RADIO SCIENCE Journal of Research NBS/USNC–URSI
Vol. 68D, No. 9, September 1964

Effect of Linear and Nonlinear Signal Processing on Signal Statistics

A. V. Balakrishnan

Department of Engineering, University of California, Los Angeles, Calif.

(Received January 20, 1964; revised February 18, 1964)

The term "Signal Processing" is interpreted as an operation on the signal that is deliberate (as distinguished from the unavoidable) and arising from:

(a) Optimization: an operation on the data to optimize extraction or detection as in radar or communication or similar applications.

(b) Routine data handling: operations such as sampling and/or quantitizing, scale changes, etc.

(c) Adaptive techniques: operations that are characteristic of adaptive systems where no a priori knowledge of signals and/or system parameters may be available and a self-learning procedure is necessary.

The main interest is in the effect of these operations on the signals; more specifically, if we represent the processor as a black-box, the signal being its "input," we study the statistics of the "output" signal. We examine certain aspects of these problems of recent significance to radio physics.

1. Introduction

The term "Signal Processing" is interpreted as an operation on the signal that is deliberate as distinguished from the unavoidable, as for instance, the effect of atmospheric turbulence on propagation and so on. The deliberate or intentional processing that is of concern can, for the purposes of this paper, be conveniently grouped as arising from:

(a) Optimization: an operation on the data to optimize extraction or detection as in radar or communication or similar applications.

(b) Routine data handling: operations such as sampling and/or quantitizing, scale changes, etc.

(c) Adaptive techniques: operations that are characteristic of adaptive systems where no a priori knowledge of signals and/or system parameters may be available and a self-learning procedure is necessary. While these can be viewed under (a), there are certain unique features of the kind of analysis involved which merit special attention.

Our main interest is in examining the effect of these operations on the signals; more specifically, if we represent the processor as a black-box, the signal being its "input," we wish to study the statistics of the "output" signal. If we interpret this broadly, the input signal being a random function of time or a stochastic process, any functional on a stochastic process can be included under the heading of this paper. This generality of course takes in too much territory to be susceptible to any sensible review and in this paper we shall examine certain aspects of these problems based on what can only be an arbitrary judgment on the part of the author as to what is of recent significance to radio physics. In the mathematics there will be no attempt at maximum of vigor with minimum of hypothesis. It will be assumed, for instance, that the signal is an extremely well-behaved stochastic process with all moments finite and correlations bounded etc., without further ado. On the other hand, on occasion, the signal may be a white noise and no special fuss need be made over this. Well-recognized mathematical techniques are available that are designed to frame our statements with the necessary rigor. This being the case it will be assumed that the interested reader can make the modifications necessary for rigor where this is important.

2. Processing Arising From Optimization

Let us represent the signal by $x(t)$, where "t" may be discrete or continuous and $x(t)$ may be a real or complex variable or an n-dimensional vector. Since the generalizations involved are more or less routine, we shall assume here that $x(t)$ is a real variable to avoid notational complexity. Also "t" will be taken as continuous and to include the discrete or sampled-data case, one has only to replace the integrals in what follows by appropriate sums. Any physical system has at any given time only a finite time-segment of signal or data at its disposal. This fact will also be assumed in what follows.

An integral representation for any operation, linear or nonlinear, that arises in optimization—such as optimal prediction, estimation or detection—can be taken as:

$$y(t)=\int_{t-T}^{t} V_1(t;\sigma;x(\sigma))d\sigma+\int_{t-T}^{t}\int_{t-T}^{t} V_2(t;\sigma_1,\sigma_2;x(\sigma_2))d\sigma_1 d\sigma_2\ldots$$

$$+\int_{t-T}^{t}\ldots\int_{t-T}^{t} V_n(t;\sigma_1,\sigma_2,\ldots\sigma_n;x(\sigma_1),x(\sigma_2)\ldots x(\sigma_n))d\sigma_1 d\sigma_2\ldots d\sigma_n \qquad (2.1)$$

where $V_k(t;\ \sigma_1\ \ldots\ \sigma_k;\ x_1,\ \ldots\ x_k)$ is a function of $(2k+1)$ variables, with $t-T\leq\sigma_1,\ \ldots\ \sigma_k\leq t, -\infty\leq x_1,\ \ldots\ x_k\leq+\infty$. The point is that any optimal operation can be approximated arbitrarily closely by choosing "n" large enough. The modifications necessary to specialize to time-invariant processors are obvious. Our first problem then can be stated as that of obtaining the statistics of the output process $y(t)$, given the statistics of the input process $x(t)$ and the functions $V_k(.\ .\ .)$. No general method is available yet that comes even close to solving this problem. Even if we restrict the processor to be linear, there are no general methods known except of course for the well-known (and now trivial) case where the input signal is a Gaussian process.

There is one method of some generality which is pertinent here. This is the theory of additive functionals on a Markov process [Balakrishnan, 1963, and Fortet, 1958]. In view of the vast literature on it [Fortet, 1958], perhaps it would be unfair to pass it over, although as we shall see, in terms of providing practically useful answers, much remains to be done even here. For this we need to assume that the signal $x(t)$ is Markovian. Let $p(x_1, t_1;\ x_2, t_2)$ be the transition density kernel of the transition to x_2 at time t_2 conditioned on $x(t_1)=x_1$. Let us first consider a system or processor represented by:

$$y(t)=\int_{s}^{t} V(\sigma;\ x(\sigma))d\sigma. \qquad (2.2)$$

The theory develops a method for determining the characteristic function of the random variable $y(t)$. Actually one considers the conditional characteristic function:

$$E\left[e^{iuy(t)}\middle|\begin{matrix}x(t)=x_2\\ x(s)=x_1\end{matrix}\right]$$

and the main result is a pair of linear integral equations for

$$r(x_1,\ s;\ x_2,\ t;\ u)=E\left[e^{iuy(t)}\middle|\begin{matrix}x(t)=x_2\\ x(s)=x_1\end{matrix}\right]p(x_1,\ s;\ x_2,\ t).$$

The integral equations themselves are

$$r(x_1,\ s;\ x_2,\ t;\ u)=p(x_1,\ s;\ x_2,\ t)+iu\int_{s}^{t}dt'\int_{-\infty}^{\infty}r(x',\ t';\ x_2,\ t;\ u)V(t',\ x')p(x_1,\ s;\ x',\ t')dx' \qquad (2.3)$$

$$r(x_1,\ s;\ x_2,\ t;\ u)=p(x_1,\ s;\ x_2,\ t)+iu\int_{s}^{t}dt'\int_{-\infty}^{\infty}r(x_1,\ s;\ x',\ t';\ u)V(t',\ x')p(x',\ t';\ x_2,\ t)dx' \qquad (2.4)$$

where (2.3) is, for obvious reasons, called the "backward" equation and (2.4) the "forward" equation.

Let us now consider some of the drawbacks in this approach. In the first place, the functional (2.2) is not general enough for our purposes. For instance, if the processor or system is time-invariant, then we would need to consider

$$y(t)=\int_s^t V(t-\sigma;x(\sigma))d\sigma$$

and this is of course not reducible to the form (2.2), except in special cases. One such case is where

$$V(t-\sigma;x(\sigma))=Ae^{k(t-\sigma)}V(x(\sigma)) \tag{2.5}$$

since one has only to consider $y(t)e^{kt}$ in place of $y(t)$. For the slightly more general case where the "memory" is all in the linear part of the processor, the linear part coming from a rational transfer function, we need to consider

$$V(t-\sigma;x(\sigma))=\sum_1^m a_ie^{k_i(t-\sigma)}V(x(\sigma)).\;.\;.\;. \tag{2.6}$$

For this case it is possible to obtain an integral equation again but one which is slightly more complicated than (2.3) and (2.4). Since this is new with this paper, we shall sketch the method of derivation as well. Let

$$r(x_1,s;x_2,t;U)=E\left[e^{i\sum_1^m u_ja_j\int_s^t e^{k_j(t-\sigma)}V(x(\sigma))d\sigma}\middle|\begin{matrix}x(t)=x_2\\ \\ x(s)=x_1\end{matrix}\right]$$

where U is the m-vector $(u_1,\;.\;.\;.\;u_m)$. We note that we obtain the characteristic function for $y(t)$ by taking $u_1=u_2\;.\;.\;.\;=u_m$. Let us consider the forward equation. For this, as in the usual derivation of (2.4), we take

$$e^{i\sum_1^m u_ja_j\int_s^t e^{k_j(t-\sigma)}V(x(\sigma))d\sigma}$$

$$=1+\int_s^t dt'\left[i\sum_1^m u_ja_jV(x(t'))+i\sum_1^m u_jk_ja_j\int_s^{t'}e^{k_j(t'-\sigma)}V(x(\sigma))d\sigma\right]e^{i\sum_1^m u_ja_j\int_s^{t'}e^{k_j(t'-\sigma)}V(x(\sigma))d\sigma}.$$

Multiplying by $p(x_1,\,s;\,x_2,\,t)$ and taking conditional expectations, and using the properties of a Markov process in dealing with the integral on the right side, we obtain:

$$r(x_1,s;x_2,t;U)=p(x_1,s;x_2,t)+\int_s^t dt'\int_{-\infty}^{\infty}\left(i\sum_1^m u_ja_j\right)V(x')r(x_1,\,s;\,x',\,t';U)p(x',\,t';\,x_2,\,t)dx'$$

$$+\int_s^t dt'\int_{-\infty}^{\infty}\sum_1^m ik_j\frac{\partial}{\partial u_j}r(x_1,\,s;\,x',\,t';U)p(x',\,t';x_2,\,t)dx'. \tag{2.7}$$

The backward equation can be obtained in an obvious manner.

But even in these cases the problem of a general solution of these equations does not appear to be an easy one. Computer solutions based on iteration (successive approximation) run into the difficulty that the successive approximations have to be restricted to be characteristic functions.

If one has to be satisfied with approximations to the characteristic functions, it may also suffice perhaps to obtain the moments. The moments, fortunately, can be calculated directly from the integral representation. Thus for

$$y(t)=\int_s^t V(t-\sigma;\,x(\sigma))d\sigma$$

we have

$$E[y(t)]=\int_s^t d\sigma \int_{-\infty}^{\infty} V(t-\sigma;\, x)p(x,\, \sigma)dx \tag{2.8}$$

where $p(x, \sigma)$ is the first order density of $x(\sigma)$. Similarly,

$$E[y(t)^2]=\int_s^t d\sigma_1 \int_s^t d\sigma_2 V(t-\sigma_1;\, x_1)V(t-\sigma_2;\, x_2)p(x_1,\, \sigma_1;\, x_2,\, \sigma_2)p(x_1,\, \sigma_1)dx_1dx_2 \tag{2.9}$$

$$E[y(t)^3]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_s^t\int_s^t\int_s^t d\sigma_1 d\sigma_2 d\sigma_3 V(t-\sigma_1;\, x_1)$$
$$V(t-\sigma_2;\, x_2)V(t-\sigma_3;\, x_3)p(x_1\sigma_1;\, x_2,\, \sigma_2;\, x_3,\, \sigma_3)dx_1dx_2dx_3$$

where $p(x_1, \sigma_1; x_2, \sigma_2; x_3, \sigma_3)$ is the joint density of $x(\sigma_1)$, $x(\sigma_2)$, $x(\sigma_3)$, and can of course be further simplified using the Markovian property. However, it is possible to avoid the use of multiple integrals by using (2.4) or the backward equation corresponding to (2.7). For this, let

$$\mu_n[x,\, s]=E[y(t)^n|x(s)=x].$$

We use the backward version of (2.7):

$$r(x_1,\, s;\, x_2,\, t;\, U)=p(x_1,\, s;\, x_2,\, t)+\int_s^t dt' \int_{-\infty}^{\infty}\left(i\sum_1^m u_ja_j\right)V(x')r(x',\, t';\, x_2,\, t;\, U)p(x_1,\, s;\, x',\, t')$$
$$+\int_s^t dt' \int_{-\infty}^{\infty}\left[\sum ik_j \frac{\partial}{\partial u_j} r(x',\, t';\, x_2,\, t;\, u)\right]p(x_1,\, s;\, x',\, t')dx'. \tag{2.10}$$

Let

$$\phi[x_1,\, s;\, t;\, U]=E\left[e^{i\Sigma u_ja_j\int_s^t e^{k_j(t-\sigma)} V(x(\sigma))d\sigma}\Big|x(s)=x_1\right].$$

Then integrating (2.10) with respect to the variable x_2, we have

$$\phi[x_1,\, s;\, t;\, U]=1+\int_s^t dt' \int_{-\infty}^{\infty}\left[\left(i\sum_1^m u;\, a_j\right)V(x')\phi(x',\, t';\, t;\, U)\right.$$
$$\left.+\sum_1^m ik_j \frac{\partial}{\partial u_j}\phi(x',\, t';\, t;\, u)\right]p(x_1,\, s;\, x',\, t')dx'.$$

To obtain a recurrence relationship for the moments we can make a Taylor expansion in U for $\phi(x_1, s; t, u)$ and equate coefficients on both sides. We omit the details here. After we find the moments we still have the problem of approximating the distribution should this be required. Moreover the question is moot as to whether the moments determine the distribution. It would be natural to assume that in all physical processes this would be so if the input signal statistics have the same property, although a general proof of this sort is not available. It is possible to state some sufficient conditions of some generality. Suppose, for instance, we consider

$$y(t)=\int_s^t k(t,\, \sigma)x(\sigma)^m d\sigma.$$

We assume that the process $x(\sigma)$ is such that

$$\sum_0^\infty \int_s^t \frac{M_n(\sigma)d\sigma}{n!} r^n < \infty \text{ for some } r>0,\, M_n(\sigma)=E[|x(\sigma)|^n]$$

which is a sufficient condition which guarantees moment-determinateness for $x(t)$. If we note that

$$|E[y(t)^n]|\leq\left[\int_s^t |k(t,\, \sigma)|^{n/n-1}d\sigma\right]^{n-1}\int_s^t M_{n+m}(\sigma)d\sigma$$

we can readily deduce a similar sufficient condition for $y(t)$ provided we assume that for $n>1$,

$$\int_s^t |k(t, \sigma)|^{n/n-1} d\sigma \leq M < \infty ,$$

which is entirely reasonable. A somewhat related condition is this: suppose the density of $x(t)$ is such that any $f(x)$ such that

$$\int_{-\infty}^{\infty} |f(x)|^2 p(x) dx < \infty .$$

Where $p(x)$ is the density corresponding to $x(t)$, we can find a sequence of polynomials $p_n(x)$ such that

$$E[|f(x) - p_n(x)|^2] \rightarrow 0.$$

The question is whether $y(t)$ will also satisfy a similar condition under reasonable restrictions on the processor, and as far as the author is aware, no general answers are available. Again, for all physical systems the answer should be affirmative.

The forms in (2.2) can be somewhat simplified [Balakrishnan, 1963] for "physically realizable" processes. We can use instead a Volterra expansion

$$y(t) = \sum_1^n \int_s^t \ldots \int_s^t W_k(t, \sigma_1, \ldots \sigma_k) x(\sigma_1) x(\sigma_2) \ldots x(\sigma_k) d\sigma_1 \ldots d\sigma_k .$$

One obvious advantage in using these forms is that the moments of $y(t)$ can be expressed in terms of the moments of the $x(t)$ process without requiring the full joint densities. The second degree form

$$y(t) = \int_s^t W_2(\sigma_1, \sigma_2) x(\sigma_1) x(\sigma_2) d\sigma_1 d\sigma_2$$

occurs in recent work involving detection of noise in noise. For a Gaussian signal and positive definite $W_2(\ldots)$, approximations to the distribution have been given recently by Grenander, Pollak, and Slepian [1959].

In view of the difficulties involved in solving (2.3) and (2.4), it would appear that for specific systems Monte Carlo methods would be a feasible computing alternative. Such methods have indeed been reported [Thaler and Meltzer, 1961] for the linear-filter-nonlinear-device-linear filter system that is of interest in radar applications.

When the Markov process is of the diffusion type, so that Fokker-Planck equations are available, it is possible [Deutsch, 1963, and Fortet, 1958] to obtain partial differential equations in addition to the integral equations. But from the point of view of practical solutions for the general cases this merely trades one difficult problem for an equally difficult one.

Systems represented by differential equations. So far we have assumed the processes arising from optimal operations to have an integral or system-function representation. In some cases the representation may be in terms of dynamical equations. Of importance to radio and communication is the phase-lock-loop system.

The "input" in this case is the slow-varying phase of a narrow-banded signal which is accompanied by additive noise. We may represent the noisy signal by

$$A \sin (w_0 t + \phi_1(t)) + x(t) \cos w_0 t + y(t) \sin w_0 t$$

$x(t)$, $y(t)$ being gaussian noise processes. The purpose of the phase-lock-loop is to produce a "clean" signal of the form

$$B \sin (w_0 t + \phi_2(t))$$

where the phase $\phi_2(t)$ is to "follow" $\phi_1(t)$ and the feedback or closed-loop is designed to achieve this. Details of the system may be found in Viterbi [1963]. Here we note that the relationship of the output phase $\phi_2(t)$ to the "input" $\phi_1(t)$ is described by a differential equation of the form

$$[a \sin \phi(t)+n(t)]=bL(D)[\dot{\phi}_1(t)-\dot{\phi}(t)]$$

where $L(D)$ is a differential operator (usually rational in D), $n(t)$ can be taken to be white Gaussian, a and b are constants and

$$\phi(t)=\phi_1(t)-\phi_2(t).$$

The difficulty in the analysis is caused by the appearance of the nonlinear function $\sin \phi$ on the left, and in the earlier literature it has been customary to use an approximate linear analysis. However, more recently, the statistics of $\phi(t)$ have been examined using Fokker-Planck equations by Viterbi [1963] and Tikhonov [1959] among others. This approach is feasible when $\phi_1(t)$ is nonrandom, since in this case $\phi(t)$ is a stationary Markovian or one component of a stationary vector Markov process. For instance, if one considers the simplest case when

$$L(D)\equiv 1, \dot{\phi}_1(t)=\nu$$

we have a stationary Markov process and the Fokker-Planck equation—the backward equation—is easily derived using standard techniques as:

$$\frac{\partial w}{\partial t}=\frac{b^2}{2}\frac{\partial^2 w}{\partial \phi^2}-[a \sin \phi+\nu]\frac{\partial w}{\partial \phi}$$

and since this contains only one space variable considerable progress can be made [Viterbi, 1963] toward obtaining the transition densities. This simplicity is lost as one considers more general forms of the operator $L(D)$, since now the Fokker-Planck equations contain more than one space variable. Some of these cases have been considered by Tikhonov [1959] and Viterbi [1963]. The former deals with

$$L(D)=(D+k)$$

while the latter examines the more realistic case (in the Communication Engineering context):

$$L(D)=\frac{D+a}{D}.$$

The analysis is shifted from a nonlinear ordinary equation to (the Cauchy Problem) a linear partial differential equation in several space variables (two in the cases above). While there is considerable recent work in the mathematical literature on the Diffusion equations that arise, much still remains to be done in specializing and applying these results to the present problem. The linearizing or quasi-linearizing techniques which have been used [Develet, 1956] appear to provide reasonable answers but a measure of the accuracy of the approximation is lacking, and must await more exact analysis.

3. Routine Data Handling

Of the many transformations of signal in more or less routine or established modes of data handling the only ones that warrant examination here are Sampling and Quantization. Let the number of quanta or levels chosen be N so that

$$f[x(t)]=\phi_i \text{ for } a_i \leq x(t) < a_{i+1}.$$

The received or reconstituted signal $y(t)$ is such that

$$y(t)=m_i$$

corresponding to the received level ϕ_i, usually,

$$m_i = \int_{a_i}^{a_{i+1}} xp(x)dx \bigg| \int_{a_i}^{a_{i+1}} p(x)dx \cdot$$

where $p(x)$ is the density corresponding to $x(t)$. Usually the quantity of interest is the error (say mean square) rather than the statistics of $y(t)$. We may calculate this error, including errors due to channel noise. Thus,

$$E[(x(t)-y(t))^2] = \sum_i \sum_j \int_{a_j}^{a_{j+1}} (x-m_i)^2 p_{ij} p(x)dx$$

where p_{ij} is the conditional probability of receiving the ith word assuming the jth word has been transmitted. This can be further simplified to

$$E[(x(t)-y(t))^2] = E[x(t)^2] - \sum m_j^2 p_{ij} + \sum\sum (m_i - m_j)^2 p_{ij} p_j \tag{3.1}$$

where

$$p_j = \int_{a_j}^{a_{j+1}} p(x)dx.$$

The first two terms, which are independent of channel characteristics, together yield the "quantization error." It must be noted that the problem remains of the proper choice of the levels $\{a_i\}$, which make (3.1) a minimum for a given N. Ordinarily, this is a rather impractical problem since the signal statistics cannot be specified a priori within the precision desired. Indeed, in practice one assumes that the channel errors can be neglected and that the signal has a uniform distribution, in which case the problem becomes trivial.

Let us next examine the effects due to sampling. Let

$$x_n = x(nT).$$

The samples represent the signal $x(t)$ in the sense that if the signal is band-limited to

$$\left[-\frac{1}{2T}, \frac{1}{2T}\right]$$

then

$$y(t) = \sum_{-\infty}^{\infty} x_n \frac{\sin \pi(2wt-n)}{\pi(2wt-n)} = x(t)$$

so that the statistics are the same. In practice, of course, one must consider the effect due to having a finite number of samples and secondly the effect due to the fact that the signal may not be band-limited. The first of these has received considerable attention and reference is made to Thomas [1963] for details. The second is the so-called folding or a biasing effect, and the error due to this should be considered well known [Balakrishnan, 1957]. A different kind of error occurs due to imperfections of the sampler. One such error is due to timing jitter. The timing is usually derived from the zero crossing of a sine wave of fixed frequency but usually there is some phase noise present which causes the axis-crossing times to jitter. This problem has been studied by Balakrishnan [1962]. Thus, the samples are now given by

$$y_n = x(nT + \phi_n)$$

where $\{\phi_n\}$ is a random sequence, which to a first approximation can be taken to be stationary. Let O be some operation desired on the samples $\{x_n\}$. Then the first step is to determine the optimal operation O′ on jittered samples $\{y_n\}$, so as to minimize the error—say, mean square—

$$E[\mathrm{O}[x_n] - \mathrm{O}'(y_n)]^2$$

and calculate this minimal error. The details may be found in Balakrishnan [1962]. Here let us note that $\{y_n\}$ is stationary and the direct error

$$E((x_n-y_n)^2)=2\int_{-w}^{w}(1-c(f))p(f)df$$

where the signal is assumed band-limited with $w=1/2T$, and $p(f)$ is the spectral density of $x(t)$, $c(f)$ is the characteristic function of the random variable ϕ_n. For the case where $p(f)$ is a constant, the normalized square error, expressed as a fraction of the average signal power, is given approximately by

$$\frac{8\pi^2\sigma_J^2W^2}{3}$$

where

$$\sigma_J^2=E[\phi_n^2].$$

This is also of course the mean square error in the direct "fitted" $y(t)$ using

$$y(t)=\sum_{-\infty}^{\infty} y_n\frac{\sin\pi(2wt-n)}{\pi(2wt-n)}.$$

We note that the error is proportional to the bandwidth. If $x(t)$ is gaussian, then $y(t)$ is again gaussian, regardless of the statistics of $\{\phi_n\}$. On the other hand, as shown by Balakrishnan [1962] for certain jitter statistics it is possible that a discrete component will arise in the spectrum of $y(t)$ even though $x(t)$ did not have any discrete components. If the jitter is "white" so that

$$E[\phi_n\phi_m]=0, \qquad n\neq m,$$

the spectral density of $y(t)$ is given by

$$|c(f)|^2p(f)+c$$

where the constant c is given by

$$c=E[x(t)^2]-\int_{-w}^{w}|c(f)|^2p(f)df,$$

$c(f)$ being the characteristic function corresponding to ϕ_n.

4. Adaptive Processing

In recent years there has been a growth of interest in adaptive methods in communication systems because of the acceptance of special purpose computer and/or computer data processing as part of the system. The effect of adaptive methods of processing on signal statistics is of interest because the analysis involved in these problems exhibits certain features that are novel.

In this section we shall examine a particular adaptive system which, while perhaps not typical, serves to illustrate the ideas involved. The system we shall consider is a means of achieving signal rate or bandwidth compression when no a priori information concerning the signal statistics is available and the processing has included a learning feature. Without going into a precise definition of what an "adaptive" system is, let us say that an adaptive system is one which monitors its own performance, and when new conditions arise which degrade the performance, the system learns how these new conditions effect the performance and adapts or makes structural changes to restore the performance level. An adaptive system thus will have a self-monitoring feature and a learning and self-adjusting feature. To be more specific, let $x(t)$ represent the continuous or discrete parameter signal. It is customary to assume that $x(t)$ can then be regarded (at least for analytical purposes) as a stochastic process. If the time-parameter t is continuous, then it is often possible to assume that

$$x(t) = \int_{-B}^{B} e^{2\pi i f t} d\psi(f)$$

in a suitable sense (that is, depending on whether we adopt the stochastic or nonstochastic viewpoint) for sufficiently large B, the bandwidth. By the well-known "sampling principle" then one can represent the continuous wave-form using periodic samples taken at $t=n/2B$. In most communication systems using sampled data of this kind, the sampling rate is determined by the nominal, highest or cutoff frequency B expected in the data. However, in many kinds of data—such as in space-telemetry data—the actual cutoff frequency is usually much smaller for most of the time, so that there is no need to sample at the nominal rate of $2B$. In other words, it is possible to "compress" the data sampling rate or channel bandwidth. A method of achieving such compression in P.C.M. systems is to exploit the redundancy or predictability of the data. Thus, we employ a "predictor" at the transmitter which predicts the data at time $t+\Delta$, based on the past up to time t. The actual value observed is then compared with the predicted. If the difference exceeds (in absolute value) a preset threshold, then the actual sample is transmitted. If it is below the threshold, a prearranged code word using, say, one or two digits is transmitted instead of the m digits, thus reducing the number of digits transmitted per second. A comma free code may be used to sort out the two kinds of words unambiguously. We shall not go into instrumentation details such as the buffering and so forth needed, but concentrate on the adaptive theory involved. The adaptive or learning feature comes in the predictor mechanism. The predictor is not operative until the threshold is exceeded, exhibiting the self-monitoring feature. The predictor itself is based on a "learning" phase, and the prediction operator being adjusted accordingly. For details on the predictor itself, reference may be made to Balakrishnan [1961]. The significance of the adaptive prediction feature lies in the fact that no a priori statistics or other assumptions concerning the data are required, and, in particular, it is realized that there may be periods in the data where prediction (to the quality set) may not be possible. (No prediction of the stock market prices is offered.)

The basic prediction philosophy may be indicated briefly. We are given a "wave-form" of duration T—a function $x(t)$, $0 \leq t \leq T$, in other words, and we are required to "predict" the value at time $T+\Delta$. Any prediction operation is to be based on this data alone, no other additional a priori knowledge being available. This is, of course, an ancient problem and here we wish to treat it strictly in the telemetry data processing context, and note two major points of view in dealing with it. One which may be considered the "numerical analysis" point of view consists in assuming that certainly any physically realized waveform must be analytic and can thus be approximated by polynomials. The data may thus be "fitted" to a polynomial of high enough degree and we then simply use this polynomial to "predict" the future values. The other and more recent view is the statistical view in which we assume (perhaps with good reason) that the data is a finite sample of a stationary stochastic process whose average properties such as moments and/or distributions are known or calculable from the data. For a process of given description we can apply the well-developed mean square prediction theory. Perhaps the main advantage with this view is that it gives us a quantitative, albeit theoretical, notion of the error in prediction. The problem of measuring the average statistics from a finite sample can be quite delicate, however. In the polynomial fitting method, the interpretation of the prediction error—which is, after all, the crucial point—is more nebulous, bound up with what degree polynomial to use and what portion of the data is to be fitted. Moreover, as ordinarily used, the fitting operations on the data are linear.

We adopt a rationale for prediction which is free from a priori assumptions concerning the "model." In a general sense what is involved in both the above methods is first "model-making" consistent with the data and as a second step using the numbers derived therefrom to perform some optimal operations. If an understanding of the mechanism generating the model is desired, the first step is essential. If what we want is prediction, then we shall show that it is possible to proceed directly (and hence more optimally) to the best prediction without the intermediate step of model-making. It may be, of course, that several philosophies lead to the same operations on the data. Even here, the present method offers some practical advantages. Moreover, it is only natural to use the philosophy that requires the least prior assumptions.

We note first that any prediction is an operation or operator on the part of the data, and in our case the finite part is all that is available. The main point of departure in our view is that if we have a prediction operator, which based on all the available input data, functions optimally in the immediate past of the point where the prediction is required, this is all that we can ask for meaningfully as a solution to the prediction problem. Thus, the only basis on which we can a priori judge any prediction method is to "back off" slightly from the present and compare the actual available data with the predicted value using the given prediction operator. Let us see how this can be formulated analytically. Let the total available data be described as a function $x(t)$, $0 \leq t \leq T$, and let it be required to predict the value at $T+\Delta$, $\Delta > 0$, being a small fraction of T. Let us next consider the data in the interval $0 < t < T-\Delta$. If for any t_0 in this interval we should choose to make a "prediction" of the function value Δ ahead from the past values up to t_0 denoting the predicted value by

$$x^*(t_0+\Delta)$$

we can explicitly observe the error

$$x^*(t_0+\Delta)-x(t_0+\Delta).$$

Let us next note that the general prediction operator will be a "function" of a finite segment of the past. Let us denote the length at this segment by S. Then, of course,

$$x^*(t_0+\Delta)=\mathrm{O}(x(\sigma):t_0-S<\sigma<t_0), \tag{4.1}$$

O representing the prediction operator. Next we have to specify the error criterion. Here we choose the mean square error, first because it is simpler analytically and is almost universally used, making comparison with other methods possible. The kind of solution presented being definitely not "analytic," based rather on successive approximation, other measures of error can be used at the risk of greater complexity. As far as the rationale of the method is concerned, this is largely a matter of detail, rather than principle. We thus use this method to determine optimality:

$$\frac{1}{T-\Delta-L}\int_L^{T-\Delta}[x^*(t+\Delta)-x(t+\Delta)]^2dt=\epsilon^2 \tag{4.2}$$

where $L \geq S$, and we proceed on the basis that the operator O is best which minimizes (4.2). We can, of course, generalize (4.2) as:

$$\frac{1}{T-\Delta-L}\int_L^{T-\Delta}\rho(t)C(x^*(t+\Delta)-x(t+\Delta))dt \tag{4.3}$$

where $\rho(t)$ is a positive weight function and $C(\cdot)$ is, say, a symmetric positive cost function. Before considering the problem of determining the optimal operator O, let us note an important consistency principle. Suppose the data is regarded as one long sample of an ergodic process. Then (4.2) yields exactly the optimal operator in the statistical sense. However, we have not needed to make any such assumption concerning the data, nor have to compute average statistics first. The point is that while the data may not be enough for determining let us say the spectrum, it may be quite adequate for the prediction itself. Unlike the polynomial fitting, the operations on the data can be as nonlinear as necessary, and at the same time (4.2) normalized to

$$\epsilon^2 \Big/ \frac{1}{T-\Delta-L} \int_L^{T-\Delta} x(t+\Delta)^2 dt \tag{4.4}$$

yields a quantitative measure of the prediction error on which to judge how good the prediction will be.

To continue with the description of the system, we assume that the receiver performs a prediction operation similar to the transmitter. In other words, the receiver predicts the values of the nontransmitted data using the transmitted as well as predicted sections of the data. This would mean that the receiver sets the same level of possible complexity of the prediction operator as the transmitter does, and in particular, the transmitter itself has to base its prediction using, as necessary, predicted data points that did not exceed threshold error.

So far we have determined the optimal adaptive structure, but there still remains the question of evaluating the system. For instance, it is natural to ask how much compression it is possible to obtain in this way on the average, and what the overall error will be. Here we have to postulate some structure for the data source and then calculate the resulting compression using the adaptive system. We shall now briefly indicate what an analysis of this type involves. Not to unduly complicate the analysis, let us consider the prediction operation in the transmitter based only on the actual samples. Let us denote the data samples by $\{x_n\}$. We assume that the data can be taken as a stationary stochastic process. We may further assume that the process is Gaussian since the maximum prediction error (and hence the minimum compression) occurs in this case because the prediction operation includes only the linear. Let us consider the case where the adaptive predictor is also constrained to be linear. In this case the optimal filter-weights $\{\alpha_j\}$ are determined by minimizing

$$\frac{1}{N} \sum_{-N+1}^{0} \left(x_n - \sum_{1}^{m} \alpha_j x_{n-j} \right)^2 \tag{4.5}$$

where the past available data consists of N samples and the prediction is based on "m" samples. We are, of course, considering the "analog" method above. The corresponding (squared) error is then

$$\epsilon_0^2 [m; N] = \left(x_1 - \sum_{1}^{m} \alpha_j x_{1-j} \right)^2. \tag{4.6}$$

The transmission is based on (4.6) exceeding a threshold "t". Hence, what we want first is to determine the statistics of (4.6). We note first that the optimal $\{\alpha_k\}$ that minimizes (4.5) will satisfy

$$\sum_{-N+1}^{0} x_n x_{n-k} = \sum_j \sum_{n=-N+1}^{0} \alpha_j x_{n-j} x_{n-k}, \qquad k=1, \ldots m.$$

Let Y_n be the N-column vector

$$[x_{+n}, x_{+n-1}, \ldots x_{n-N+1}].$$

Then the minimum of expressıon (4.5) becomes

$$\frac{1}{N} \frac{D_{m+1}}{D_m}$$

where D_{m+1} is the determinant of the $m+1$ by $m+1$ matrix with entries

$$Y_i \cdot Y_j, i, j = 0, -1, \ldots -m$$

and D_m is the determinant of the m by m matrix with entries

$$Y_i \cdot Y_j, i, j = -1, -2, \ldots -m.$$

On the other hand, we are interested in the error (4.6), which is

$$\epsilon_0^2[m;N]=\left(\frac{D'_{m+1}}{D_m}\right)^2 \tag{4.7}$$

where D'_{m+1} has first row

$$x_1,\ x_0,\ x_{-1},\ .\ .\ .,\ x_{-m+1}$$

and is otherwise the same as D_{m+1}. Our first interest is in the statistics of (4.7). We need to know

$$Pr\cdot[\epsilon_0^2[m;N]\geq t]$$

which is the probability of exceedance of the threshold, and the attainable compression ratio is then readily deduced from this.

We shall not go into the details of these calculations. However, if we simplify matters and assume that N is large enough so that we can replace the "time" average in (4.5) by a phase average, the $\{\alpha_j\}$ of course become the optimal regression coefficients that minimize

$$E\left[\left(x_0-\sum_1^m x_{-j}\alpha_j\right)^2\right]$$

and the error $\epsilon_0^2[m;\infty]$ is now the residual

$$\epsilon_0^2[m_1;\infty]=\left[x_0-\sum_1^m x_{-j}\alpha_j\right]^2.$$

But

$$\left(x_0-\sum_1^m x_{-j}\alpha_j\right)$$

is Gaussian and the threshold probability we want can be calculated simply from

$$\sigma^2[m,\infty]=E[\epsilon_0^2[m;\infty]].$$

Hence, the first step is to calculate this. This is already a nonstandard problem, in that explicit expressions for this error are not known. Some asymptotic estimates are given. The complete analysis thus involves some labor, unless simplifying approximations can suffice. For instance, in computing the statistics of (4.7), we may be content to compute the first moments. We omit the details of these calculations since our purpose here is merely to illustrate the kind of analysis is involved.

We note that the adaptive theory winds up with an "optimum" *procedure.* To evaluate how good the system actually is, we have to specify the class of input or system parameters—or their statistics if they are regarded as randomly varying. Very often there may not be any clear-cut "optima." This means that we may have to be satisfied with suboptimal systems and there will usually be many of these, and the problem of deciding among them by analysis in any quantitative way can be a hard one.

In conclusion, let us note that adaptive processing methods in communication theory are still in their formative stage. We have discussed an example which illustrates most of the features that characterize the theory involved in these methods without any pretense at being exhaustive.

5. References

Balakrishnan (1957), A note on the sampling principle for continuous signals, IRE Trans. Inform. Theory **IT–3,** No. 2, 143–146.

Balakrishnan, A. V. (1961), An adaptive nonlinear data predictor, Proc. of the National Telemetry Symposium.

Balakrishnan, A. V. (April 1962), On the problem of time jitter in sampling, IRE Trans. Inform. Theory, **IT–8,** No. 3, 226–236.

Balakrishnan, A. V. (1963), A general theory of nonlinear estimation problems in control systems, Journal of Math Analysis and Applications, **8,** No. 1.

Deutsch, R. (1963), Nonlinear Transformations of Random Processes (Prentice-Hall, Englewood, Cliffs, N.Y.).

Develet, J. A., Jr. (Feb. 1956), A threshold criterion for phase-lock modulation, Proc. IRE **44,** No. 2.

Fortet, R. (1958), Recent advances in probability theory, Some Aspects of Analysis and Probability (John Wiley Sons, Inc., New York, N.Y.).

Grenander, V., H. Pollak, and D. Slepian (Dec. 1959), Distribution of quadratic forms in normal variates SIAM J. **7,** No. 4.

Thomas, J. B. (1963), A survey of sampling theorem expansions, Conference on Identification Problems in Communication and Control Systems, Princeton University.

Tikhonov, V. I. (1959), The effects of noise on phase-lock oscillation operation, Automatika i Telemakhanika **22,** No. 9.

Viterbi, A. J. (Dec. 1963), Phase-locked loop dynamics in the presence of noise by Fokker-Planck techniques, Proc. of the IEEE.

6. Additional Related Reference

Rosenblatt, M. (Nov. 1957), Some purely deterministic processes, J. Math. Mech. **6,** No. 6, 801–810.

Paper 68D9–395