;t)$ as a constant.

When N varies in time with period T, its system function $K(s;t)$ is likewise a periodic function of time with period T. Consequently, $K(s;t)$ may be expanded into a Fourier series of the form

$$K(s;t)=\sum_{n=-\infty}^{\infty} K_n(s)\epsilon^{jn\omega_0 t} \tag{42}$$

where $\omega_0=2\pi/T$ and the $K_n(s)$ represent the coefficients of the series. Thus, in the case of a periodically varying network, the problem of determination of $K(s;t)$ may be reduced to that of finding the coefficients of the Fourier series expansion of $K(s;t)$. In practice, a few terms in equation 42 usually provide an adequate approximation to $K(s;t)$, so that in many cases only two or at most three coefficients in the Fourier series expansion of $K(s;t)$ need be determined. (Note that $K_{-n}(s)$, is the conjugate of $K_n(s)$ since $K(s;t)$ is a real function of time.) Once $K(s;t)$—or, rather, an approximation to it—has been determined, the system function $K(s;t)$ can be used in the conventional manner for the purpose of obtaining the response of N to a specified input, for the investigation of the stability of N, for the calculation of the mean-square value of the response of N to a random input, for the determination of the ripple in the output, and many other purposes that are not pertinent to the present analysis.

The application of the general approach just outlined to the analysis of a sampled-data system is quite straightforward. For simplicity, the third system in Table II will be considered first. An input of the form ϵ^{st} is transformed by the sampler into a series of exponential terms which may be written as

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} \epsilon^{(s+jn\omega_0)t} \tag{43}$$

The response of the network N (following the sampler) to this input is

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} G(s+jn\omega_0)\epsilon^{(s+jn\omega_0)t} \tag{44}$$

where $G(s)$ is the system function (transfer function) of N. Consequently, from the definition of the system function $K(s;t)$ of the over-all system, it follows that

$$K(s;t)=\frac{1}{T}\sum_{n=-\infty}^{\infty} G(s+jn\omega_0)\epsilon^{jn\omega_0 t} \tag{45}$$

which is in effect a complex Fourier series expansion of $K(s;t)$. It is seen that the coefficient $K_n(s)$ of $\epsilon^{jn\omega_0 t}$ is equal to $1/T\ G(s+jn\omega_0)$.

As a simple illustration of the use of this expression suppose that the input is a unit-step, that $G(s)=1/(s+a)$, and that ω_0 is such that $K(s;t)$ is adequately approximated by the first two terms in equation 45. In this case, $R(s)=1/s$ and

$$K(s;t)=\frac{1}{T}\left[\frac{1}{s+a}+\frac{\epsilon^{j\omega_0 t}}{s+a+j\omega_0}+\frac{\epsilon^{-j\omega_0 t}}{s+a-j\omega_0}\right] \tag{46}$$

Substituting these expressions in equation 41 and performing the inverse Laplace transformation with the help of a standard table of Laplace transforms, with t treated as a constant, one readily obtains

$$c(t)=\frac{1}{Ta}(1-\epsilon^{-aT})+\frac{2}{T(a^2+\omega_0^2)}\times [a\cos\omega_0 t+\omega_0\sin\omega_0 t-a\epsilon^{-aT}] \tag{47}$$

which is the desired expression for the output.

Turning to feedback systems, consider the fourth system in Table II. In this case, it is expedient to obtain first the expression for $e^*(t)$ corresponding to an exponential input $r(t)=\epsilon^{st}$. In view of equation 22, this is

$$e^*(t)=\frac{\epsilon^{st}}{1+GH^*(s)}\,\frac{1}{T}\sum_{n=-\infty}^{\infty}\epsilon^{jn\omega_0 t} \tag{48}$$

where the second factor represents $\delta_T(t)$. To deduce $K(s;t)$ from this expression, it is sufficient to find the response of the forward circuit, characterized by $G(s)$, to $e^*(t)$ and divide the result by ϵ^{st}. This yields

$$K(s;t)=\frac{1}{T[1+GH^*(s)]}\sum_{n=-\infty}^{\infty} G(s+jn\omega_0)\epsilon^{jn\omega_0 t} \tag{49}$$

which is in effect a complex Fourier series expansion of $K(s;t)$, with the coefficient of $\epsilon^{j\omega_0 t}$ being

$$K_n(s)=\frac{G(s+jn\omega_0)}{T[1+GH^*(s)]} \tag{50}$$

It will be noted that, as should be expected, at the sampling instants $t_m=mT$, $K(s;t)$ reduces to

$$K(s;mT)=\frac{G^*(s)}{1+GH^*(s)} \tag{51}$$

which will be recognized as the starred transfer function of the over-all system.

In the case under consideration, the determination of the response of the system to a given input is complicated somewhat by the fact that the denominator of $K_n(s)$ is a rational function in ϵ^{sT}, rather than in s. For purely numerical computations this is generally not objectionable. However, in analytical work it is usually necessary to approximate the term $GH^*(s)$ in equation 50 by a few terms in its expansion

$$GH^*(s)=\frac{1}{T}\sum_{n=-\infty}^{\infty} G(s+jn\omega_0)\times H(s+jn\omega_0) \tag{52}$$

In this way, one obtains a rational function approximation to $K_n(s)$.

In the final analysis, the results obtainable by the method outlined in this section are also obtainable, although less conveniently, from the expression for the Laplace transform of the output $C(s)$. The chief advantage of the system function approach is that the function $K(s;t)$ is essentially a time-varying transfer function and as such provides a clear picture of the state of the system at each instant.

Response to a Random Input

The expression for the system function obtained in the preceding section has an immediate application in connection with the important problem of determining the statistical characteristics of the response of a sampled-data system to a random input. Since a general discussion of this problem is outside the scope of the present paper, the following analysis is limited to the case where the input is a stationary time series, and it is desired to obtain the expression for the mean-square value of the output at a specified instant of time. This particular problem has considerable bearing on the design of sampled-data systems that are optimum in the sense of the minimum rms error criterion.

A general expression for the mean-square value of the output of a time-variant system is given in reference 14; it reads

$$\sigma^2(t) = \int_0^\infty |K(j\omega;t)|^2 S(\omega) df, \quad \omega = 2\pi f \tag{53}$$

where $S(\omega)$ is the power spectrum function of the input; $K(j\omega;t)$ is the system function with s replaced by $j\omega$; and $\sigma^2(t)$ is the mean-square value of the output at a specified instant t. To apply this equation to a sampled-data system it is only necessary to substitute the expressions for the system function of the system and the power spectrum of the input into equation 53 and carry out the necessary integration. In the majority of practical cases, the integration in question is most readily carried out by graphical means.

When the specified instant of time t coincides with a sampling instant, $t = nT$, the formula given above assumes a much simpler form. Thus, as was shown in the preceding section, for $t = nT$ the system function $K(s;t)$ becomes identical with the starred transfer function $K^*(s)$ of the over-all system. Consequently, equation 53 reduces to

$$\sigma^2 = \int_0^\infty |K^*(j\omega)|^2 S(\omega) df \tag{54}$$

where $K^*(j\omega)$ is the starred transfer function with s replaced by $j\omega$; and σ^2 is the mean-square value of the output at any sampling instant.

Sometimes it is more convenient to express σ^2 in terms of the autocorrelation function[2] of the input, rather than in terms of its power spectrum. It can be shown readily that, in terms of the autocorrelation function of the input, $\psi(\tau)$, σ^2 is given by the following expression

$$\sigma^2 = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} k(nT)k(mT)\psi[(n-m)T] \tag{55}$$

where $k(t)$ is the inverse z-transform of $K^*(z)$. This expression is useful chiefly in those cases where the autocorrelation function $\psi(\tau)$ drops off rapidly with increase in τ.

The Hold System

As stated previously, the function of the hold circuit is to reconstruct approximately the original time function from the impulse train generated by the sampler. It is evident that if it were possible to realize a perfect hold circuit, a sampled-data system incorporating such a circuit would be identical with a continuous-data system. However, in general, a perfect hold circuit is not realizable because of the random nature of the time function which has to be reconstructed. Furthermore, a very important consideration in the design of hold circuits is the fact that a close approximation of the original time function requires, in general, a long time delay, which is undesirable in view of its adverse effect on the stability of the system. Consequently, the design of a hold circuit involves a compromise between the requirements of stability and over-all dynamic performance on the one hand, and on the other hand, the desirability of a close approximation to the original time function and the reduction of ripple content in the output of the system.

It should be remarked that the hold circuits commonly employed in practice are generally of the so-called clamp type, which is one of the simplest forms of hold circuits. More sophisticated types of hold circuits based on the use of polynomial interpolating functions have been described by Porter and Stoneman.[7]

A complete treatment of hold circuits cannot ignore the random nature of the time function which the hold circuit is called upon to reconstruct. Such a treatment is outside the scope of the present paper. The brief discussion which follows is concerned chiefly with some of the more basic aspects of hold circuit design.

The generation of an approximation to the original time function between two

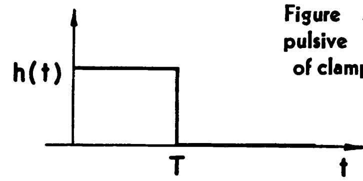


Figure 5. Impulsive response of clamp circuit

sampling instants t_n and t_{n+1}, from its values at the preceding sampling instants $t_n, t_{n-1}, t_{n-2}, \ldots$, is essentially a problem in extrapolation or prediction. An effective, though not optimum, method of generating the desired approximation is based on the consideration of the power series expansion of $r(t)$ in the typical interval from $t_n = nT$ to $t_{n+1} = (n+1)T$

$$r(t) = r(nT) + r'(nT)(t-nT) + \frac{r''(nT)}{2}(t-nT)^2 + \ldots \tag{56}$$

where the primes indicate the derivatives of $r(t)$ at $t_n = nT$. To evaluate the coefficients of this series it is necessary to obtain the derivatives of the function $r(t)$ at the beginning of the interval in question. Since the information concerning $r(t)$ is available only at the sampling instants, these derivatives must be estimated from the sampled data. For instance, an estimate of the first derivative involving only two data pulses is given by

$$r'(nT) = \frac{1}{T}\{r[nT] - r[(n-1)T]\} \tag{57}$$

and the second derivative is given by

$$r''(nT) = \frac{1}{T}\{r'[nT] - r'[(n-1)T]\} = \frac{1}{T^2}\{r[nT] - 2r[(n-1)T] + r[(n-2)T]\} \tag{58}$$

Thus to obtain an estimate of a derivative of $r(t)$ the minimum number of data pulses which must be considered is equal to the order of the desired derivative plus one. This implies that the higher the order, the greater the delay before a reliable estimate of that derivative can be obtained. For this reason, an attempt to utilize the higher order derivatives of $r(t)$ for purposes of extrapolation meets with serious difficulties in maintaining system stability. Generally, only the first term in equation 56 is used, resulting in what is sometimes described as a boxcar or clamp circuit but which will be referred to here as a zero-order hold system. More generally, an nth order hold system is one in which the signal between successive sampling instants is approximated by an nth order polynomial.

Considering the zero-order hold system, it is evident that its impulsive response $h(t)$ must be as shown in Figure 5. By inspection, the Laplace transform of this time function is seen to be

$$H(s)=\frac{1}{s}-\frac{1}{s}\epsilon^{-Ts} \tag{59}$$

The frequency response of this hold system is obtained by replacing the complex frequency s by $j\omega$, resulting in the following expression for the magnitude of the transfer function

$$|H(j\omega)|=\frac{2}{\omega}\left|\sin\frac{\omega T}{2}\right| \tag{60}$$

This relation is plotted in Figure 6. It is observed that the hold system is essentially a low-pass filter which passes the low frequency spectrum of the impulse train and rejects the displaced high-frequency spectra resulting from the sampling process. An important property of the zero-order hold system is that the ripple at the output is zero if the input is a constant. In a similar way, the first order hold system has zero ripple output for an input function whose slope is a constant.

As is well known, if the signal does not contain frequencies higher than one-half the sampling frequency, perfect reproduction of the signal is obtained with an ideal low-pass filter (that is, one with unity gain and linear phase shift up to its cutoff frequency) whose cutoff frequency is equal to one-half the sampling frequency. For that matter, any low-pass network having roughly this characteristic can be used to extract most of the useful spectrum from the impulse train. It is even possible, though not advisable, to dispense with the hold circuit altogether, and rely on the low-pass characteristics of the forward circuit to perform the necessary smoothing of the sampled data.

System Design Using z-Transforms

The primary objectives in the design of closed-cycle control systems include the achievement of stability and an acceptable over-all dynamic performance. The standard technique of achieving these objectives consists in plotting the Nyquist diagram of the loop transfer function and adjusting the system parameters until the plot does not enclose the point -1.0. The margins by which such an enclosure is avoided constitute a measure of the damping of the system. Other more sophisticated techniques may be used if a better dynamic performance is desired.

In view of the fact that the starred transfer function is analogous to the conventional transfer function, the same basic technique may be applied in the case of sampled-data control systems. The function which is plotted is $GH^*(s)$ or $G^*(s)H^*(s)$ according as the denominator of the z-transform of the output is of the form $1+GH^*(s)$ or $1+G^*(s)H^*(s)$, see Table II. As usual, the complex frequency s is varied along a contour in the s-plane consisting of the imaginary axis and a semicircle enclosing the right-half of the plane. Since the starred transfer functions are periodic with period $j\omega_0$, the loci of the functions $GH^*(s)$ or $G^*(s)H^*(s)$ retrace themselves at each cycle, so that s need be varied only from $-j\omega_0/2$ to $j\omega_0/2$ in order to obtain the shape of the locus. Critical regions such as the vicinity of the origin in the s-plane are handled in the same manner as in the case of continuous-data systems.

Since the auxiliary variable z is defined as ϵ^{Ts}, it is evident that as s is assigned imaginary values over one complete cycle, z traces a unit circle in the z-plane. Thus, the loop transfer function $GH^*(s)$ can be plotted directly by expressing it in the form $GH^*(z)$ and varying z along the unit circle. To demonstrate the technique, the $GH^*(z)$ locus for the system shown in Figure 7 will be plotted. This system is seen to consist of a zero-order hold circuit and a simple linear component in the forward circuit. The feedback transmission is unity so that the loop transfer function is

$$G(s)H(s)=\frac{(1-\epsilon^{-s})}{s}\,\frac{1}{s(s+1)},\ T=1 \tag{61}$$

Expanding this transform into partial fractions, there results

$$G(s)H(s)=(1-\epsilon^{-s})\left(\frac{1}{s^2}-\frac{1}{s}+\frac{1}{s+1}\right) \tag{62}$$

On considering each term separately and obtaining the corresponding z-transforms from Table I, the starred loop transfer function is found to be

$$GH^*(z)=(1-z^{-1})\left[\frac{z^{-1}}{(1-z^{-1})^2}-\frac{0.632z^{-1}}{(1-z^{-1})(1-0.368z^{-1})}\right] \tag{63}$$

which can be simplified to

$$GH^*(z)=\frac{0.264+0.368z}{(0.368-z)(1-z)} \tag{64}$$

This function is plotted in Figure 8 where it is seen that for the constants chosen, the system is stable; but if the loop gain is increased by a factor of 1.5, the system becomes unstable.

If it were desired to improve the margin of stability or, for that matter, stabilize this system with a higher loop gain, the procedure would be to add lead networks just as in the case of a continuous-data system. The major difficulty encountered in this procedure is that the resulting starred loop transfer function is not related in a simple manner to the original function, for as was shown previously the starred transfer function of two networks in tandem is not equal to the product of respective starred transfer functions. Consequently, the insertion of a corrective network in the feedback loop requires the recalculation of the starred loop transfer function. The need for recalculation is inherent in the stabilization of sampled-data systems by the insertion of a corrective network in the feedback loop. Thus, as the art exists at the present time, the shaping procedure involves essentially a trial and error method with the plot of the Nyquist diagram for each trial set of system parameters used to assess the stability of the system.

Conclusions

The sampled-data feedback control system may be analyzed in a systematic manner by applying the z-transform method. Techniques of locus shaping similar to those commonly used in continuous systems may be applied by plotting the starred loop transfer function on the complex plane. This is done with no approximations other than those relating to the narrowness of the pulses constituting the pulse train. Once satisfactory loci are obtained by the addition of ap-

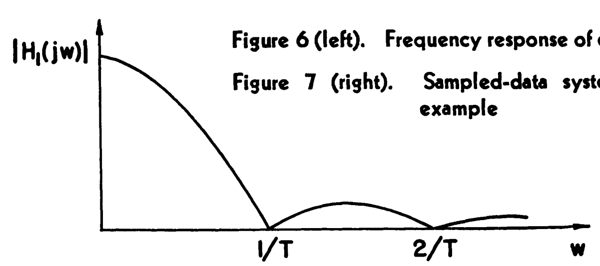
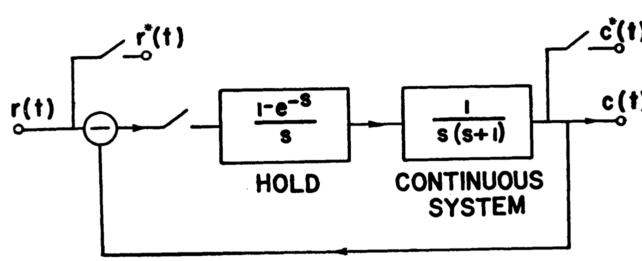


Figure 6 (left). Frequency response of clamp circuit

Figure 7 (right). Sampled-data system used in example



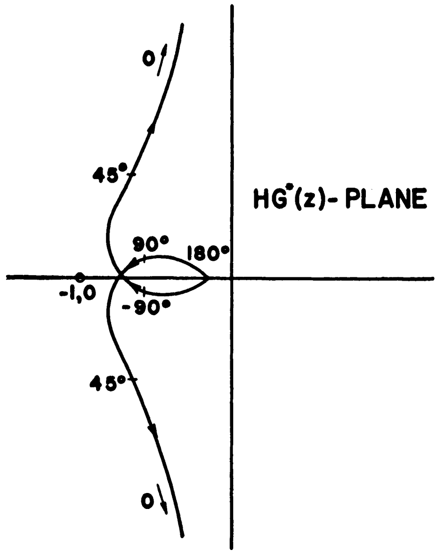


Figure 8. Plot of HG*(z) locus for system used in example

propriate networks, the transient performance of the system can be assessed by obtaining the time functions from a table of z-transforms or by expanding the z-transform of the output function into a power series which gives the ordinates at the sampling instants. The smoothness of the output can be estimated by use of the variable network analysis described in this paper.

One complication in the design of sampled-data systems is the relative difficulty of evaluating the effect of inserting corrective networks in the control loop. If this complication could be removed, locus shaping would be no more difficult than with continuous systems. Research now in progress is directed toward the devising of practical methods of locus shaping in the loop transfer function plane. Results to date indicate that the z-transform method, in conjunction with design techniques somewhat analogous to those used with conventional servomechanisms, furnishes a powerful tool for the analysis of linear sampled-data systems.

References

1. Fundamental Theory of Servomechanisms (book), **L. A. MacColl.** D. Van Nostrand Company, Inc., New York, N. Y., chapter 10, 1945.

2. Theory of Servomechanisms (book), **H. M. James, N. B. Nichols, R. S. Phillips.** McGraw-Hill Book Company, Inc., New York, N. Y., chapter 5, 1947.

3. Sampled-Data Control Systems Studied Through Comparison of Sampling with Amplitude Modulation, **W. K. Linvill.** *AIEE Transactions*, volume 70, part II, 1951, pages 1779–88.

4. Analysis of Discontinuous Servomechanisms, **F. H. Raymond.** *Annales de Telecommunication* (Paris, France), volume 4, 1949, pages 250–56, 307–14, 347–57.

5. Theory of Intermittent Regulation, **Y. Z. Tsipkin.** *Avtomatika i Telemekhanika* (Moscow, USSR), volume 10, number 3, 1949, pages 189–224.

6. The Dynamics of Automatic Controls (book), **R. C. Oldenbourg, H. Sartorius.** American Society of Mechanical Engineers New York, N. Y., 1948, chapter 5.

7. A New Approach to the Design of Pulse-Monitored Servo Systems, **A. Porter, F. Stoneman.** *Journal*, Institution of Electrical Engineers (London, England), volume 97, part II, 1950, pages 597–610.

8. A General Theory of Sampling Servo Systems, **D. F. Lawden.** *Proceedings*, Institution of Electrical Engineers (London, England), volume 98, part IV, October 1951, pages 31–36.

9. Application of Finite Difference Operators to Linear Systems, **B. M. Brown.** *Proceedings of the D.S.I.R. Conference on Automatic Control*, Butterworths Scientific Publications (London, England), 1951.

10. Analysis of a Sampling Servomechanism, **K. S. Miller, R. J. Schwarz.** *Journal of Applied Physics* (New York, N. Y.), volume 21, April 1950, pages 290–294.

11. Theorie Analytique des Probabilités, Part 1: Du Calcul Des Fonctions Generatrices (book), **P. S. Laplace.** Paris, France, 1812.

12. A List of Generalized Laplace Transforms, **W. M. Stone.** *Journal of Science*, Iowa State College (Ames, Iowa), volume 22, April 1948, pages 215–225.

13. Transients in Linear Systems (book), **M. F. Gardner, J. L. Barnes.** John Wiley and Sons, Inc., New York, N. Y., 1942, pages 354–356.

14. Frequency Analysis of Variable Networks, **L. A. Zadeh.** *Proceedings*, Institute of Radio Engineers (New York, N. Y.), volume 38, March 1950, pages 291–299.

Discussion

John M. Salzer (Hughes Aircraft Company, Culver City, Calif.): The objectives of this paper are: (a) to unify the z-transform (Shannon, Hurewicz) and s-transform (Linvill) approaches to sampled-data systems; (b) to formulate input-output relations for various types of such systems; and (c) to treat such systems by considering the sampler a time-variant element.

The paper fills a clear need in this underpublished field, and it presents its topic concisely and illuminatingly. It brings together two viewpoints, (a) and (c), to bear on the same problem and shows the way toward systematization (b) of the solution.

Perhaps the most significant contribution of the paper is the variable network approach. Although the applicability of this approach was recognized before,[1] the authors' more recent investigations in the general field of time-varying systems are now made to bear directly on the problem of sampled-data systems. The importance of the generalization thus afforded should not be overlooked.

It is interesting to note that the variable network approach leads directly to the same expressions as the s-transform method. Both methods are predicated on the concept of using an arbitrary, characteristic input function $[r(t) = \epsilon^{st}]$, but by distinguishing a time-variant transfer function certain generalizations are made possible, as already noted. An interesting illustration of such a generalization is given in the section titled "Response to a Random Input."

There are several points and results in this paper which bear interesting relations to some of the work done by the discussor in connection with digital-analogue systems.[2] The use of starred s-transforms, such as $R^*(s)$, in place of infinite sums, is a welcome convenience.[3] This notation facilitates the manipulations in dealing with sampled-data systems and makes the systematizations of Table II of the paper easier to comprehend.

As noted in the paper, in the case of starred transforms (that is, in the case of sampled functions) the s- and z-transforms are exactly equivalent. Thus, the use of z-transforms involves no approximation, and it offers certain conveniences. One advantage is notational, for z is easier to write than ϵ^{st}. The other is that the infinite number of poles and zeros of the starred s-transforms are replaced by a finite set in the z-plane. Furthermore, the ambiguity at the point of infinity in the s-plane (due to essential singularity) is circumvented by the use of z-transforms.

In case the s-transform is not a starred transform, information is lost by the use of z-transform because this amounts to representing a continuous function, $c(t)$, by its samples, $c^*(t)$, just as it is done in numerical mathematics. But whereas in numerical work the sampling (or tabular) interval may be adjusted until such a representation is justified, in analyzing a given control system one is faced with a sampling rate already determined and the analysis must be made correct for the existing physical situation. For particular systems the sequence of output samples may not give a satisfactory picture so that it becomes necessary to study the behavior of the output also between sampling instants. Furthermore, and this is important, it is not always *a priori* obvious whether this is or is not the case.

Where the output behavior between samples is also of interest, the z-transform method is still applicable but must be augmented by separate investigation of the output during the sampling period. A separate solution based on initial conditions at the sampling instant may be used in analysis, but a synthesis procedure would hardly be fruitful along these lines.

The alternative solution is the s-transform approach. As noted by the authors, this method leads to transforms which are products of rational functions of s and z, and tables of corresponding transform pairs are not available. Nevertheless, the exact analysis is straightforward, even if somewhat laborious, and moreover, the frequency characteristics obtained may give a hint as to the nature of compensation needed to improve the response.

Whether the z-transform method is applicable or not in a particular case depends on the question of bandwidth. If the sampling rate is many times higher than the bandwidth of the input or of the system, then the z-transform solution is expected to be a suitable representation of the continuous output. It is presumed that the applications with which the authors concerned themselves were of this type. However, in the design of certain systems it is often desirable to use the lowest permissible sampling rate consistent with the specifications. In finding this limiting rate, one does not get a complete answer by the use of only z-transforms.

It may be noted that as far as stability is concerned either the z- or the s-transforms lead to exactly the same result, for only a divergent continuous function has a divergent sequence of samples. That this is so is demonstrated in Table II of the paper, where the corresponding denominators in both the s- and z-transform columns are identical.

To the numerous examples of sampled-data systems mentioned by the authors, the discussor wishes to add one: systems in which a digital computer is incorporated. As generally conceived today, digital computers operate on sampled data; therefore, their presence in the system requires sampling. Since the output of the digital computer is sampled also, the computer fits in between the sampler and the holding unit. If the digital computer is instructed to perform a linear difference equation on its sampled input, it can be represented by a transfer function which is rational in z. This result ties in with the methods of the paper.

For example, in the case of a digital computer equation 27 of the paper applies, because the data stay sampled through all numerical work. Thus, the transfer function of a composite digital program equals the product of the transfer functions of the component programs. The implication of this fact in system design is to be noted. Suppose that the stability of the system illustrated in Figure 8 of the paper is to be improved by a digital compensator, $W^*(s)$, inserted in either the forward or the feedback section. Since

$$WHG^*(z) = W^*(z)HG^*(z)$$

the new stability diagram is directly related to the old one so that the synthesis procedure is facilitated greatly. Of course, this is not meant to imply that digital compensation can do a better job than analogue; it only means that it is easier to see what a digital unit does. A somewhat academic example of digital compensation is worked out in chapter 4 of reference 2 of this discussion.

There is one comment concerning notation which may be found of interest: namely, it may be preferable to define z as being equal to ϵ^{-sT} rather than ϵ^{+sT}, when dealing with sampled-data systems. This is so because the latter corresponds to a time-advance operation, which has no physical meaning in a real-time application. In purely mathematical work one definition is as good as the other, and it is just unfortunate that in previous operational and transform work with difference equations the advance (ϵ^{+sT} rather than the delay (ϵ^{-sT}) operator was given a symbol. In consequence of this choice, z naturally will appear raised to negative powers, thereby diminishing the manipulative advantage of its usage. Table I of the paper illustrates this point. Of course, multiplication of both numerator and denominator by an appropriate power of z always can eliminate the negative powers of z (as was done in line 11 of Table I), but this is an additional step.

There are further reasons for which ϵ^{-sT} should be regarded the fundamental variable. For instance, in the investigation of stability of digital programs by conformal mapping the use ϵ^{-sT} leads to much simpler rules that that of ϵ^{+sT}.

REFERENCES

1. ANALYSIS AND DESIGN OF SAMPLED DATA CONTROL SYSTEMS (thesis), **W. K. Linvill.** Project Whirlwind *Report R-170*, Massachusetts Institute of Technology (Cambridge, Mass.), 1949.

2. TREATMENT OF DIGITAL CONTROL SYSTEMS AND NUMERICAL PROCESSES IN THE FREQUENCY DOMAIN (thesis), **John M. Salzer.** Digital Computer Laboratory (microfilmed report), Massachusetts Institute of Technology (Cambridge, Mass.), 1951.

3. Compares with the notation $R(s)$ used in reference 2; for example, pages 56 and 176.

B. M. Brown (Royal Naval College, London, England): An alternative technique is available for handling the general theory and problems of the type discussed in this paper. Instead of using Laplace transforms, which imply a limitation to input and output functions which are zero for negative time, this technique assumes functions of general type, and uses as transfer functions operators which are functions of the operator $D = d/dt$. Such a function $F(D)$ is usually a fraction, in which numerator and denominator are polynomials in D. Thus if input and output, u and x respectively, are connected by the relation $x=F(D)u$, where $F(D)=P(D)/Q(D)$, then x is understood to denote the general solution of the differential equation

$$Q(D)x=P(D)u$$

Such operators have provided a classical method for solving linear differential equations. If the coefficients are constant the operators in general can be manipulated algebraically, and this type of manipulation proves to be a very powerful tool.

If a linear system has a transfer function $F(D)$, then the stability is determined by the roots of $Q(\lambda)=0$. A steady-state solution can be obtained by expanding $F(D)$ in a series of ascending powers of D and operating on u. In many cases an adequate approximation is given by the first term of this series, which always can be written down by inspection. Thus if $F(D)$ is given, the main characteristics of the response can be inferred without the labour of evaluating particular solutions to particular inputs, whether by Laplace transform or other methods.

Operators of the type $F(D)$ have parallels in the form of functions of the operators E and Δ, with a complete set of analogous properties. These operators are defined by

$$Eu(t)=u(t+T)$$

and

$$\Delta u(t)=u(t+T)-u(t)$$

where T is constant. The three basic operators are connected by the relations

$$E=1+\Delta=\epsilon^{TD}$$

the latter being obtained by Taylor's theorem. Functions of E and Δ can be used for solving difference equations and for discussing systems based on such equations. A short account of the appropriate methods is given in reference 9 of the paper.

Now operators of the types $F(D)$, $F(E)$, and $F(\Delta)$ are all linear, which is to say that, denoting such an operator by Φ

$$\Phi(u+v)=\Phi u+\Phi v$$

They also have another important property of being time-invariant, which means that if $\Phi u(t)=x(t)$, then $\Phi u(t+t_0)=x(t+t_0)$, where t_0 is any constant. Because of this, they are commutative. They can be, and are in practice, applied both to continuous functions of t and to pulse trains.

Another operator S, the sampling operator, now is introduced whose effect is to convert a continuous function into a train of pulses. It is of course equivalent to multiplication by $\delta_T(t)$ in the notation of the paper. It easily is seen that S is linear, but not time-invariant and not in general commutative with functions of D, E, or Δ.

It will be shown briefly how the various processes described in the paper can be represented operationally. Consider first the effect of sampling a function $u(t)$ and then passing it through a network with transfer function $F(D)$. The resulting function is $F(D)Su(t)$ which is in general a continuous function. If this is sampled, either physically or for the purpose of analysis, the result is the pulse train $SF(D)Su(t)$. To determine the relation between the two pulse trains, let $f(t)$ be the response to a unit impulse input to the network. Then

$$\begin{aligned} SF(D)Su(t) &= SF(D)\sum_m u(mT)\delta(t-mT) \\ &= S\sum_m u(mT)F(D)\delta(t-mT) \\ &= S\sum_m u(mT)f(t-mT) \\ &= \sum_n\sum_m u(mT)f[(n-m)T]\delta(t-nT) \\ &= \sum_n\sum_k f(kT)u[(n-k)T]\delta(t-nT), \\ &\qquad\qquad \text{putting} \quad n-m=k \\ &= \sum_n\left\{\sum_k f(kT)E^{-k}u(nT)\right\}\delta(t-nT) \\ &\doteq \left\{\sum_k f(kT)E^{-k}\right\}Su(t) \end{aligned}$$

The operator in brackets is a function of E. If it is denoted by $F^*(E)$ we have the operational relation

$$SF(D)S=F^*(E)S \tag{1}$$

where

$$F^*(E)=\sum_k f(kT)E^{-k} \tag{2}$$

Consider the special case when $F(D) = 1/(D+a)$. Then $f(t)=\epsilon^{-at}$, so that

$$F^*(E)=\sum_{k=0}^{\infty}\epsilon^{-kaT}E^{-k}=\frac{1}{1-\epsilon^{-aT}E^{-1}} \tag{3}$$

As in the paper, the general operator can be dealt with by using partial fractions and a table similar to Table I.

It is now apparent that the operator E plays a part analogous to that of the transform variable z, the association being similar to that of D and s. It will be found further that most of the equations involving starred and unstarred functions of s given in the first part of the paper can be expressed in operational form. In particular, equation 1 of this discussion corresponds to equation 13 of the paper, while the equivalent of equation 27 in the paper would be written in the operational form

$$SF_1(D)SF_2(D)S = SF_1(D)F_2^*(E)S = F_1^*(E)F_2^*(E)S \quad (4)$$

It is not easy to make a comparison of the relative merits of the two alternative approaches. So much depends on the notation with which a particular individual is familiar. It was claimed earlier that the use of operators implies greater generality, but this is perhaps of small account from a practical point of view. However, it may be an advantage to represent the operation of sampling by a special symbol.

As an example of the use of operators, consider system 4 in Table II of the paper. The circuit equation is easily seen to be

$$G(D)S[r(t)-H(D)c(t)] = c(t) \quad (5)$$

Operating with $H(D)$ and sampling

$$HG^*(E)[Sr(t)-SH(D)c(t)] = SH(D)c(t)$$

so that

$$SH(D)c(t) = \frac{HG^*(E)}{1+HG^*(E)}Sr(t)$$

Substituting in equation 5 and simplifying gives

$$c(t) = \frac{G(D)}{1+HG^*(E)}Sr(t)$$

and

$$Sc(t) = \frac{G^*(E)}{1+HG^*(E)}Sr(t)$$

It is of interest to observe that the process of clamping can be represented by the operator $(E^{-1}\Delta/D)S$. To prove this it is only necessary to point out that a clamped function is the integral of the sequence formed by the first differences of the sampled function.

William K. Linvill (Massachusetts Institute of Technology, Cambridge, Mass.): This paper makes a concise mathematical summary of the analysis of sampled-data systems. Table II is particularly helpful and illustrates the applicability of the analysis to a wide variety of system configurations. The easy interchangeability between the z-transform method and the Laplace transform method should receive more emphasis than the paper gives. When the sampler output is considered to be a train of modulated impulses rather than a sequence of ordinates of the sampler input, the sampler has all the properties of the familiar pulse-amplitude modulator and can be treated by conventional Laplace transforms. When the whole system is treated in the frequency domain from this point of view, there is no difference between the z-transform approach and the "old-fashioned" Laplace transform approach other than a change in variable $z=\epsilon^{sT}$. When the signals at any point are discrete samples, their transforms are periodic and the $z(=\epsilon^{sT})$ variable is the convenient one to use. When the signals at a point are continuous, their transforms are aperiodic and the s variable is the convenient one to use. The attitude on the part of the engineer that the z-transform and the Laplace transform methods are equivalent and that he can change from one to the other as the situation demands is mathematically correct as well as convenient.

Freedom in changing from the z to the s domain is often particularly convenient in obtaining time responses. For example, the result of equations 46 and 47 of the paper can be seen from elementary considerations in the frequency domain. The sampler is an impulse modulator. Since $G(s)$ represents a low-pass filter, only the pure signal and the pair of lowest frequency side bands from the sampler result in significant output. The pure signal input to the filter $G(s)$ is a step of amplitude $1/T$ and the two low-frequency side bands combine to form a cosine wave: $1/T(\epsilon^{j\omega_0 t}+\epsilon^{-j\omega_0 t})=2/T \cos \omega_0 t$ for $t>0$. The output resulting from the step is $(1/aT)(1-\epsilon^{-at})$ and the output from the cosine wave is calculated from elementary transient theory to be $2/T(a^2+\omega_0^2)\,[a \cos \omega_0 t+\omega_0 \sin \omega_0 t]$ for the steady-state component and $-2a\epsilon^{-at}/T(a^2+\omega_0^2)$ for the transient. Superposing all the output components gives

$$c(t) = \frac{1}{Ta}(1-\epsilon^{-at}) + \frac{2}{T(a^2+\omega_0^2)}[a \cos \omega_0 t + \omega_0 \sin \omega_0 t - a\epsilon^{-at}]$$

The statement that the Laplace transformation procedure requires more labor than the other methods should be modified. To calculate the exact continuous time response is more laborious than to calculate a sequence of samples regardless of the method. The Laplace transform approach embraces both the s-domain and the z-domain pictures and from it the engineer can calculate the sampled response as easily as he can calculate it by any other method, but in addition he can get continuous signals exactly at the cost of considerable labor. Procedures have been worked out for obtaining simply an approximate continuous response and they will be described in a forthcoming paper.

J. R. Ragazzini and **L. A. Zadeh**: The authors wish to thank Dr. Salzer, Dr. Linvill, and Dr. Brown for their constructive discussions.

With regard to Dr. Salzer's comment on the variable network approach, it should be noted that the main feature of this approach is the characterization of a sampled-data system in terms of a transfer function $K(s;t)$ which involves both frequency and time. Such transfer functions do not appear in the report referred to by the discusser.[1]

In defining z as ϵ^{+sT} rather than ϵ^{-sT}, we have been motivated first by a desire to avoid conflict with the notation used by W. Hurewicz and others, and second by the fact that the alternative choice would make it inconvenient to use the only extensive table of z-transforms now available, namely, the table of so-called generalized Laplace transforms compiled by W. M. Stone. Otherwise, we are in complete agreement with Dr. Salzer's suggestion that it would be preferable to define z as being equal to ϵ^{-sT} rather than ϵ^{+sT}.

With regard to Dr. Linvill's statement to the effect that there is no difference between the z-transform and the conventional Laplace transform approaches, we believe that it would be more precise to say that the z-transform approach is a technique of analysis based on the Laplace transformation which is best suited for applications in which it is sufficient to know the values of the output at the sampling instants. By contrast, Dr. Linvill's approach leads to the expression for the Laplace transform of the continuous output, and does not yield directly the values of the output at the sampling instants. The method described in our paper achieves a unification of these approaches in the sense that it furnishes a systematic procedure for determining both the Laplace and z-transforms of the output, as illustrated in Table II. The connecting link between the two approaches is contained in the statement that a relation of the form $C(s)=G(s)R^*(s)$, where $C(s)$ and $G(s)$ are ordinary Laplace transforms and $R^*(s)$ is a starred transform (that is, z-transform), implies the relation $C^*(s)=G^*(s)R^*(s)$.

As we have pointed out in the paper, the expression for the output obtained by the use of the variable network approach also may be obtained from the expression for the Laplace transform of the output, and, in simple cases, the same results may be derived from elementary considerations in the manner indicated in Dr. Linvill's discussion. The main advantage of the variable network approach is that it yields the expression for a time-varying transfer function $K(s;t)$ which constitutes a much more explicit and flexible means of characterizing a system than the expression for the Laplace transform, $C(s)$ of the response to some particular input. Furthermore, $K(s;t)$ is, in general, more convenient to work with than $C(s)$. For example, one can readily express the mean-square value of the response to a random input in terms of $K(s;t)$, but not in terms of $C(s)$.

With reference to Dr. Brown's statement that the use of Laplace transforms implies limitation to input and output functions which vanish for negative time, it should be noted that, when such is not the case, it is merely necessary to employ the Fourier or bilateral Laplace transforms in place of the unilateral Laplace transforms. Thus, the applicability of the methods described in our paper is not restricted to input functions which vanish for negative t.

The operational approach presented by Dr. Brown is related to the z-transform approach in much the same manner as Heaviside's operational calculus is related to the conventional Laplace transformation. In particular, the operator E corresponds to z, the relation $SF(D)S=F^*(E)S$ is equivalent to the relation $C^*(s)=G^*(s)R^*(s)$, and the operation with a sampling operator S corresponds to the z-transformation. It is of interest to note that the operators employed by Dr. Brown may be regarded as special forms of so-called time-dependent Heaviside operators.[2]

A useful feature of the operational approach is that it places in direct evidence the operations performed on the operand. However, working with operators is more difficult than with z-transforms, since the latter require only purely algebraic manipulations. This is indeed the chief advantage of the z-transform method.

References

1. See reference 1 of John M. Salzer's discussion.

2. Time-Dependent Heaviside Operators, L. A. Zadeh. *Journal of Mathematics and Physics* (Cambridge, Mass.), volume 30, 1951, pages 73–78.