Linear Least-Squares Smoothing and Prediction, with Applications

By SIDNEY DARLINGTON

(Manuscript received April 29, 1958)

This paper describes the calculation of smoothing and prediction operators of the linear least-squares sort using techniques derived from a circuit theory point of view. The techniques are developed explicitly for time series which are continuous and statistically stationary. Other situations are explored more briefly, however, in which the time series are either discrete or statistically nonstationary.

For the most part, functions of time are replaced by functions of frequency, representing their transforms. Mathematical complications are avoided by restricting statistical ensembles to those which have rational power spectra. In practice, actual spectra can be approximated sufficiently well by rational spectra, and the simplified methods are sufficiently general for engineering applications of many different sorts. Both finite and semi-infinite smoothing intervals are permitted, with or without constraints of various sorts. The assumption of rational spectra does not apply directly to nonstationary time series, but it may be replaced by a closely analogous restriction which does apply. Then there are nonstationary operations which are closely analogous to the stationary operations, developed for stationary systems. A brief examination of these analogies is of interest, even though the nonstationary operations are usually too complicated for engineering purposes.

The general techniques are developed in terms of specific problems, chosen for purposes of exposition and because of their engineering interest.

I. INTRODUCTION

For a number of years, beginning roughly at the end of World War II, there has been a growing interest in theories of optimum smoothing and prediction. Much of the work has been concerned with optimum smoothing and prediction of the linear least-squares sort applied to statistically stationary time series — a subject which is both attractive to mathematicians and important in various engineering problems.

This paper describes techniques for solving smoothing and prediction problems of the stationary, linear, least-squares sort using a circuit theory point of view. It avoids the more difficult mathematics of the very general, completely rigorous treatments, but maintains sufficient generality for many engineering applications. It develops general techniques in terms of specific engineering problems, which are of real interest in themselves and may also serve as patterns to be followed in solving other problems. Among the problems considered explicitly are the following: classical smoothing and prediction problems solved by Wiener,[1] Kolmogoroff,[2] Zadeh and Ragazzini,[3] etc.; the simultaneous use of different instruments, with different error spectra, for the observation of single physical variables; applications of the *mathematics* of data smoothing to circuit design problems which do not actually involve data smoothing as such.

The general techniques described here have been developed over a period of years. Some of the results have already been stated, in specialized reports describing specific applications of one sort or another.[4, 5, 6]

1.1 *Further Background*

Present-day theories of smoothing and prediction may be said to have started with the classic papers of Wiener[1] and Kolmogoroff,[2] which were written during World War II. They assumed linear least-squares

operations, stationary statistics and observations available for all past times. Zadeh and Ragazzini[3] modified the theory for observations which are available only over a past interval of finite duration. By now, many other papers have been published, aimed at generalizing, modifying, interpreting or applying the original theories. A complete bibliography would be very extensive, and will not be attempted in this paper.*

Differences in points of view can result in quite different formulations of smoothing and prediction theory, even though the formulations must reflect the same mathematical fundamentals. This is important, because the classic papers of Wiener, Kolmogoroff, Zadeh and Ragazzini, etc. contain quite formidable mathematics, which is not generally accessible to engineers. Much of this mathematics can be avoided by imposing certain additional restrictions, which are generally minor in terms of resulting restrictions on engineering applications. Further complications can be avoided by not requiring perfect rigor in regard to all singular, or mathematically "pathological" situations. This point of view is quite different from that adopted, for example, by Doob[8] in his very general treatment of smoothing and prediction in terms of the general theory of stochastic processes.

Bode and Shannon[9] simplified the derivation of Wiener and Kolmogoroff's most important result, using circuit theory concepts to interpret the mathematical operations in physical terms. Their physical interpretations are very powerful tools for engineers who must solve the mathematical problems and, in fact, their paper is our principal reference in what follows. Their method of solving the Wiener-Kolmogoroff problem, however, does not apply to the Zadeh-Ragazzini problem or to various other generalizations of engineering interest (unless it is complicated in ways which destroy most of its advantages). Furthermore, their solution of the Wiener-Kolmogoroff problem itself is not simple in numerical applications.

This paper uses a circuit theory point of view in a somewhat different way, which leads to more general applications and to simpler computations. The advantages are obtained at the price of an additional mathematical restriction. Functions of frequency representing statistical "power spectra" are required to be *rational*, whereas more general theories allow more general functional forms.† This is a minor restriction in most engineering applications, where nonrational functions can be replaced by rational approximations.

* For an extensive bibliography (as of 1955) see Stumper.[7]

† Bode and Shannon mention rational spectra as simplifying one part of their method — the evaluation of a loss-phase integral — but they do not seek other simplifications, which may be realized by a rather different method of solution.

Under the assumption of rationality, most of the analysis can be carried out in the frequency domain, in terms of the more usual operations of circuit theory. The concepts of generalized Borel fields, measurable spaces and even Hilbert spaces need not be used at all. Usually, Wiener-Hopf equations can be replaced by contour integrals in the complex plane. When Wiener-Hopf equations do appear, they may be replaced quickly by conditions applied to the analytic properties of functions of frequency. This avoids the usual difficulties with "δ functions" and their derivatives, and states conditions in forms more familiar to circuit theorists. Frequently, end results may be expressed as conditions which determine network zeros and poles more or less directly. These circumstances all depend, however, on the basic assumptions of linear least-squares smoothing. For the simple methods, the assumption of stationary statistics also is essential; more complicated analogous methods apply to time-variable situations, at least in principle. Nonstationary systems are discussed in this paper only briefly, in Section 3.6.

For the most part, continuous-data systems are assumed. However, the techniques developed for continuous data can readily be adapted to sampled-data problems, by methods which are outlined in Section 3.5. These methods have not yet been compared in detail with more direct methods of handling sampled-data problems such as, for example, those of Levinson (Ref. 1, Appendix B), and Lloyd and McMillan.[10]

Chang[11] has described a frequency-domain equivalent of Wiener and Kolmogoroff's central result. He starts with contour integration, but does not simplify the solution by assuming rational spectra. He does not extend the method to the finite memory problem solved by Zadeh and Ragazzini. Zadeh and Ragazzini themselves describe a solution which assumes rational spectra, but they use a time-domain analysis which is less simple than an analysis in the frequency domain. Laning and Batten[12] also describe smoothing and prediction in time-domain terms, subject to the assumption of rational spectra.

In Ref. 11 Chang also points out that the *mathematics* of smoothing and prediction may be applied to network synthesis problems which do not actually involve data smoothing or prediction. Basically, Chang proposes designing to a least-squares error criterion, where the error is the magnitude of the complex difference between a physical transfer function and a nonphysical ideal function which it is to approximate. This appears to be a quite rewarding approach to various nonstochastic problems in network synthesis, particularly after the frequency-domain method has been extended to more general smoothing and prediction problems.

Within the general field of data smoothing, an important variation of the classical problem is as follows: In the classical problem, one is concerned with the optimum smoothing and prediction of a statistical signal, contaminated with a statistical noise, when the statistics of the signal and noise are known. In the variation, one is concerned with the simultaneous use of *two different* instruments to measure a single physical variable or signal. In the simplest form, the readings of the two instruments are combined through optimum linear operations, subject to the condition that the net error is to depend only on the instrumental error. Then the signal statistics do not enter at all, but two different statistics must still be considered, corresponding to the two different instrumental errors. If the errors of the two instruments have quite different frequency characteristics, the two-instrument combination can give much greater accuracy than either instrument alone. This may be compared with the use of "woofer" and "tweeter" speakers in high fidelity sound systems.

The techniques described here were developed, to a considerable extent, in connection with specific applications of the two-instrument problem described in Refs. 4, 5 and 6. The two-instrument optimization problem was suggested by previous uses of two kinds of instruments, combined through arbitrary, non-optimum linear operations.* More recently, two-instrument optimization principles have been described in papers by Bendat,[14] and Stewart and Parks.[15]

1.2 *Organization of the Paper*

The remainder of this paper is organized as follows: Section II formulates a fairly general smoothing and prediction problem in mathematical and physical terms. At the same time it reviews certain mathematical relations which will be needed in the sequel, including some elementary Fourier transforms, some properties of "physical" networks and some properties of stationary Gaussian noise.

Section III develops techniques for solving the general problem. The techniques are explained in terms of specific problems, which are special cases of the general problem (or reasonable variations of it) and are especially suitable for purposes of explanation. Section IV describes other specific problems and engineering applications which have been chosen primarily for their practical interest.

Some of the specific problems illustrate existing engineering applications. Others are merely potentially useful or of interest for largely the-

* Examples are: an instrument made by North American Aviation for the Sandia Corporation (which furnished a starting point for Ref. 4) and a proposal of Crooks.[13]

oretical reasons. Some of the problems may not have been solved before. Others have well-known solutions, in one form or another, and are included purely to illustrate the generality and efficiency of the techniques under discussion.

A more detailed outline of the paper may be seen in the table of contents at the beginning of this paper.

II. FORMULATION OF GENERAL RELATIONS

In this section we formulate a central problem, in about the same way as Bode and Shannon, and review some mathematics which will be needed in the sequel. In later sections, we shall modify some of the details, but within a set of fundamental restrictions which are included in the formulation described below.

2.1 *The Central Problem*

The central problem is as follows: We are given a time function $f(t)$, representing a signal $s(t)$ contaminated by noise $n(t)$:

$$f(t) = s(t) + n(t). \tag{1}$$

The time functions $s(t)$ and $n(t)$ are drawn from statistical ensembles of such functions, and we assume that the pertinent statistical characteristics of the ensembles are known.

We are to derive from $f(t)$ an estimate $g(t)$ of $s(t + \alpha)$. When α is positive, $g(t)$ is a prediction of what the true signal s will be α seconds from present time t. When α is negative, $\alpha = -\beta$ and $g(t)$ is an estimate of what the true signal was β seconds before present time t. The operations to be used in deriving $g(t)$ from $f(t)$ are restricted in various ways. Then $g(t)$ generally will not match $s(t + \alpha)$ exactly, but will be in error, by an amount $\epsilon(t)$:

$$g(t) = s(t + \alpha) + \epsilon(t). \tag{2}$$

The permitted operations are to be used in such a way that $g(t)$ is an *optimum* estimate of $s(t + \alpha)$, as judged by a specific criterion applied to the statistics of the error $\epsilon(t)$. The problem is to find the specific combination of operations within the permitted operations which will, in fact, yield the optimum $g(t)$.

An engineering representation of the problem is illustrated in Fig. 1. The "observed" signal $f(t)$ differs from the "true" signal $s(t)$ by the noise $n(t)$. The observed signal is to be modified by passing it through some sort of device, such as an electrical network, to obtain the output signal $g(t)$, which is to represent an estimate of $s(t + \alpha)$. The action of the de-

Fig. 1 — A physical representation of the smoothing and prediction problem.

vice is described by a mathematical "operator", which is simply a symbolic representation of the corresponding mathematical process relating $g(t)$ to $f(t)$. If the device must be chosen from some class of permitted devices, this class of permitted devices will determine a class of permitted operators. The problem is to determine the optimum operator within the permitted class, as a first step in designing an optimum device.

2.2 *General Assumptions*

The specific method of analysis depends on specific assumptions regarding the statistics of the signal and noise, the criterion used to define optimum estimates and the class of permitted operators. The six conditions stated below will be assumed throughout this paper, except in a few instances where certain specific departures will be noted. Other conditions will vary with different circumstances, considered in different sections, and will be noted when needed.

i.* The signal and noise statistics are assumed to be *stationary*. Thus, statistical characteristics which refer to a single time are the same for all times, correlations involving more than one time depend only on time differences, etc.

ii. The criterion used to define an optimum estimate is to be the *average square error*, or variance $\sigma^2 = \text{ave } \epsilon^2$. In other words, the permitted operations are to be used in such a way that σ^2 is a minimum.

iii. The permitted operations are to be *linear*. A linear operation, applied to $f(t)$, may yield a sum of terms of the following sorts: values of $f(t)$ at specific times, derivatives of $f(t)$ of any order, weighted integrals of $f(t)$.

iv.† The "power spectra" of the signal and noise are to be *rational* functions of frequency ω — real when ω is real. (Power spectra are Fourier transforms of covariance functions.)

v. The statistics and the permitted operations must be such that estimates with *bounded average square errors* are possible.

vi. In general, the permitted operations will be an *assigned subset* of the class of all linear operations (for example, the class of all "physical"

* Except in Section 3.6, which shows how the methods used for stationary statistics may be applied to time-variable systems of certain very special kinds.

† Except for possible departures in Sections 3.2.3 and 4.3.

linear operations). Sums and differences of permitted operations will always be linear operations, but they will not always be *permitted linear* operations. However, we will assume that the class C_Y, of permitted operations Y, always has the following property: If Y_1 and Y_2 are permitted operations and k is an arbitrary positive or negative real constant, there must always exist a permitted operation Y_3 such that*

$$Y_3 - Y_1 = k(Y_2 - Y_1). \tag{3}$$

Conditions i, ii and iii are fundamental to the theories of Wiener,[1] Kolmogoroff,[2] Zadeh and Ragazzini,[3] and Bode and Shannon.[9] It is these conditions which make the mathematics tractable. Conditions i and iii are clearly appropriate for a treatment using conventional theories of fixed linear circuits. Under condition ii, the optimization depends only on linear correlations, as will be confirmed in Section 2.6. Then the actual statistics may be replaced by any more convenient statistical models which have the same linear correlations. Bode and Shannon discuss ways in which these three assumptions do and do not limit engineering applications. The limits should be clearly understood before practical applications are attempted.

Under conditions i, ii and iii, with no further restrictions, mathematically "pathological" situations must be accounted for, and these lead to quite formidable (although tractable) mathematics. Condition iv, assumed here, excludes the more pathological situations. The resulting simplifications in the necessary mathematics are very substantial. While the requirement of rational spectra is an arbitrary restriction, it does not restrict engineering applications to a serious extent. The nonrational spectra usually encountered can be approximated sufficiently well with rational functions.

Condition v is not a significant restriction on the usefulness of a design method. The convergence of certain integrals in which we will be interested depends on this condition, and it is stated here for ready reference. Note that v does *not* require convergence of the integrals of the signal and noise spectra alone, provided the permitted operations can lead to estimates with bounded average square errors. (See, for example, Section 4.1.1.)

Condition vi guarantees that the optimum permitted operator will correspond to a "stationary point" in the usual calculus of variations

* L. A. MacColl has shown that condition vi is equivalent to the following: Let Y_0 be any one permitted operation, and let V be the class of operations $Y - Y_0$. Then V is a "linear subspace" and C_Y a "flat subspace" in the linear vector space of all linear operations.

sense. It can frequently be simplified to the following: If Y_1 is a permitted operator, then kY_1 is a permitted operator. We are going to examine certain constraints, however, which exclude the simple form. The typical constraint of this sort permits only linear operations Y which make no change in in some particular (specified) time function (for example, a constant, or dc, signal). If $f_0(t)$ is the particular time function, and $Y \cdot f_0(t)$ is the result of applying operator Y to $f_0(t)$, the constraint requires $Y \cdot f_0(t) = f_0(t)$. But then $(kY) \cdot f_0(t)$ becomes $kf_0(t)$, and is not permitted. On the other hand, $(Y_2 - Y_1) \cdot f_0(t)$ becomes $[f_0(t) - f_0(t)] = 0$, and the same is true of $k(Y_2 - Y_1)$.

2.3 *Substitution of Gaussian Ensembles*

We now replace the actual signal and noise ensembles by Gaussian ensembles with the same linear correlations (as permitted under assumption ii). For a more specific physical representation, we may think of the new $f(t)$ and $g(t)$ as electrical signals (voltages or currents), provided the pertinent statistical characteristics are retained. Stationary Gaussian ensembles may be generated by passing white noise through (idealized) linear networks. Under assumption iii, the operations used to derive $g(t)$ from $f(t)$ also correspond to some *linear* network. Then, Fig. 1 may be replaced by Fig. 2. The two white noise ensembles are uncorrelated (assuming that signal and noise are uncorrelated). Their spectral densities are unity (scale factors appear as gains or losses associated with the networks, which are permitted to include amplifiers). The linear operations performed by the networks may be represented by frequency functions $Y_s(p)$, $Y_N(p)$, $Y_G(p)$, where $p = i\omega$. Responses to any time functions may be found from the frequency functions by means of Fourier transforms. We shall say more about these later on, and also about the properties of white noise.

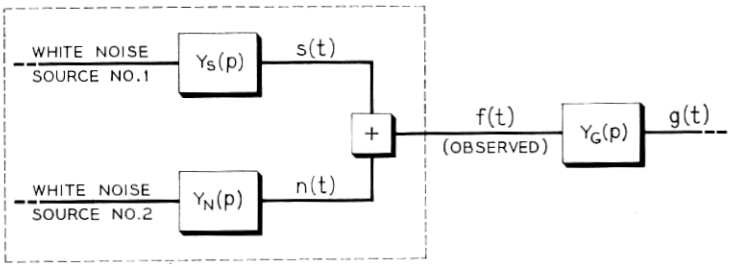


Fig. 2 — A Gaussian physical model.

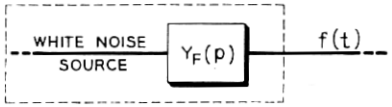


Fig. 3 — An alternative physical model for $f(t)$.

The noise sources themselves are not available to the observer, who sees only $f(t)$. The noise sources and associated networks corresponding to $Y_S(p)$ and $Y_N(p)$ are merely imagined devices which permit the pertinent signal and noise statistics to be described in physical terms. Our problem is to find that particular permitted $Y_G(p)$ (regarded as a linear operator) which converts $f(t)$ into the optimum $g(t)$.

In Fig. 2, $f(t)$ is the sum of two Gaussian ensembles. Since the sum of two Gaussian ensembles is a Gaussian ensemble, $f(t)$ may be represented more simply, as in Fig. 3. This representation does not show the correlation between $f(t)$ and $s(t)$ or $n(t)$, but it will be useful in some of our analysis. The power spectrum of $f(t)$, viewed as a single ensemble, is the sum of the power spectra of the (uncorrelated) signal and noise.

The auto-covariance of any one of the ensembles (signal, noise or signal plus noise) may be specified in any one of three ways: directly as a function of time difference [average of $s(t)\ s(t+\tau)$, etc.]; by means of the power spectrum (which is the Fourier transform of the auto-covariance function); or by means of a network function, Y_N, Y_S or Y_F of Figs. 2 or 3 (from which the power spectrum can easily be computed). We will use the following notation:

Ensemble	Auto-covariance Function	Power Spectrum	Network Function
$s(t)$	$\Phi_S(\tau)$	S	$Y_S(p)$
$n(t)$	$\Phi_N(\tau)$	N	$Y_N(p)$
$f(t) = s(t) + n(t)$	$\Phi_F(\tau)$	$F = S + N$	$Y_F(p)$

Let E and Y_E be any of the three spectra and its corresponding network function. Let $\tilde{Y}(p)$ designate $Y(-p)$:

$$\tilde{Y}(p) = Y(-p). \tag{4}$$

Then an elementary property of power spectra requires:*

$$E = Y_E(p)\,\tilde{Y}_E(p). \tag{5}$$

We need to know how to find Y_E when E is given. In general, the relation of Y_E to E involves the general loss-phase integrals, as described by

* Related to equations (T-9) and (T-14) of Table I, Section 2.5, and the properties of white noise.

Bode.[16] Under our assumption of rationality, however, the loss-phase integrals can be replaced by simple relations between zeros and poles (also in accordance with Bode[16]).

Equation (5) makes E an even function of p, and also an even function of ω (since $p^2 = -\omega^2$). At real *frequencies*, $E = |Y_E(i\omega)|^2$ and is non-negative. The zeros of E occur in positive and negative pairs, like $+p_\sigma$ and $-p_\sigma$, and so do the poles. Of each pair, one is a zero or pole of Y_E and the other a zero or pole of $\tilde{Y}_E$. When E is given, the zeros and poles can be arranged in + and − pairs, and one of each pair can be assigned to Y_E. The possible assignments are not unique, however, unless some further restriction is imposed. For our purposes, we will need the specific assignment described below.

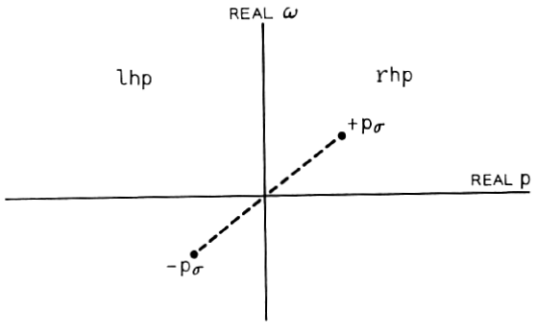


Fig. 4 — The complex plane for $p = i\omega$.

Referring to the complex plane for p oriented as in Fig. 4, if p_σ lies on one side of the real ω axis, $-p_\sigma$ lies on the other side. Using "lhp" to designate "left-half plane", we require:

$$\textit{The zeros and poles of } Y_E \textit{ are the lhp zeros and poles of } E, \quad \textit{where } E = S, N, F. \tag{6}$$

There may also be zeros of E *on* the real ω axis, but these always occur in *identical* pairs, one of which goes to Y_E.*

2.4 *Properties of Physical Frequency Functions*

Let "rrhp" designate "regular in (the finite part of) the right-half plane." Then an immediate consequence of (6) is

$$Y_E \text{ and } 1/Y_E \text{ are rrhp where } E = S, N, F. \tag{7}$$

* As an immediate consequence of the nonnegative character of $E = |Y(i\omega)|^2$ at real frequencies.

Then both Y_E and $1/Y_E$ are "physically realizable" in the general sense of Bode.[16]* Under assumption iv they are realizable with finite networks of lumped elements, provided these are idealized to include multiple poles at $p = \infty$ (which can be approximated within reasonable limits by active circuits).

The function $Y_G(p)$ in Fig. 2, which converts the observed $f(t)$ into the estimate $g(t)$, may not be rational even though Y_S and Y_N are rational. When a nonrational Y_G is required to be "physical", it must still be regular in the finite part of the rhp. Now, however, it may have an essential singularity at $p = \infty$, and this must meet an additional restriction.

A general definition of a "physical" frequency function, $Y(p)$, merely requires that it be "causal". A causal $Y(p)$, applied as an operator to any time function $f(t)$, produces a response $g(t)$ which depends only on present and past values of $f(t)$. Causal frequency functions have been studied in very general terms, by Wiener,[1] Beurling,[17] Nyman[18] and Youla, Castriota, and Carlin,[19] but the general mathematics is relatively complicated. A less comprehensive description of the conditions for physicalness will be sufficient for our purposes. Let the definition of rrhp be extended, arbitrarily, to

$$\begin{aligned} &\text{rrhp means:} \\ &a.\ \text{Regular in the finite part of the rhp,} \\ &b.\ \text{Approaches } \textstyle\sum_{m,\gamma} (C_{m,\gamma} p^m e^{-\gamma p}) \text{ as } p \to \infty, \\ &\qquad m = \pm\text{integer}, \\ &\qquad \gamma = \text{real and } \geqq 0, \\ &\qquad C_{m,\gamma} = \text{real}. \end{aligned} \tag{8}$$

Then, for purposes of this paper, it is sufficient to define "physical" by:

A physical $Y(p)$ is any real function of p which is rrhp. (9)

(A real function of p is one which is real when p is real.)

In (8), the second condition permits $Y(p)$ to behave like $p^m e^{-\gamma t}$ at $p \to \infty$, provided γ is *positive*. Behaviors of this kind can be obtained with networks of lumped elements and ideal delay lines.

We will need corresponding relations describing $\tilde{Y}(p) = Y(-p)$. Let rlhp be defined by

* As used here "physically realizable" includes the requirement of stability. This is the usual interpretation of the conventional theory of linear networks. An unstable linear device can in fact exist and it can be driven by an input which is a general function of time, but only over a finite time interval.

rlhp means:

$$\begin{aligned} &a.\ \text{Regular in the finite part of the lhp,} \\ &b.\ \text{Approaches } \textstyle\sum_{m,\gamma} (C_{m,\gamma} p^m e^{+\gamma p}) \text{ as } p \to \infty, \\ &\qquad m = \pm\text{integer}, \\ &\qquad \gamma = \text{real and} \geqq 0, \\ &\qquad C_{m,\gamma} = \text{real}. \end{aligned} \tag{10}$$

Then, it follows from (8) and (9) that:

A physical $\tilde{Y}(p)$ is any real function of p which is rlhp. (11)

2.4.1 *An Essential Integral Theorem*

Bode[16] has derived a number of special properties of "physical" functions, in terms of integrals in the complex p plane. These include the loss-phase relations, the integral in the definition of resistance efficiency and other similar integrals. One particularly simple theorem of this sort is essential to our method of solving smoothing and prediction problems.

For our purposes, the theorem may be stated as follows:

$$\begin{aligned} &\text{If: } a.\ H(p) \text{ is either rrhp or rlph and} \\ &\qquad b.\ |H(i\omega)| = O\omega^{-2} \text{ when } \omega \text{ is real and} \to \infty^*, \\ &\text{then: } \int_{-\infty}^{+\infty} H(i\omega)\, d\omega = 0. \end{aligned} \tag{12}$$

The function $H(p)$ is not necessarily one of our network functions, $Y_E(p)$ or $Y_G(p)$, provided it is rrhp or rlhp in the sense of (8) or (10) and also meets the convergence condition. Generally, it will be a combination of our network and spectral functions.

The theorem is easily proved by closing the path of integration with an arc at ∞. The arc encloses either the rhp or the lhp, as in Fig. 5(a) or 5(b), depending upon whether $H(p)$ is rrhp or rlhp. Because of (12a) the integration around the closed contour is 0. Because of (12b), the integral over the arc at ∞ is 0 [provided $\gamma \geqq 0$ in (8) or (10), as required].

2.5 *Fourier Transforms*

Equations (1) and (2) describe our problem in terms of time functions, while Fig. 2 describes it in terms of frequency functions (and the proper-

* The symbol O has the following meaning: if $q(\omega) = Or(\omega)$, then $q(\omega)/r(\omega)$ is bounded; thus, $H = O\omega^{-2}$ means $\omega^2 H$ is bounded.

ties of white noise). In general, a facility at transforming quickly between time-domain and frequency-domain formulations is an important tool in smoothing and prediction problems. Time-domain and frequency-domain formulations are, of course, mathematically equivalent, and are related by Fourier transformations. A few elementary theorems regarding Fourier transforms which will be referred to in later sections are reviewed in Table I.

In order for the Fourier transforms to exist, certain mild conditions must be met. We will assume, a priori, that the conditions are satisfied wherever we use the transforms. This is a departure from strict rigor, but our use of the transforms will be entirely reasonable, under our assumptions iv and v.

When $Y(p)$ is the frequency function of a network or equivalent device, $Y(i\omega)$ indicates the steady state response to a sinusoidal input. The inverse transform, $K(t)$, is the response to an ideal unit impulse, or δ function, applied at time $t = 0$. A general input time function, $f(t)$, may be thought of as a series of impulses. The effect of any one impulse on the response $g(t)$ at a given time t depends on the amplitude of the impulse and on the length of time which has elapsed since the impulse was applied. Then

$$g(t) = \int_{-\infty}^{+\infty} f(t - \tau)K(\tau)\, d\tau = f(t) * K(t). \tag{13}$$

Thus, $g(t)$ is a weighted integral of $f(t)$, in which the weight factor is $K(\tau)$ and τ represents the "age of data".

The response of a physical device cannot depend on future inputs.

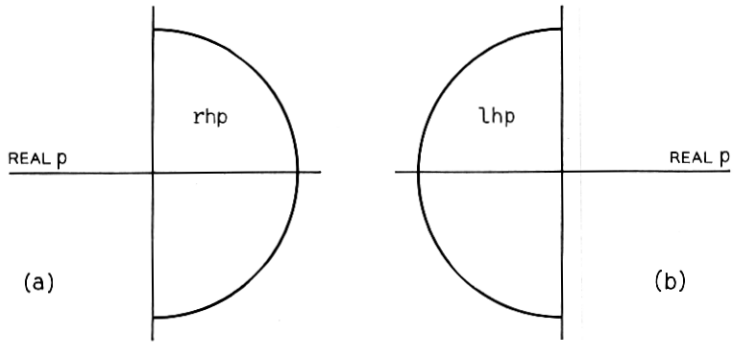


Fig. 5 — Arcs at infinity, enclosing half-planes.

TABLE I — FOURIER TRANSFORMS

Definitions

Fourier Transform:

$$Y(i\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} K(t)\, e^{-i\omega t}\, dt \tag{T-1}$$

Inverse Transform:

$$K(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} Y(i\omega)\, e^{+i\omega t}\, d\omega \tag{T-2}$$

Convolution:

$$K_1(t) * K_2(t) = \int_{-\infty}^{+\infty} K_1(t-\tau)\, K_2(\tau)\, d\tau \tag{T-3}$$

Delta Function:

$$\delta(t) = 0 \quad \text{when} \quad t \neq 0, \int_{-\epsilon}^{+\epsilon} \delta(t)\, dt = 1 \tag{T-4}$$

Some Fourier Transform Pairs

$K(-t)$	$Y(-p) = \tilde{Y}(p)$	(T-5)
$kK(t)$	$kY(p)$	(T-6)
$K_1(t) + K_2(t)$	$Y_1(p) + Y_2(p)$	(T-7)
$\delta(t)$	$1/\sqrt{2\pi}$	(T-8)
$K(t-T)$	$Y(p)e^{-Tp}$	(T-9)
$K_1(t) * K_2(t)$	$Y_1(p)Y_2(p)$	(T-10)
$K(t) * K(-t)$	$Y(p)\tilde{Y}(p)$	(T-11)
$K(t) = 0, \quad t < 0$	$Y(p)$ rrhp*	(T-12)
$K(t) = 0, \quad t > 0$	$Y(p)$ rlhp*	(T-13)

Some Related Formulas

Parseval's Equation:

$$\int_{-\infty}^{+\infty} K_1(t)\, K_2(t)\, dt = \int_{-\infty}^{+\infty} Y_1(i\omega)\, \tilde{Y}_2(i\omega)\, d\omega \tag{T-14}$$

If $Y(p)$ = real when p is real:

$$\int_{-\infty}^{+\infty} [K(t)]^2\, d\omega = \int_{-\infty}^{+\infty} |\, Y(i\omega)\, |^2\, d\omega \tag{T-15}$$

Power Spectrum = Fourier Transform of Covariance. (T-16)

If $g(t)$ is the response of a linear device to an input $f(t)$ and if $g(t) = Y(p)e^{pt}$ when $f(t) = e^{pt}$ and $g(t) = K(t)$ when $f(t) = \delta(t)$, then $Y(p)$ and $K(t)$ are a Fourier transform pair and, for a general $f(t)$,

$$g(t) = f(t) * K(t) = K(t) * f(t) \tag{T-17}$$

* Provided certain pathological $Y(p)$ and $K(t)$ are excluded, in a way consistent with our definitions of "rrhp," "rlhp," and "physical."

Thus, (13) requires:

$$\textit{If } K(t) \textit{ is physical, } K(t) = 0 \textit{ when } t < 0. \tag{14}$$

When $t \geqq 0$, $K(t)$ need not be particularly well behaved, for it can include δ functions and their derivatives. If it contains nothing worse than derivatives of δ functions it can be approximated with combinations of transversal filters and differentiators.

2.6 *A Variational Condition, Equivalent to the Optimization Requirement.*

In Figs. 2 and 3 our inputs are drawn from unit-level white noise ensembles. White noise may be described in either frequency-domain or time-domain terms, in accordance with Rice.[20] In the frequency domain it is a sum of sinusoids of all frequencies, with phases that are completely random. The "spectral density" is constant over all frequencies. If the white noise is applied to a network described by $Y(p)$ the corresponding output has similar properties, except that the amplitudes of the sinusoids of different frequencies, ω, are changed by a factor $| Y(i\omega) |$. Then the power spectrum of the ensemble has density E given by the following (at real frequencies):

$$E = | Y(i\omega) |^2 = Y(p)\tilde{Y}(p). \tag{15}$$

Since phases are initially entirely random, phases added by the network do not change the character of the ensemble. Changing the phase of $Y(i\omega)$ without changing $| Y(i\omega) |$ merely maps individual time functions into other, equally probable time functions of the same ensemble.

The average square, σ^2, of the response to the white noise, is simply the sum of the average squares for the individual frequencies. Thus,

$$\sigma^2 = \int_{-\infty}^{+\infty} | Y(i\omega) |^2 \, d\omega = 2 \int_{0}^{+\infty} | Y(i\omega) |^2 \, d\omega. \tag{16}$$

Here, σ^2 represents both the "time average" and the "ensemble average", since the two are identical when they refer to a stationary Gaussian ensemble. The two ranges of integration are equally permissible, as shown, because $| Y(i\omega) |^2$ is an even function of ω.*

In the time domain, the white noise may be described as a sequence of impulses, with infinitesimal spacing along the time scale. The amplitudes of the impulses are uncorrelated Gaussian random variables. An

* It is assumed here that a spectrum E is so scaled that integrating E from $\omega = -\infty$ to $+\infty$ gives the variance, σ^2. "Unit level" white noise is to be consistent with this assumption and (15). Sometimes the scale of E is doubled, so that integrating E from $\omega = 0$ to ∞ gives σ^2.

impulse at time $t - \tau$ contributes to the response of a network at time t in proportion to $K(\tau)$ and σ^2 may now be expressed as the sum of the average squares contributed by the individual (uncorrelated) impulses. When limits are taken properly, the result is:

$$\sigma^2 = \int_{-\infty}^{+\infty} [K(t)]^2 \, dt. \tag{17}$$

Equations (16) and (17) are, of course, consistent with (T-15), in Table I, which is a special form of Parseval's equation.

Referring again to Fig. 2, if the output $g(t)$ is interpreted as an estimate of the true future signal $s(t + \alpha)$, the error $\epsilon(t)$ must be

$$\epsilon(t) = g(t) - s(t + \alpha). \tag{18}$$

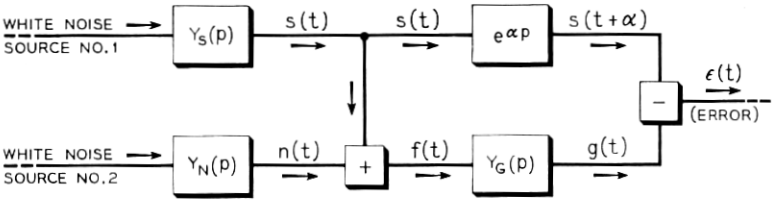


Fig. 6 — A physical model for the calculation of the error $\epsilon(t)$.

We may think of $\epsilon(t)$ as the output of the (unattainable) circuit shown in Fig. 6, in which the responses of the different parts are again described in frequency domain terms.* The average squared error σ^2 is now the sum of the contributions from the two uncorrelated white noise sources. Evaluating these in terms of (16) gives

$$\sigma^2 = \int_{-\infty}^{+\infty} (|\, Y_G \,|^2 N + |\, Y_G - e^{\alpha i\omega} \,|^2 S) \, d\omega. \tag{19}$$

The optimization problem may now be stated as follows: Given the class of permitted functions Y_G corresponding to all networks of the permitted sort, find the particular Y_G, say Y_M, such that σ^2 is a minimum. Assume tentatively that Y_M exists and let $\Delta_Y(p)$ be defined by

$$Y_G(p) = Y_M(p) + \Delta_Y(p). \tag{20}$$

* When $\alpha > 0$, the box marked $e^{\alpha p}$ is nonphysical, since $e^{\alpha p}$ corresponds to a negative delay, or ideal prediction. So long as the noise is present, the box is connected to an unavailable signal source, $s(t)$ without $n(t)$, whatever the value of α. These circumstances are what makes $\epsilon(t)$ the error, instead of an observable correction which could be used to determine $s(t + \alpha)$ exactly.

Under assumption vi, if $Y_G(p)$ in (20) is a permitted $Y_G(p)$, then

$$Y_M(p) + k\Delta_Y(p)$$

is a permitted $Y_G(p)$, where k is any positive or negative real constant.

Substituting $Y_M + k\Delta_Y$ for Y_G in (19) gives:

$$\begin{aligned}\sigma^2 = &\int_{-\infty}^{+\infty} (|\, Y_M \,|^2 N + |\, Y_M - e^{\alpha i\omega} \,|^2 S)\, d\omega \\ &+ k^2 \int_{-\infty}^{+\infty} (N + S)\, |\, \Delta_Y \,|^2\, d\omega \\ &+ k \int_{-\infty}^{+\infty} [Y_M(N + S) - e^{\alpha i\omega} S]\tilde{\Delta}_Y\, d\omega \\ &+ k \int_{-\infty}^{+\infty} [\tilde{Y}_M(N + S) - e^{-\alpha i\omega} S]\Delta_Y\, d\omega.\end{aligned} \tag{21}$$

In this expression the last two integrals are equal, for the following reasons: First, each of the two integrals is *real*, for the imaginary part of the integrand is always an *odd* function of ω. Second, the two integrals can at most be conjugates of each other, since their integrands are conjugates at real ω. Replacing the two integrals by twice the first leaves

$$\begin{aligned}\sigma^2 = &\int_{-\infty}^{+\infty} (|\, Y_M \,|^2 N + |\, Y_M - e^{\alpha i\omega} \,|^2 S)\, d\omega \\ &+ k^2 \int_{-\infty}^{+\infty} (N + S)\, |\, \Delta_Y \,|^2\, d\omega \\ &+ 2k \int_{-\infty}^{+\infty} [Y_M(N + S) - e^{\alpha i\omega} S]\tilde{\Delta}_Y\, d\omega.\end{aligned} \tag{22}$$

When $k = 0$, $Y_M + k\Delta_Y = Y_M$, and σ^2 must be a minimum. This will be true only if the coefficient of k in (22) is zero; hence the third integral must be zero. Furthermore, for Y_M to be a true optimum, the integral must be zero whatever permitted Y_G is used in (20) in determining Δ_Y. If the variable of integration, ω, is replaced by $p = i\omega$ (for convenience in what follows), this requires:

For every permitted Δ_Y :

$$\int_{-i\infty}^{+i\infty} [Y_M(N + S) - e^{\alpha p} S]\tilde{\Delta}_Y\, dp = 0. \tag{23}$$

Our principal concern, in Sections III and IV, will be the solution of (23) for Y_M, for different classes of permitted functions Y_G. Generally,

the class of permitted functions will exclude the obvious solution which makes the bracket, [ ], in (23) identically zero. Then the integrand will not be identically zero, but will have to be such that the integral will be zero for every $\tilde{\Delta}_Y$ which may be derived by using a permitted Y_G in (20).

When (23) is true, (22) may be written as follows, for every permitted Y_G:

$$\begin{aligned} \sigma^2 &= \sigma_M{}^2 + k^2 \int_{-\infty}^{+\infty} (N + S) \mid \Delta_Y \mid^2 d\omega \\ \sigma_M{}^2 &= \int_{-\infty}^{+\infty} (\mid Y_M \mid^2 N + \mid Y_M - e^{\alpha i \omega} \mid^2 S)\, d\omega. \end{aligned} \tag{24}$$

Here $\sigma_M{}^2$ is the σ^2 achieved with $Y_G = Y_M$, and the second term in σ^2 is clearly nonnegative.

Equation (24) implies the following situation: Under assumptions v and vi, a true minimum variance $\sigma_M{}^2$ exists (at least as a limit of a sequence of σ^2's corresponding to a sequence of permitted functions Y_G). Any solution of (23) for Y_M within the class of permitted functions Y_G must yield the true $\sigma_M{}^2$. Since $N + S$ is nonnegative in (23), no other permitted Y_G can yield a smaller σ^2. When $N + S$ is also nonzero at real frequencies, no other Y_G can yield as small a σ^2, and the solution for Y_M is at once unique.

When $N + S$ is zero at one or more real frequencies, the situation regarding uniqueness is not so clear, for any Δ_Y which is nonzero at those frequencies but zero at all other real frequencies will lead to a new Y_G, yielding the same $\sigma^2 = \sigma_M{}^2$. Under assumption iv, $N + S$ can be zero only at discrete frequencies. As a result, if there are two solutions for Y_M in (23), at least one must include transfer functions of filters with infinitesimal bandwidths. It is questionable whether zero-bandwidth filters may be called "physical"; in any event, they cannot be built and alternative solutions do not reduce $\sigma_M{}^2$. Accordingly, any solution of (23) for Y_M which does not include transfer functions of zero-bandwidth filters will be called unique, whether or not $N + S$ has zeros at real frequencies.

Certain statements about convergence will be useful later on. Under assumption v, Y_M will make σ^2 finite. In seeking Y_M, then, we may start by excluding all Y_G for which $\sigma^2 = \infty$. Since the two integrals in (24) are nonnegative, σ^2 will be bounded only if both integrals converge. Each of the two integrands is a sum of two nonnegative terms. It follows that each term must meet convergence conditions. Combining these yields an additional useful condition, which will be satisfied by the integrand in

(23). The five conditions may be written collectively as follows:

$$\begin{aligned}&\text{When } \omega \text{ is real and } \to \infty \text{ the following five functions } = O\omega^{-2}:\dagger\\&|\, Y_M \,|^2 N; \qquad |\, Y_M - e^{\alpha p} \,|^2 S\\&|\, \Delta_Y \,|^2 N; \qquad |\, \Delta_Y \,|^2 S\\&[Y_M(N + S) - e^{\alpha p} S]\tilde{\Delta}_Y\end{aligned} \tag{25}$$

2.6.1 *An Equivalent Formulation in the Time Domain*

Time-domain equivalents of (23) and (24) can easily be derived. The above analysis can be paralleled in time-domain terms or, alternatively, (T-14) and (T-15) of Table I can be applied directly to (23) and (24). Let $K_G(t)$, $K_M(t)$, $\Delta_K(t)$ be the inverse transforms of $Y_G(p)$, $Y_M(p)$, $\Delta_Y(p)$. Then (23) becomes:

For every permitted Δ_K,

$$\int_{-\infty}^{+\infty} [K_M(\tau) * \Phi_F(\tau) - \Phi_S(\tau + \alpha)]\Delta_K(\tau)\, d\tau = 0. \tag{26}$$

Equation (24) can be transformed into various time-domain equivalents, of which the following is perhaps the most interesting:

$$\begin{aligned}\sigma^2 &= \sigma_M{}^2 + \int_{-\infty}^{+\infty} [K_F * \Delta_K]^2\, d\tau,\\ \sigma_M{}^2 &= \int_{-\infty}^{-\infty} \{(K_M * K_N)^2 + [(K_M - \delta(\tau + \alpha)) * K_S]^2\}\, d\tau.\end{aligned} \tag{27}$$

2.7 *Breadth of the Optimum*

The Y_M determined by (23) may or may not be realizable with a finite network of lumped elements, even though assumption iv insures that Y_N and Y_S of Fig. 2 could be so realized. When Y_M cannot be realized with a finite network of lumped elements, it may be necessary to replace Y_M by a reasonable approximation to it, which can be so realized (and similarly for equivalent nonelectrical devices). A reasonable approximation will be one which can be realized in a reasonable way, and which makes σ^2 only a little greater than the minimum, $\sigma_M{}^2$.

The approximation of general "physical" network functions with "finite network approximations" is a familiar problem in general network

† Recall the meaning of O noted in Section 2.4.1.

theory, which need not be developed here. In this connection, however, it is important to note the situation described briefly below.

In engineering problems, the minimum exhibited by σ^2 as a function of Y_G is usually quite broad. In other words, Y_G may be made quite a little different from Y_M without increasing σ^2 very much. This has not been proved but is simply a matter of experience in problems of the engineering sort (and, in fact, it cannot even be stated exactly without assigning a more quantitative meaning to the expression "broad minimum").

A broad minimum does not mean that *all* small departures from Y_M have small effects on σ^2. For example, a change from order c/p^m at $p = \infty$, to order of c/p^{m-1} may yield only small departures from $Y_M(p)$ at all real frequencies, but it is likely to change $\sigma^2 = {\sigma_M}^2$ into $\sigma^2 = \infty$. Generally, reasonable *percentage* changes, relative to the magnitude of Y_M at corresponding frequencies, may be tolerated. The percentage changes may be frequency-dependent, and may be real or complex. The specific sensitivities to changes, however, will depend on the specific values of N and S.

The effect of specific departures from Y_M, in specific problems, may be calculated by means of (24) or (27).

III. GENERAL TECHNIQUES, IN TERMS OF SPECIFIC PROBLEMS

The remainder of the paper describes the calculation of Y_M from (23), and modifications thereof. The specific Y_M determined by (23) depends on the class of permitted functions Y_G, within which Y_M is to be the optimum choice. A number of different classes are of interest, on both theoretical and practical grounds. Furthermore, the appropriate techniques for calculating Y_M vary with the permitted class, in ways which are likely to be nontrivial.

In this section we consider some fairly general classes of permitted functions, which will illustrate general techniques. In Section IV we shall examine variations and special cases, chosen primarily for their engineering interest.

In describing the properties of the different classes, it will be convenient to use the following general notation:

$$\begin{aligned} &a.\ C_Y = \text{the class of permitted functions } Y_G, \text{ within which } Y_M \\ &\quad \text{is to be the optimum choice;} \\ &b.\ C_\Delta = \text{the class of functions } Y_{G1} - Y_{G2}, \text{ where } Y_{G1} \text{ and } Y_{G2} \\ &\quad \text{are any two } Y_G \text{ in } C_Y. \end{aligned} \tag{28}$$

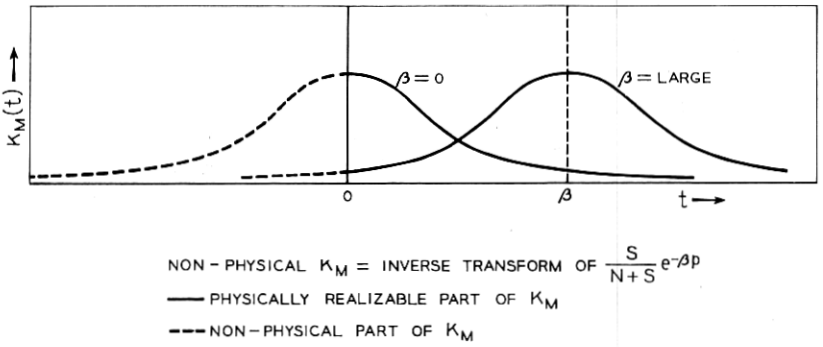


Fig. 7 — The optimum nonphysical impulse response when β is large.

The class C_Δ is completely determined by the class C_Y. In terms of C_Δ, assumption vi of Section 2.2 may be written as follows:

$$\text{If } \Delta_Y \text{ is in } C_\Delta, \qquad k\Delta_Y \text{ is in } C_\Delta. \tag{29}$$

3.1 *Optimum Nonphysical Network*

The integral in (23) will surely vanish if the integrand is zero identically for every permitted Δ_Y. The integrand will be zero if

$$Y_M = \frac{S}{N+S} e^{\alpha p}. \tag{30}$$

This Y_M, however, is generally nonphysical. First, it is not generally regular in the finite part of rhp. Second, when $\alpha > 0$ it behaves improperly at $p \to \infty$.

When α is negative, Y_M behaves properly at $p \to \infty$, but there are still singularities in the rhp [except when $S/(N+S)$ is a constant]. Let α equal $-\beta$. Then $g(t)$ is an estimate of $s(t)$ for a time β seconds in the past. If β is sufficiently large, Y_M can be approximated closely with a physical Y.* This is illustrated in Fig. 7, in time-domain terms. Note that the inverse transform of Y_M is symmetrical about the time $t = \beta$, which represents the length of the time interval from the time for which the signal $s(t - \beta)$ is estimated up to present time t.

A large β is appropriate for reducing data long after they are collected, such as reconstructing the flight of an experimental missile from recorded positions or velocities. Then β is the length of time by which the interval of observation extends beyond the time for which S is to be estimated.

* See Ref. 9, Section V.

When the Y_M of (30) is used in (24), the minimum average square error is (after some simple manipulation)

$$\sigma_M^2 = \int_{-\infty}^{+\infty} \frac{NS}{N+S}\, d\omega. \tag{31}$$

Then assumption v of Section 2.2 requires

$$\frac{NS}{N+S} = O\omega^{-2} \quad \text{when } \omega \to \infty. \tag{32}$$

The same restriction on N and S will surely still be necessary (although perhaps not sufficient) when the choice of Y_M is restricted to any subset of the present C_Y.

3.2 *Optimum Physical Network*

In this section, we restrict the function class of C_Y, from which Y_M is to be chosen, to the "physical" subset of the class permitted in the previous section.

$$C_Y = \text{the class of all physical frequency functions.} \tag{33}$$

by "physical", we again mean rrhp as in (8) and (9). Since the difference of two rrhp functions is also rrhp, (33) implies

$$C_\Delta = \text{the class of all physical frequency functions.} \tag{34}$$

The integrand in (23) can no longer be identically zero for all the $\tilde{\Delta}_Y$ permitted by C_Δ, but the restrictions on $\tilde{\Delta}_Y$ are such that they may be taken advantage of, in one way or other. Note that the C_Δ of (34) obeys (29), and hence also our assumption vi.

The optimum Y_M may be determined in the following way: First, (33) is applied to (23), to obtain tentative conditions on Y_M which are consistent with (33) and in which Δ_Y does not appear. As derived, these are sufficient conditions. Any corresponding Y_M, if one exists, must be the correct Y_M, but it is not at once apparent that one does exist. The necessity of the conditions is then established by demonstrating that a corresponding Y_M does, in fact, exist. (Recall that the correct Y_M is unique.) This is a procedure which is useful in many variations of the smoothing and prediction problem.

The integral in (23) will be zero if its integrand behaves like $H(p)$ of (12). By (25), the integrand behaves properly when ω is real and $\to \infty$. By (33), the factor $\tilde{\Delta}_Y$ in the integrand in (23) is rlhp. Therefore, if the remaining factor in the integrand in (23) is made rlhp both (12*a*) and

(12*b*) will be true, and (23) will be satisfied. Assembling this condition with (25) and (33) gives the following set of *sufficient* conditions on Y_M :

$$\begin{aligned} &a.\ Y_M && \text{is rrhp,} \\ &b.\ Y_M\,(N+S) - e^{\alpha p}S && \text{is rlhp,} \\ &c.\ \text{When } \omega \text{ is real and} \to \infty, \\ &\quad |\,Y_M\,|^2\,N \quad \text{and} \quad |\,Y_M - e^{\alpha p}\,|^2\,S = O\omega^{-2}. \end{aligned} \tag{35}$$

It remains to be shown that a Y_M meeting (35) does, in fact, exist. The demonstration is somewhat different for positive and negative values of α.

3.2.1 *Prediction for a Present or Future Time* ($\alpha \geqq 0$)

When α is positive in (35*b*) the behavior of $e^{\alpha p}S$ as $p \to \infty$ is consistent with rlhp. Then $Y_M(p)$ may reasonably be rational (under our assumption of rational N and S). When Y_M must be physical and α is nonzero and positive, $|\,Y_M - e^{\alpha p}\,|^2$ cannot $\to 0$, when ω is real and $\to \infty$. Then assumption v, as reflected in condition (25), requires $S = O\omega^{-2}$ as $\omega \to \infty$. (The limiting case of $\alpha = 0$ will be examined later.) When $S = O\omega^{-2}$ as $\omega \to \infty$, conditions (35) can easily be translated into the following set, which refers to poles and zeros in the finite part of the p plane:

a. The poles of Y_M are the lhp zeros of $N + S$,

b. At the lhp poles of N and S

$$Y_M = \frac{S}{N+S}\,e^{\alpha p}, \tag{36}$$

c. The degree of the numerator of Y_M is a minimum, within conditions *a* and *b*.

These conditions determine a Y_M uniquely. The Y_M so determined will also satisfy (35), provided the degree determined by (36*c*) is such that (35*c*) is satisfied. It is shown below that (36*c*) is, in fact, just consistent with (35*c*).

Recall that N and S are even functions, with half of their zeros and poles in each half plane. Also, $Y_M\tilde{Y}_M$, which is $|\,Y_M\,|^2$ at real ω, is an even function, in which exactly half of the zeros and poles are zeros and poles of Y_M. Then, under (36*a*), the number of (finite) poles of $Y_M\tilde{Y}_M$ is exactly the number of (finite) zeros of $(N + S)$. On the other hand,

the number of zeros of Y_M is *one less* than the number of lhp poles of $N + S$, for the scale of Y_M can be adjusted, as well as the zeros, in meeting (35*b*). Then the number of zeros of $Y_M \tilde{Y}_M$ is exactly *two less* than the number of poles of $N + S$. As a result,

$$| Y_M |^2 (N + S) \to c/\omega^2 \text{ as } \omega \to \infty. \tag{37}$$

Finally, since N and S are nonnegative at real ω (and hence cannot cancel each other at $\omega \to \infty$), both

$$| Y_M |^2 N \quad \text{and} \quad | Y_M |^2 S = O\omega^{-2} \text{ as } \omega \to \infty. \tag{38}$$

Since we already know that $S = O\omega^{-2}$, this is sufficient to establish (35).

Conditions (36) need further interpretations for special cases. When zeros or poles of $N + S$ occur *on* the axis of real frequencies, they occur in identical pairs.* One of each pair is to be interpreted as in the lhp. Certain zeros of $N + S$ may coincide with poles common to N and S, when they are computed from the numerators and denominators of the (rational) N and S. Lhp zeros of this sort are to be retained as poles of Y_M.†

When $\alpha = 0$, the second condition in (35*c*) becomes $| Y_M - 1 |^2 S = O\omega^{-2}$ as $\omega \to \infty$. So long as $S = O\omega^{-2}$ as $\omega \to \infty$, conditions (36) are still appropriate. Now, however, S may be nonzero, or even unbounded at $\omega = \infty$. When S is nonzero but bounded, $Y_M \to 1$ as $\omega \to \infty$. When S has poles at ∞, $(Y_M - 1) \to c/p^m$. When (36) is properly modified to include the new conditions, a unique Y_M is again determined, which again satisfies (35), provided N is such that (32) is still satisfied. When $S \not\equiv 0$ at ∞, however, (32) will not be satisfied unless $N = O\omega^{-2}$ as $\omega \to \infty$.

3.2.2 *Estimation for a Past Time* ($\alpha < 0$)

If $\alpha = -\beta$, (35) becomes:

$$\begin{aligned} &a.\ Y_M && \text{is rrhp,} \\ &b.\ Y_M(N + S) - e^{-\beta p} S && \text{is rlhp,} \\ &c.\ \text{When } \omega \text{ is real and} \to \infty, \\ &\quad | Y_M |^2 N \quad \text{and} \quad | Y_M - e^{-\beta p} |^2 S = O\omega^{-2}. \end{aligned} \tag{39}$$

When β is positive, $e^{-\beta p} S$ does not behave at infinity in an rlhp manner,

* A necessary consequence of the *positiveness* of spectra N and S.

† This is confirmed by (36*b*), which makes Y_M infinite at a lhp pole of S which is cancelled by a like pole of N in $N + S$.

as defined by (10). Hence $Y_M(p)$ can no longer be rational, but must contain a term which annuls the forbidden behavior at infinity.

Let Y_M be represented as

$$Y_M = A + \frac{S}{N+S} e^{-\beta p}. \tag{40}$$

Then, if (39*c*) is rearranged with due regard for (32) and the positiveness of N and S, (39) now becomes

$$\begin{aligned} &a.\quad A + \frac{S}{N+S} e^{-\beta p} \text{ is rrhp,} \\ &b.\quad A(N+S) \qquad\qquad \text{is rlhp,} \\ &c.^*\ \text{When } \omega \text{ is real and} \to \infty, \\ &\qquad |A|^2(N+S) = O\omega^{-2}. \end{aligned} \tag{41}$$

The function A can now be factored, recalling Y_F of Section 2.3 and Fig. 3:

$$\begin{aligned} &Y_F \tilde{Y}_F = N + S, \\ &Y_F\,,\ 1/Y_F \text{ are rrhp.} \end{aligned} \tag{42}$$

We can multiply the function in (41*a*) by Y_F, without changing its rrhp character. (If rrhp, it will remain rrhp; if not rrhp, it will remain not rrhp.) Similarly, we can divide the function in (41*b*) (arbitrarily) by $\tilde{Y}_F$. Then (40) and (41) may be written as follows [if (42) is again used]:

$$Y_M = \frac{1}{Y_F}\left(B + \frac{S}{\tilde{Y}_F} e^{-\beta p}\right),$$

$$\begin{aligned} &a.\ B + \frac{S}{\tilde{Y}_F} e^{-\beta p} \text{ is rrhp,} \\ &b.\ B \qquad\qquad \text{is rlhp,} \\ &c.\ \text{When } \omega \text{ is real and} \to \infty,\ |B|^2 = O\omega^{-2}. \end{aligned} \tag{43}$$

The conditions (43) can only be realized with a unique rational B. The poles of B are in the rhp but they are cancelled in Y_M by poles of

* This may be derived in the following way: Use the Y_M of (40) in the two functions in (39*c*). Add the two functions, rearrange, and separate out the function in (41*c*), making use of the following theorem: Let U_1 and U_2 be two functions of ω; if $U_1 = O\omega^{-2}$ and $U_2 = O\omega^{-2}$, then $U_1 + U_2 = O\omega^{-2}$; conversely, if $U_1 \vdash U_2 = O\omega^{-2}$, and both U_1 and U_2 are nonnegative, then $U_1 = O\omega^{-2}$ and $U_2 = O\omega^{-2}$.

$S/\tilde{Y}_F$. Because the poles of B are in the rhp, Y_M cannot be realized exactly by a finite network of lumped elements and delay lines. It can be approximated arbitrarily closely, however, by transversal filter techniques.

3.2.3 *A Time-Domain Interpretation*

A time-domain interpretation of Y_M may be derived from (35) which coincides with, for example, Bode and Shannon's interpretation.[9] This may be accomplished by using the functions Y_F and $\tilde{Y}_F$ much as in the previous section.

Since $\tilde{Y}_F$ and $1/\tilde{Y}_F$ are both rlhp, the function in (35*b*) may be multiplied by $1/\tilde{Y}_F$ without altering its rlhp character. (If rlhp, it will remain rlhp; if not rlhp, it will remain not rlhp.) Then (35*b*) becomes

$$Y_M Y_F - \frac{S}{\tilde{Y}_F} e^{\alpha p} \text{ is rlhp.} \tag{44}$$

By (T-13), the inverse transform must be 0 when $t > 0$. Hence, the two terms in the difference must be equal when $t > 0$. Since Y_M and Y_F are both physical, $Y_M Y_F$ is physical and its inverse transform must be zero when $t < 0$. Then (41) becomes, by (T-10),

$$\begin{aligned} K_M * K_F &= \text{inverse transform of } \frac{S}{\tilde{Y}_F} e^{\alpha p} \text{ when } t > 0 \\ &= 0 \text{ when } t < 0. \end{aligned} \tag{45}$$

Equation (45) determines $K_M * K_F$ uniquely. The transform is $Y_M Y_F$, and Y_M can be found by dividing out the (minimum phase, physical) factor Y_F. Note that (45) can be solved for K_M even though the spectra are nonrational: $F = N + S$ must be such that Y_F can be found from $F = Y_F \tilde{Y}_F$ by means of the loss-phase integral; and various transforms and convolutions must be evaluated.

3.2.4 *A Further Interpretation*

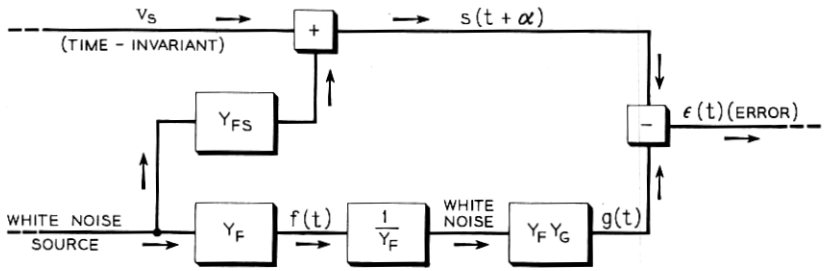
A further interpretation of (45) is useful when one seeks certain generalizations of the present problem. The time series which are fundamental to the problem are the true signal $s(t)$ and the observed signal $f(t)$. When $s(t)$ is to be predicted for a specific time, $t + \alpha$, from values of $f(t)$ observed over a specific interval, $(-\infty < \tau < t)$, only certain of the linear covariances, describing statistical characteristics of s and

f, are actually pertinent to the problem. The pertinent linear covariances are:

1. the auto-covariance of the observed signal f at any two times, t_1, t_2, in the observation interval;
2. the cross-covariance between f at any time, t_1, in the observation interval and s at the special time $t + \alpha$;
3. the expected value of s^2, which is the zero lag covariance of s, at the special time $t + \alpha$.

The auto-covariance function Φ_s, or the corresponding spectrum S, is pertinent to our present problem only as it affects the covariances of f through the relation $f(t) = s(t) + m(t)$.

It follows from the above that the original Gaussian model may be replaced by any other Gaussian model which retains the same pertinent linear covariances, without affecting either Y_M or ${\sigma_M}^2$. For example, the model represented by Fig. 6 may be replaced by that shown in Fig. 8. The network with frequency function Y_F generates an $f(t)$ with the correct auto-covariance (in accordance with Section 2.3 and Fig. 3). The network with frequency function Y_{FS}, in the presence of the other network, supplies the correct cross-covariance between $f(t_1)$ and $s(t + \alpha)$. (The cross-covariance arises from the sharing of the same white noise, by s and f.) Note that Y_{FS} in Fig. 8 is exactly the function $(S/\tilde{Y}_F)e^{\alpha p}$ in (44). The single (time-invariant) Gaussian random variable v_s contributes only to s, and gives the correct variance of s when added to the contribution from the white noise. Note that the variance of v_s is exactly the ${\sigma_M}^2$ of Section 3.1, corresponding to the optimum nonphysical Y_M.



$$Y_F \tilde{Y}_F = S + N \qquad Y_{FS} = \frac{S}{\tilde{Y}_F} e^{\alpha p} \qquad E\left[v_S^2\right] = \int_{-\infty}^{+\infty} \frac{NS}{N+S} d\omega$$

Fig. 8 — An alternate physical model, which retains the pertinent covariances.

Because Y_F has been chosen in such a way that $1/Y_F$ is physical, the white noise of Fig. 8 can be recovered. If $Y_M Y_F = Y_{FS}$, the estimate $g(t)$, of $s(t + \alpha)$, is exactly the contribution of the white noise to $s(t + \alpha)$, and the error in the estimate is exactly v_s. If Y_M must be physical, $Y_M Y_F$ can yield only the contributions to $s(t + \alpha)$ from present and past values of the white noise. This is roughly the description of the solution used by Bode and Shannon (in Ref. 9, Section VII).

3.3 *Optimum Network with a Finite Memory*

In Section 3.2 we assumed that all past values of $f(t)$ (the signal plus noise) were available, back to $t = -\infty$. Furthermore, the $Y_M(p)$ of (35) does generally lead to the utilization of all past values. In practical applications, the corresponding $K_M(\tau)$ is usually very small when the age of data τ is sufficiently large, say $\tau > \tau_m$; and data older than τ_m may reasonably be neglected. Frequently, however, $f(t)$ is available only for a smaller interval, say $0 < \tau < T$; and then the procedure must be revised. This is the central problem considered by Zadeh and Ragazzini.[3] It may be referred to as the "finite memory" problem, as opposed to the "infinite memory" considered in Section 3.2.

We can restrict ourselves to values of $f(t)$ no older than T by making our function class C_Y correspond to physical networks with memories which extend only T seconds into the past. With fixed networks of this sort, T is constant, and the used data begins at a variable time $t - T$. When $f(t)$ is available from a fixed starting time, t_0, to present time t, T is a function of t. Then a variable network must be designed with a response which equals or approximates the response of a different fixed network at each different t.

Our smoothing and prediction device will remember only T seconds into the past if the impulse response K_G is restricted by

$$K_G(t) = 0, \quad \text{except when} \quad 0 < t < T. \tag{46}$$

An equivalent frequency domain restriction may be derived as follows: Referring to Fig. 9, if $K_G(t) = 0$ except when $0 < t < T$, it follows that $K_G(-t) = 0$ except when $-T < t < 0$. But then $K_G[-(t - T)] = 0$ except when $0 < t < T$, which meets the conditions for physicalness. If Y_G is the transform of K_G, then $\tilde{Y}_G e^{-Tp}$ is the transform of $K_G[-(t - T)]$, by (T-5) and (T-9). Thus (46) corresponds to

$$\begin{aligned} &C_Y \text{ is the class of functions } Y_G(p) \text{ such that:} \\ &\quad a.\ Y_G \qquad \text{is rrhp,} \\ &\quad b.\ \tilde{Y}_G e^{-Tp} \text{ is rrhp.} \end{aligned} \tag{47}$$

Since differences of K_G's of the form (46) also obey (47), it follows that

C_Δ is the class of functions Δ_Y such that:

$$\begin{aligned} &a.\ \Delta_Y \quad \text{is rrhp,} \\ &b.\ \tilde{\Delta}_Y e^{-Tp} \text{ is rrhp.} \end{aligned} \tag{48}$$

Note that (48) is consistent with (29) and hence also with assumption vi of Section 2.2.

The following is an important property of functions in the class (47): Replacing Y_G by $\tilde{Y}_G$ maps singularities from either half plane into the other. Then (47) excludes singularities from all finite parts of the p plane. Let rfpp mean "regular in the finite part of the plane". Then, (47) implies that

$$Y_G \text{ is rfpp.} \tag{49}$$

There will be no Y_M of the class (47) which satisfies (35). Hence, a new relation must be derived from (23). To take advantage of (47), Y_M may be expressed in the following arbitrary way:

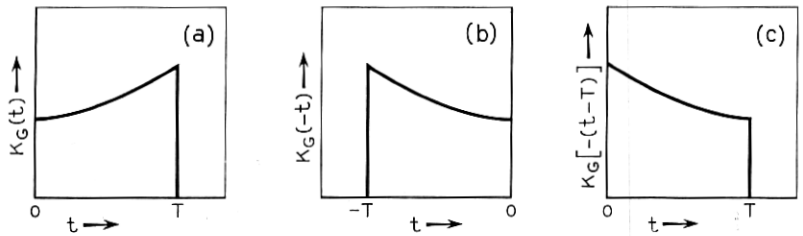
$$Y_M = A + Be^{-Tp}. \tag{50}$$

If this expression is used in (23), the integral in (23) may be divided into two parts, as follows:

$$\int_{-i\infty}^{+i\infty} [A(N+S) - e^{\alpha p}S]\tilde{\Delta}_Y\, dp + \int_{-i\infty}^{+i\infty} [B(N+S)]e^{-Tp}\tilde{\Delta}_Y\, dp = 0. \tag{51}$$

The two integrals in (51) will be zero, individually, if the two integrands behave like $H(p)$ of (12). This will be true under conditions noted below.

In the first integrand, $\tilde{\Delta}_Y$ is rlhp. In the second, $\tilde{\Delta}_Y e^{-Tp}$ is rrhp by



$K_G(t)$ = INVERSE TRANSFORM OF $Y(p)$ = 0 WHEN $t < 0$

$K_G(-t)$ = INVERSE TRANSFORM OF $\tilde{Y}(p)$ = 0 WHEN $t > 0$

$K_G[-(t-T)]$ = INVERSE TRANSFORM OF $\tilde{Y}(p)e^{-Tp}$ = 0 WHEN $t < 0$

Fig. 9 — Impulse responses of finite duration.

(48). For (12*a*) to be true, the other factor in each integrand must be regular in the same half plane. The convergence condition (12*b*) must also be applied, and it can be applied if (25) is properly taken into account. When (47) also is transcribed the following conditions may be assembled as sufficient conditions on A and B:

$$\begin{aligned}
&a.\ A + Be^{-Tp} && \text{is rrhp,}\\
&b.\ \tilde{B} + \tilde{A}e^{-Tp} && \text{is rrhp,}\\
&c.\ A(N + S) - e^{\alpha p}S && \text{is rlhp,}\\
&d.\ B(N + S) && \text{is rrhp,}\\
&e.\ \text{When } \omega \text{ is real and } \to \infty\text{, the following functions } = O\omega^{-2}\text{:}\\
&\qquad |A|^2N, \qquad |A - e^{\alpha p}|^2S\\
&\qquad |B|^2N, \qquad |B|^2S.
\end{aligned} \tag{52}$$

Functions A and B which satisfy (52) do, in fact, exist; hence the conditions are necessary as well as sufficient. As in Section 3.2, the solution takes different forms for positive and negative values of α.

3.3.1 *Prediction for a Present or Future Time* ($\alpha \geqq 0$).

When α is positive in (52*c*) A and B are rational. By the definition of rrhp in (8), conditions *a* and *b* will be satisfied if the rational A and B satisfy (49). When $S = O\omega^{-2}$ as $\omega \to \infty$ (as it must if α is nonzero and positive), (52) may be translated into the following set of conditions, which refer to poles and zeros in the finite part of the p plane:

$$\begin{aligned}
&a.\ \text{The poles of } A \text{ and } B \text{ are the zeros of } N + S\text{, in both half planes,}\\
&b.\ \text{At the lhp poles of } N \text{ and } S\\
&\qquad A = \frac{S}{N + S}\, e^{\alpha p},\\
&c.\ \text{The rhp poles of } N \text{ and } S \text{ are zeros of } B,\\
&d.\ \text{At the poles of } A \text{ and } B\\
&\qquad -\frac{A}{B} = e^{-Tp},\\
&e.\ \text{The combined degrees of the numerators of } A \text{ and } B \text{ are to be as small as possible and the relative degrees are to be adjusted as required by (52}e\text{).}
\end{aligned} \tag{53}$$

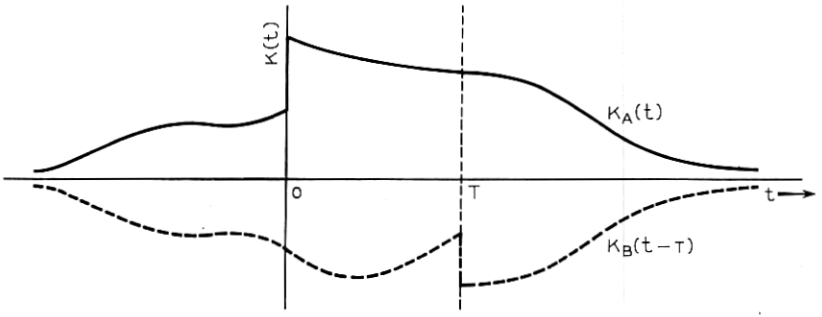


Fig. 10 — Inverse transforms of A and Be^{-Tp}.

An analysis of degrees, as in Section 3.2.1, shows that (53) determines a unique A and B, which satisfy (52) when $S = O\omega^{-2}$ as $\omega \to \infty$. When $\alpha = 0$, S is not necessarily 0 at ∞. Then (52*e*) imposes additional conditions on A. When (53) is suitably modified to include these new conditions, a unique A and B are again determined.

It is interesting to examine the forms of the inverse transforms of A and Be^{-Tp}, which together make up the inverse transform $K_M(t)$ of $Y_M(p)$. The inverse transforms of general rational functions are equal piecewise to sums of exponentials. The exponentials are different for $t > 0$ and $t < 0$, and there may be discontinuities (and also δ functions) at $t = 0$. In Be^{-Tp}, the discontinuities are displaced to $t = T$. Then the two time functions may take the form suggested in Fig. 10.

As in Section 3.2.2, Y_M cannot generally be realized with a finite network of lumped elements and ideal delay lines. It can be approximated by a transversal filter; plus differentiations if $K_M(t)$ is not a bounded function. The transversal filter need not involve transmission times greater than T. If preferred, the rrhp parts of A and B can be separated out and realized with a finite network of lumped elements plus a single delay line.

3.3.2 *Prediction for a Past Time* ($\alpha < 0$).

When $\alpha = -\beta$, (52*c*) excludes a rational A. Recalling (40), we let A be

$$A = D + \frac{S}{N + S} e^{-\beta p}. \tag{54}$$

Then when $\alpha = -\beta$, (52) becomes the following:*

$$\begin{aligned}
&a.\ D + \frac{S}{N+S}\, e^{-\beta p} + Be^{-Tp} && \text{is rrhp,}\\
&b.\ \tilde{B} + \tilde{D}e^{-Tp} + \frac{S}{N+S}\, e^{-(T-\beta)p} && \text{is rrhp,}\\
&c.\ D(N+S) && \text{is rlhp,}\\
&d.\ B(N+S) && \text{is rrhp,}\\
&e.\ \text{When } \omega \text{ is real and } \to \infty, \text{ the following functions } = O\omega^{-2}:\\
&\qquad |D|^2N, \qquad |D|^2S\\
&\qquad |B|^2N, \qquad |B|^2S.
\end{aligned} \tag{55}$$

Consider the term $e^{-(t-\beta)p}$ in (55*b*). When $\beta < T$, the exponential behaves at ∞ in a manner suitable for an rrhp function, as defined in (8). Then D and B are both rational, and can be calculated from zero and pole conditions implied by (55).

When $\beta > T$, the exponential $e^{-(T-\beta)p}$ is not suitable for an rrhp function, and it must be annulled by an exponential term in B. Let

$$B = J - \frac{S}{N+S}\, e^{-(\beta-T)p}. \tag{56}$$

Using both (54) and (56) in (50) and (52) gives

$$Y_M = D + Je^{-Tp}$$

$$\begin{aligned}
&a.\ D + Je^{-Tp} && \text{is rrhp,}\\
&b.\ \tilde{J} + \tilde{D}e^{-Tp} && \text{is rrhp,}\\
&c.\ D(N+S) && \text{is rlhp,}\\
&d.\ J(N+S) - Se^{-(\beta-T)p} && \text{is rrhp,}\\
&e.\ \text{When } \omega \text{ is real and } \to \infty, \text{ the following functions } = O\omega^{-2}:\\
&\qquad |D|^2N, \qquad |D|^2S\\
&\qquad |J|^2N, \qquad |J - e^{-(\beta-T)p}|^2S.
\end{aligned} \tag{57}$$

When $\beta > T$, D and J are rational, and can be found from (57).

A comparison of (57) and (52) indicates the following relationship:†

* Because N and S are even, $\tilde{N} = N$ and $\tilde{S} = S$.

† In making the comparison, recall that Y is rrhp when $\tilde{Y}$ is rlhp, and vice versa.

If Y_M is the optimum Y_G for estimating $s(t + \alpha)$, then $\tilde{Y}_M e^{-Tp}$ is the optimum Y_G for estimating $s(t - T - \alpha)$. (58)

The two times, $t + \alpha$ and $t - T - \alpha$, are symmetrically located relative to the interval $t - T < \tau < t$, for which values of $f(t)$ are available. The relation (58) may be viewed as a direct consequence of the symmetry of auto-covariance functions.

3.3.3 *A Time-Domain Interpretation*

In Section 3.2.3 we derived Bode and Shannon's explicit time-domain solution of the infinite memory problem. When memories must be finite there are difficulties which exclude an explicit time-domain solution of the same sort. The difficulties themselves are informative, however, and a simple time-domain analysis also furnishes an alternative derivation of (52).

The starting point is now (26) in place of (23). Like K_G, Δ_K is now restricted by (46). Then the limits of integration may just as well be reduced to $0 < \tau < T$ and (26) becomes

$$\int_0^T [K_M(\tau) * \Phi_F(\tau) - \Phi_S(\tau + \alpha)]\Delta_K(\tau)\, d\tau = 0. \tag{59}$$

The variation function Δ_K can have any arbitrary values at times within the interval of integration. Therefore, the other factor in the integrand must be zero, over the same interval $0 < \tau < T$. Referring to Fig. 11, we may divide it into two parts, one of which is zero when $\tau > 0$, and the other when $\tau < T$. These may be written as follows:

$$K_M(\tau) * \Phi_F(\tau) - \Phi_S(\tau + \alpha) = K_U(\tau) + K_V(\tau - T), \tag{60}$$

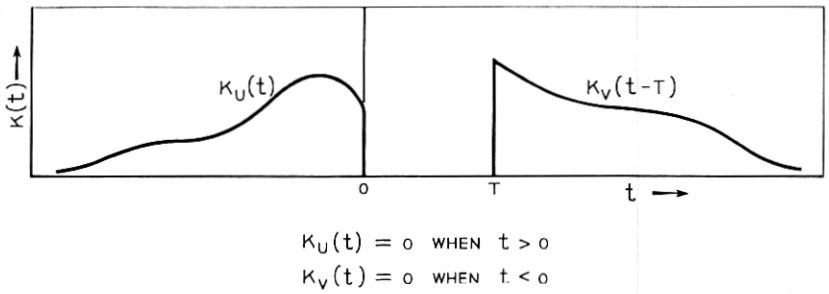


Fig. 11 — The functions K_u and K_v.

in which

$$K_U(\tau) = 0 \quad \text{when} \quad \tau > 0,$$
$$K_V(\tau) = 0 \quad \text{when} \quad \tau < 0.$$

Take the transform of both sides of (60), and recall that the transforms of K_M, Φ_F, Φ_S are Y_M, $N + S$, S. Then

$$Y_M(N + S) - e^{\alpha p}S = U + Ve^{-Tp},$$
$$U \text{ is rlhp,} \tag{61}$$
$$V \text{ is rrhp.}$$

Compare this with (35*b*) of Section 3.2.

In Section 3.2.3, we divided (35b) by $\tilde{Y}_F$ to obtain (44) and a simple physical interpretation. What happens if (61) is divided by $\tilde{Y}_F$? The result is as follows:

$$Y_M Y_F - \frac{S}{\tilde{Y}_F} e^{\alpha p} = \frac{U}{\tilde{Y}_F} + \frac{V}{\tilde{Y}_F} e^{-Tp},$$
$$U/\tilde{Y}_F \text{ is rlhp,} \tag{62}$$
$$V/\tilde{Y}_F \text{ is not rrhp.}$$

Because $V/\tilde{Y}_F$ is not rrhp, the inverse transform of the left-hand side is *not* zero in the interval $0 < \tau < T$ (nor in any other significant interval). Furthermore, while a $Y_M Y_F$ which is rrhp still implies that Y_M is rrhp, it is not simple to solve for $Y_M Y_F$ in such a way that $\tilde{Y}_M e^{-Tp}$ is also rrhp. Thus division by $\tilde{Y}_F$ is no longer useful.

To obtain an alternative derivation of (52), use the Y_M of (50) in (61). Then

$$(A + Be^{-Tp})(N + S) - e^{\alpha p}S = U + Ve^{-Tp},$$
$$U \text{ is rlhp,} \tag{63}$$
$$V \text{ is rrhp.}$$

Equate, separately, the terms which do and do not involve e^{-Tp}, and cancel the rrhp factor e^{-Tp} from those that do. Then

$$A(N + S) - e^{\alpha p}S = U, \text{ which is rlhp,}$$
$$B(N + S) = V, \text{ which is rrhp.} \tag{64}$$

These equations correspond exactly to conditions *c* and *d* of (52).

3.3.4 *A Somewhat More General Method*

In some problems (59) must be replaced by a slightly more general expression. In particular, $\Phi_S(\tau + \alpha)$, the last term in the square brackets, sometimes must be replaced by a more general function of τ, say $K_c(\tau)$. This is characteristic of problems involving linear constraints, which will be discussed further in Section IV. The methods described above may be modified in such a way that the transform of $K_c(\tau)$ need not be rational, provided Φ_F still corresponds to a rational spectrum.

Suppose Y_M is to be determined from conditions of the following sort (in which Y_M, Δ_Y, F, Y_C are of course the transforms of K_M, Δ_K, Φ_F, K_C):

$$\begin{aligned}
&a.\ \int_0^T [K_M(\tau) * \Phi_F(\tau) - K_C(\tau)]\Delta_K(\tau)\, d\tau = 0;\\
&b.\ K_M(t),\ \Delta_K(t) = 0 \text{ except when } 0 < t < T\\
&\quad [\text{then } Y_M,\ \Delta_Y \text{ belong to the classes (47), (48)}];\\
&c.\ |\, Y_M\,|^2 F = O\omega^{-2}, \text{ when } \omega \text{ is real and } \to \infty;\\
&d.\ \Phi_F(t) = \text{an auto-covariance function},\\
&\quad F = Y_F(p)\tilde{Y}_F(p) = \text{a rational function};\\
&e.\ K_C(t) = \text{a given time function, } Y_C(p) \text{ need not be rational.}
\end{aligned} \tag{65}$$

Let a new time function $\hat{K}_C(t)$ with transform $\hat{Y}_C(p)$ be defined as follows:

$$\begin{aligned}
\hat{K}_C &= K_C(t) \quad \text{when} \quad 0 < t < T,\\
&= 0 \quad \text{when} \quad t < 0 \text{ or } > T.
\end{aligned} \tag{66}$$

Then, exactly as with Y_M, we must have

$$\hat{Y}_C(p) \quad \text{and} \quad \tilde{\hat{Y}}_C e^{-Tp} \text{ are rrhp and rfpp.} \tag{67}$$

In these terms (65) implies the following, replacing our previous (60):

$$\begin{aligned}
K_M(\tau) * \Phi_F(\tau) - \hat{K}_C(\tau) &= \hat{K}_U(\tau) + \hat{K}_V(t - \tau),\\
\hat{K}_U(\tau) &= 0 \quad \text{when} \quad \tau > 0,\\
\hat{K}_V(\tau) &= 0 \quad \text{when} \quad \tau < 0.
\end{aligned} \tag{68}$$

The functions $\hat{K}_U$ and $\hat{K}_V$ are like K_U and K_V of (60), except that they include the "tails" of K_C lying outside the interval $0 < \tau < T$.

Now take the transforms of all functions in (68), and solve for $Y_M(p)$. If $\hat{U}$ and $\hat{V}$ are the transforms of K_U and K_V, the result is

$$Y_M = \frac{\hat{Y}_C + \hat{U} + \hat{V}e^{-Tp}}{F},$$

$$\begin{aligned} &a.\ \hat{U} \text{ is rlhp,} \\ &b.\ \hat{V} \text{ is rrhp,} \\ &c.\ |\, Y_M \,|^2 F = O\omega^{-2} \text{ when } \omega \text{ is real and } \to \infty. \end{aligned} \tag{69}$$

The functions $\hat{U}$ and $\hat{V}$ must be such that Y_M is rfpp. But $\hat{Y}_C$ itself is rfpp. As a result,

$$\begin{aligned} &a.\ \text{The finite poles of } \hat{U} = \text{the rhp poles of } F; \\ &b.\ \text{The finite poles of } \hat{V} = \text{the lhp poles of } F; \\ &c.\ \text{If } p = p_k \text{ at a finite zero of } F, \\ &\qquad \hat{Y}_C(p_k) + \hat{U}(p_k) + \hat{V}(p_k)e^{-Tp_k} = 0; \\ &d.\ \text{When } \omega \text{ is real and } \to \infty, \hat{U} \text{ and } \hat{V} \text{ must behave} \\ &\qquad \text{in such a way that } |\, Y_M \,|^2 F = O\omega^{-2}. \end{aligned} \tag{70}$$

These conditions determine a unique rational $\hat{U}$ and $\hat{V}$ when F is rational, provided $K_C(t)$ satisfies certain requirements relating to continuity. If $F \to C\omega^{-2m}$ as $\omega \to \infty$, then $K_C(t)$ and its first $m - 1$ derivatives must be continuous in the interval $0 < t < T$. When the continuity condition is violated, there will be no Y_M which meets all the conditions (65), unless (65c) is modified. (See Section 4.3 for an interpretation in connection with a particular application.)

The transform $Y_C(p)$ of $K_C(t)$ need not be rational. Furthermore, when $K_C(t)$ is given the time function can be calculated without actually evaluating $Y_C(p)$, except at the special points $p = p_k$. In particular, the following two relations may be used in evaluating the inverse transform of the right hand side of (69), with help of (70):

$$\begin{aligned} &a.\ \hat{Y}_C(p_k) = \frac{1}{\sqrt{2\pi}} \int_0^T K_C(\tau)e^{-pk\tau}\, d\tau; \\ &b.\ \text{Inverse transform of } \frac{\hat{Y}_C}{F} = \frac{1}{\sqrt{2\pi}} \int_0^T K_C(\tau)K_{(1/F)}(t - \tau)\, d\tau, \\ &\qquad \text{where } K_{(1/F)}(t) = \text{inverse transform of } 1/F. \end{aligned} \tag{71}$$

3.4 *Simultaneous Optimization of Two Network Functions*

In each of our problems so far we have been required to find a single frequency function, $Y_M(p)$, which is to represent the optimum choice of a single linear operator, $Y_G(p)$. There are other problems, however, in which two or more frequency functions are to be found, corresponding to the simultaneously optimum choices of two or more different but related linear operators. This section develops general methods in terms of one such problem.

Suppose the same signal can be observed in two different ways, or at two different places, involving contamination with noise from two different (uncorrelated) sources. An important example which will be discussed later (but not the only example) is the observation of a single physical variable with instruments of two different kinds. There are now two observed time functions, $f_1(t)$ and $f_2(t)$, related to $s(t)$ by

$$\begin{aligned} f_1(t) &= s(t) + n_1(t), \\ f_2(t) &= s(t) + n_2(t). \end{aligned} \tag{72}$$

The two different time functions are to be modified by (different) linear operations, and the results are to be added, to obtain an optimum estimate of $s(t + \alpha)$. All the assumptions of Section 2.2 are to be retained. Thus, Gaussian ensembles may be substituted for $s(t)$, $n_1(t)$ and $n_2(t)$, and Fig. 2 may be replaced by Fig. 12.

From Fig. 12, the following integral is easily obtained, in place of (19):

$$\sigma^2 = \int_{-\infty}^{+\infty} (|\, Y_{G1}\,|^2 N_1 + |\, Y_{G2}\,|^2 N_2 + |\, Y_{G1} + Y_{G2} - e^{\alpha i\omega}\,|^2 S)\, d\omega. \tag{73}$$

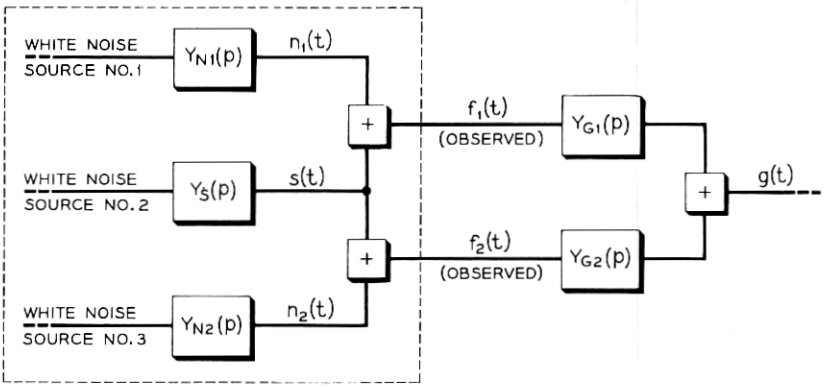


Fig. 12 — A physical model for a signal which is observed in two different ways.

Here N_1 and N_2 are, of course, the power spectra of the two noise ensembles (assumed to be uncorrelated). Let Y_{M1} and Y_{M2} be the optimum choices of Y_{G1} and Y_{G2}, and define Δ_{Y1} and Δ_{Y2} by

$$\begin{aligned} Y_{G1} &= Y_{M1} + \Delta_{Y1}\,, \\ Y_{G2} &= Y_{M2} + \Delta_{Y2}\,. \end{aligned} \tag{74}$$

Then (73) may be replaced by

$$\begin{aligned} \sigma^2 = &\int_{-\infty}^{+\infty} (|\, Y_{M1}\,|^2 N_1 + |\, Y_{M2}\,|^2 N_2 + |\, Y_{M1} + Y_{M2} - e^{\alpha p}\,|^2 S)\, d\omega \\ &+ \int_{-\infty}^{+\infty} (|\, \Delta_{Y1}\,|^2 N_1 + |\, \Delta_{Y2}\,|^2 N_2 + |\, \Delta_{Y1} + \Delta_{Y2}\,|^2 S)\, d\omega \\ &+ 2\int_{-\infty}^{+\infty} \{[Y_{M1}(N_1 + S) + Y_{M2}S - e^{\alpha p}S]\tilde{\Delta}_{Y1} \\ &\qquad + [Y_{M1}S + Y_{M2}(N_2 + S) - e^{\alpha p}S]\tilde{\Delta}_{Y2}\}\, d\omega. \end{aligned} \tag{75}$$

The first integral in (75) is σ_M^2, corresponding to a Δ_{Y1} and Δ_{Y2} which are zero identically. The second integral is positive for any Δ_{Y1} and Δ_{Y2} which are not zero identically. Then, if Δ_{Y1} and Δ_{Y2} each obeys (29), the last integral must be zero. This replaces (23) by the following:

For every permitted Δ_{Y1} and Δ_{Y2}:

$$\begin{aligned} \int_{-i\infty}^{+i\infty} &\{[Y_{M1}(N_1 + S) + Y_{M2}S - e^{\alpha p}S]\tilde{\Delta}_{Y1} \\ &\qquad + [Y_{M1}S + Y_{M2}(N_2 + S) - e^{\alpha p}S]\tilde{\Delta}_{Y2}\}\, dp = 0. \end{aligned} \tag{76}$$

The one integral can often be split into two, each involving one $\tilde{\Delta}_Y$, and each equal to zero. The separation may not be permissible however, when $N_1/S = O\omega^{-2}$ at $\omega = \infty$. An examination of convergence rules, noted below, will clarify this statement. Convergence conditions may again be obtained, by excluding at once, all Y_M's and Δ_Y's for which $\sigma^2 = \infty$. Conditions are easily derived from the first two integrals in (75). These may be combined to obtain a condition on the integrand in (76). The result is

When ω is real and $\to \infty$, the following seven functions $= O\omega^{-2}$:

$$\begin{aligned} &|\, Y_{M1}\,|^2 N_1\,, \quad |\, Y_{M2}\,|^2 N_2\,, \quad |\, Y_{M1} + Y_{M2} - e^{\alpha p}\,|^2 S, \\ &|\, \Delta_{Y1}\,|^2 N_1\,, \quad |\, \Delta_{Y2}\,|^2 N_2\,, \quad |\, \Delta_{Y1} + \Delta_{Y2}\,|^2 S, \\ &\text{and the integrand in (76).} \end{aligned} \tag{77}$$

Note that $|\Delta_{Y1}|^2 S$ and $|\Delta_{Y2}|^2 S$ do *not* necessarily $= O\omega^{-2}$, provided they are restricted by a linear relation at $\omega \to \infty$.*

3.4.1 *Optimum Physical Networks When* $\alpha \geqq 0$

Suppose Y_{M1} and Y_{M2} are to be selected from the class of all "physical" frequency functions, that is, from the C_Y of (33). Then Δ_{Y1} and Δ_{Y2} belong to the C_Δ of (34). Following our usual procedure, we fit the integrand of (76) to $H(p)$ of (12). Behavior at ∞ is already proper, and $\tilde{\Delta}_{Y1}$ and $\tilde{\Delta}_{Y2}$ are both rlhp. Then we have only to make the two associated factors also rlhp. Assembling these requirements with conditions of physicalness and convergence gives:

$$\begin{aligned}
&a.\ Y_{M1} && \text{is rrhp,}\\
&b.\ Y_{M2} && \text{is rrhp,}\\
&c.\ Y_{M1}(N_1 + S) + Y_{M2}S - e^{\alpha p}S && \text{is rlhp,}\\
&d.\ Y_{M1}S + Y_{M2}(N_2 + S) - e^{\alpha p}S && \text{is rlhp,}\\
&e.\ \text{When } \omega \text{ is real and } \to \infty, \text{ the following three functions} = O\omega^{-2}:
\end{aligned} \tag{78}$$

$$|Y_{M1}|^2 N_1\,, \qquad |Y_{M2}|^2 N_2\,, \qquad |Y_{M1} + Y_{M2} - e^{\alpha p}|^2 S.$$

When $\alpha \geqq 0$, Y_{M1} and Y_{M2} are rational (assuming rational N_1, N_2, S). It is not at once apparent, however, what the poles of Y_{M1} and Y_{M2} will be. From (78c) and (78d), we can define two rlhp functions, say $\tilde{U}_1$ and $\tilde{U}_2$, by:

$$\begin{aligned}
Y_{M1}(N_1 + S) + Y_{M2}S - e^{\alpha p}S &= \tilde{U}_1\,,\\
Y_{M1}S + Y_{M2}(N_2 + S) - e^{\alpha p}S &= \tilde{U}_2\,.
\end{aligned} \tag{79}$$

Solving for Y_{M1} and Y_{M2} gives:

$$\begin{aligned}
Y_{M1} &= \frac{\tilde{U}_1 N_2 + (\tilde{U}_1 - \tilde{U}_2)S + e^{\alpha p}SN_2}{N_1N_2 + (N_1 + N_2)S}\,,\\
Y_{M2} &= \frac{\tilde{U}_2 N_1 - (\tilde{U}_1 - \tilde{U}_2)S + e^{\alpha p}SN_1}{N_1N_2 + (N_1 + N_2)S}\,.
\end{aligned} \tag{80}$$

Since Y_{M1} and Y_{M2} are to be rrhp, while $\tilde{U}_1$ and $\tilde{U}_2$ are to be rlhp, the

* For example, suppose N_1, $N_2 = O\omega^{-4}$ as $\omega \to \infty$, and $S = O\omega^{-2}$. Then $|\Delta_{Y1}|^2$ and $|\Delta_{Y2}|^2$ may $= O\omega^{+2}$, provided Δ_{Y1} and Δ_{Y2} are so related that $|\Delta_{Y1} + \Delta_{Y2}|^2 = OC$. But this Δ_{Y1} and Δ_{Y2} cannot be chosen entirely independently, and then the integral in (76) must not be split in two.

two pairs of functions can have no finite poles in common. Then every finite pole of Y_{M1} or Y_{M2} in (78) must belong to one or the other of the following classes:

a. The lhp zeros of $N_1N_2 + (N_1 + N_2)S$;

b. For Y_{M1}, any lhp poles of N_2 or S which are not poles of $N_1N_2 + (N_1 + N_2)S$; for Y_{M2}, any lhp poles of N_1 or S which are not poles of $N_1N_2 + (N_1 + N_2)S$. (81)

The second class occurs only in special or degenerate cases, such as when zeros of $N_1 + N_2$ happen to coincide with poles of S.

After the permissible poles have been determined by (81), (78) may be used to find the numerators of Y_{M1} and Y_{M2}. Simultaneous linear equations in the numerator coefficients may be derived from required behavior at lhp poles of S, N_1, N_2 and also at poles of Y_{M1} and Y_{M2} themselves. For the latter, take the difference of (78c) and (78d) to get:

$$(Y_{M1}N_1 - Y_{M2}N_2) \text{ is rlhp.} \tag{82}$$

Numerators of the minimum degree are determined uniquely, and also turn out to be just compatible with (78*e*). In special or degenerate cases, certain of the poles, permitted by (81), may coincide with zeros of the corresponding numerators of Y_{M1} and Y_{M2} and can be cancelled out.

3.4.2 *Optimum Physical Networks When* $\alpha < 0$

When $\alpha = -\beta$, with $\beta > 0$, Sections 3.2.2 and 3.4.1 may be combined to get

$$\begin{aligned} Y_{M1} &= A_1 + \frac{SN_2e^{-\beta p}}{N_1N_2 + (N_1 + N_2)S}, \\ Y_{M2} &= A_2 + \frac{SN_1e^{-\beta p}}{N_1N_2 + (N_1 + N_2)S}. \end{aligned} \tag{83}$$

Here, A_1 and A_2 are two different rational functions, not necessarily rrhp (even though adding the exponential terms must make Y_{M1}, Y_{M2} rrhp). Substituting in (78) and using a generalization of (32) gives*

* Paralleling Section 3.1 gives, in place of (32):

$$\frac{N_1N_2S}{N_1N_2 + (N_1 + N_2)S} = O\omega^{-2} \text{ as } \omega \to \infty$$

$$a.\ A_1 + \frac{SN_2e^{-\beta p}}{N_1N_2 + (N_1 + N_2)S} \text{ is rrhp,}$$

$$b.\ A_2 + \frac{SN_1e^{-\beta p}}{N_1N_2 + (N_1 + N_2)S} \text{ is rrhp,}$$

$$c.\ A_1(N_1 + S) + A_2S \quad \text{is rlhp,} \tag{84}$$

$$d.\ A_1S + A_2(N_2 + S) \quad \text{is rlhp,}$$

$e.$ When ω is real and $\to \infty$, the following three functions $= O\omega^{-2}$:

$$|A_1|^2N_1, \qquad |A_2|^2N_2, \qquad |A_1 + A_2|^2S.$$

In (84), conditions a and b put the rhp poles of A_1 and A_2 at the rhp poles of the functions which multiply $e^{-\beta p}$. Conditions c and d permit the same lhp poles as the poles of Y_{M1} and Y_{M2} for $\alpha > 0$, which are determined by (81). Combining the lhp and rhp conditions shows that every finite pole of A_1 or A_2 must belong to one or the other of the following classes:

$a.$ The zeros of $N_1N_2 + (N_1 + N_2)S$;

$b.$ For Y_{M1}, any poles of N_2 or S which are not poles of $N_1N_2 + (N_1 + N_2)S$; for Y_{M2}, any poles of N_1 or S which are not poles of $N_1N_2 + (N_1 + N_2)S$. (85)

These are merely the extensions of the classes (81), to include poles in both half planes.

The poles of A_1 and A_2 are twice as numerous as the poles of Y_{M1} and Y_{M2} for $\alpha > 0$. As a result there are more numerator coefficients to be determined, but they are still determined by simultaneous *linear* equations.

3.4.3 *A Time-Domain Interpretation*

Paralleling Section 3.2.3 gives a time-domain interpretation. Although it does not represent a significant simplification of the present problem, it does have theoretical interest, and also a potential usefulness in variations of the present problem.

In Section 3.2.3, we multiplied the rlhp function in (35b) by the rlhp function $1/\tilde{Y}_s$, so as to obtain the rlhp function (44), in which the rrhp term Y_MY_F may be regarded as the only unknown. We now multiply the two rlhp functions in (78) by two different rlhp functions and add

the result. The object is to obtain an rlhp function with a single unknown term which is itself rrhp.

Let Q_1 and Q_2 be two (as yet unknown) functions of p, both required to be rrhp. Then $\tilde{Q}_1$ and $\tilde{Q}_2$ are both rlhp. If we multiply the two rlhp functions in (78c) and (78d) by $\tilde{Q}_1$ and $\tilde{Q}_2$ and add the result the sum will have to be rlhp. The sum can be written as follows:

$$\begin{aligned}(Y_{M1}Y_a + Y_{M2}Y_b) - e^{\alpha p}S(\tilde{Q}_1 + \tilde{Q}_2) &\text{ is rlhp,}\\ Y_a = \tilde{Q}_1(N_1 + S) + \tilde{Q}_2 S,&\\ Y_b = \tilde{Q}_1 S + \tilde{Q}_2(N_2 + S).&\end{aligned} \tag{86}$$

If Y_a and Y_b are both rrhp, as well as Y_{M1} and Y_{M2}, the function $(Y_{M1}Y_a + Y_{M2}Y_b)$ will be rrhp and can be found from (86) by the Bode-Shannon method. This function can be used to express (82) in terms of Y_{M1} or Y_{M2} alone. Thus, Y_{M1} and Y_{M2} may be found by a sequence of straightforward calculations, as soon as $\tilde{Q}_1$ and $\tilde{Q}_2$ are known. The problem is to find a $\tilde{Q}_1$ and a $\tilde{Q}_2$ such that Y_a and Y_b are, in fact, rrhp. It may be more informative to use the equivalent requirement that $\tilde{Y}_a$ and $\tilde{Y}_b$ must be rlhp. Then (since N_1, N_2, S are even), (86) requires

$$\begin{aligned}&a.\ \tilde{Y}_a = Q_1(N_1 + S) + Q_2 S \text{ is rlhp,}\\ &b.\ \tilde{Y}_b = Q_1 S + Q_2(N_2 + S) \text{ is rlhp,}\\ &c.\ Q_1, Q_2 \qquad\qquad \text{are rrhp.}\end{aligned} \tag{87}$$

Equations (87) are about as hard to solve for Q_1 and Q_2 as (78) is for Y_{M1} and Y_{M2}. The poles of Q_1 and Q_2 are exactly the poles of Y_{M1} and Y_{M2}. The numerators are different and are not uniquely determined.

Note that the calculation of Q_1 and Q_2 does not involve the terms $e^{\alpha p}S$ of (78). These enter only in the Bode-Shannon type of analysis. The method has potential usefulness in variations of the present problem, in which the terms $e^{\alpha p}S$ are replaced by more complicated functions. (Compare this section with Section 3.3.4).

3.5 *Sampled-Data Systems*

This section shows how the methods which we applied in the previous sections to continuous-data systems may be modified for application to discrete, or sampled-data systems. Methods appropriate for discrete systems can, of course, be derived without reference to continuous-data systems; and they have been so derived by, for example, Levinson (Ref.

1, Appendix B), and Lloyd and McMillan.[10] It is interesting to observe, however, how simply the transformation may be accomplished, from techniques for continuous-data systems to techniques for discrete-data systems. For this purpose, it will be sufficient to outline an appropriate procedure, without filling in details.

Two different kinds of discrete-data systems may be considered. The signal and noise may be discrete time series, described by statistics appropriate for such series. On the other hand, the signal, $s(t)$, and the noise, $n(t)$, may themselves be continuous time series, with the observations of signal-plus-noise, $f(t)$, limited to discrete "sampling" times. Methods appropriate for both kinds of discrete-data systems may be derived from a study of continuous-data systems of a special kind.

3.5.1 *A Special Kind of Continuous-Data System*

Following the usual methods of "z transform" theory, let $z = e^{-Tp}$, in which T is to be the sampling interval. Rational functions of z, when treated as functions p, now have inverse transforms which are zero except at the discrete times σT. The methods which we have applied to continuous-data systems do not apply directly to spectra, S and N, which are simply rational functions of z. The behavior at $p = \infty$ leads to divergent integrals such as, for example, the integral in (31). On the other hand, spectra of the sort described below meet the convergence conditions and also lead to $Y_M(p)$ which are exactly rational functions of z. Then the only data which are actually utilized are the data observed at the discrete times $t - \sigma T$.

Note that $\tilde{z} = 1/z$, and that $|z| > 1$ in the left half of the p plane. Use the notation $Y(p) = Y_z(z)$ and $\tilde{Y}_z(z) = Y_z(1/z)$. In these terms, N and S are to be

$$\begin{aligned} &a.\ S = \frac{q}{1 + \epsilon^2\omega^2} Y_S \tilde{Y}_S\,, \\ &b.\ N = \frac{q}{1 + \epsilon^2\omega^2} Y_N \tilde{Y}_N\,, \\ &c.\ Y_{Sz}(z),\ Y_{Nz}(z) \text{ are rational in } z, \\ &d.\ Y_{Sz}\,,\ Y_{Nz}\,,\ 1/Y_{Sz}\,,\ 1/Y_{Nz} \text{ are regular at } |z| < 1. \end{aligned} \tag{88}$$

The constants q and ϵ are to be real and positive.

Suppose $\alpha > 0$, in $s(t + \alpha)$, and Y_M is to be the optimum "physical" Y_G, when N and S are described by (88). The procedure explained in Section 3.2 is easily adapted to the new problem. Corresponding to

(35), one gets:

$$\begin{aligned} &a.\ Y_M(p) && \text{is rrhp,} \\ &b.\ \frac{q}{1-\epsilon^2 p^2}\, Y_\Sigma(p) && \text{is rlhp} \\ &\qquad Y_{\Sigma z} = Y_{Mz}(Y_{Sz}\tilde{Y}_{Sz} + Y_{Nz}\tilde{Y}_{Nz}) - \frac{Y_{Sz}\tilde{Y}_{Sz}}{z^{\alpha/T}}, \\ &c.\ |\, Y_{\Sigma z}\,| < \infty \text{ at } z = \infty. \end{aligned} \tag{89}$$

These conditions are satisfied by a $Y_M(p)$ such that $Y_{Mz}(z)$ is rational and meets the following conditions:

$$\begin{aligned} &a.\ Y_{Mz}(z) \text{ is regular} && \text{when } |\, z\,| < 1, \\ &b.\ Y_{\Sigma z}(z) \text{ is regular} && \text{when } |\, z\,| > 1, \\ &c.\ Y_{\Sigma z}(z) = 0 && \text{when } 1 + \epsilon p = 0, \\ &d.\ 0 < |\, Y_{\Sigma z}(z)\,| < \infty && \text{at } z = \infty. \end{aligned} \tag{90}$$

The factor $q/1 + \epsilon^2\omega^2$ was applied to N and S in (88) to obtain convergence of integrals in the derivation of (89) and (90). However, we can make ϵ as small as we wish, and Y_M approaches a reasonable limit as $\epsilon \to 0$. The poles of Y_{Mz} are completely independent of ϵ. The numerator approaches a limit such that condition (90*c*) disappears, and (90*d*) changes to

$$Y_{\Sigma z}(z) = 0 \quad \text{at} \quad z = \infty. \tag{91}$$

This follows from the fact that (90*c*) requires $Y_{\Sigma z}$ to have a numerator factor in z which is zero at the zero of $1 + \epsilon p$ — namely, a factor

$$1 - e^{-(T/\epsilon)}z = 1 - e^{-(T/\epsilon)(1+\epsilon p)}. \tag{92}$$

When ϵ is small, but not zero, the autocovariances corresponding to the N and S of (88) take the form shown in Fig. 13. The "spikes" become sharper as ϵ becomes smaller. As $\epsilon \to 0$, the integrals of the N and S of (88) become infinite unless q also $\to 0$ to order of ϵ. The function Y_M, however, is entirely independent of q, which need be considered quantitatively only in calculating the corresponding variance, σ_M^2.

3.5.2 *Discrete Time Series*

Suppose S and N are the spectra of continuous time series $s(t)$ and $n(t)$, such that the covariances are exactly zero except when the lag

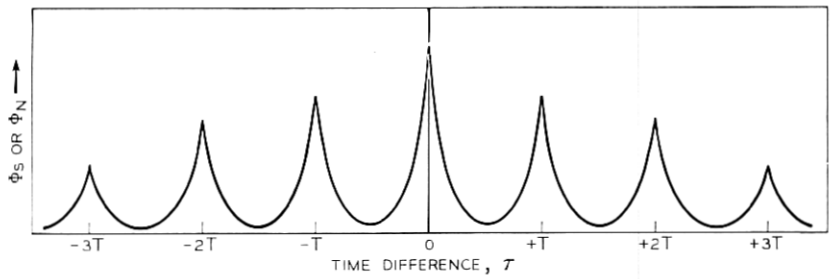


Fig. 13 — A special form of auto-covariance function.

time is σT, with σ an integer. Let the prediction time α also be an integral multiple of T. The corresponding Y_M will be such that the estimate $g(t)$, of $s(t + \alpha)$, will depend only on the values assumed by $s(t)$, and $n(t)$ at the discrete series of times $t - \sigma T$. Since the values of $s(t)$ and $n(t)$ which occur at intermediate times are not utilized, and are not correlated to the values which are utilized, they can have no significant relation to the problem; they may be regarded as undefined without altering the solution. Thus the special continuous-data system yields the same Y_M as though $s(t)$ and $n(t)$ were discrete, ordered sequences of random variables matching the values which are assumed by the continuous time series at the discrete times σT.

Conversely, when s and n are initially discrete series, corresponding continuous ensembles may be constructed for calculating the optimum smoothing and prediction. When the covariances of the discrete series are sums of exponentials in $m_2 - m_1$, where m_1 and m_2 are the order numbers of the samples involved, the corresponding S and N will be rational functions of $z = e^{-Tp}$, and Y_M and K_M may be found by applying Section 3.5.1. Many details, of course, remain to be filled in.

3.5.3 *Discrete Samples of Continuous Time Series*

Now suppose that $s(t)$ and $n(t)$ are continuous time series with rational spectra S and N, but that $f(t) = s(t) + n(t)$ may be observed only at discrete "sampling times", $t = \sigma T$. Let "present time" t coincide with one of the sampling times and consider the calculation of the optimum Y_M for estimating $s(t + \alpha)$, with $\alpha > 0$.

In Section 3.2.4, we replaced the Gaussian statistical model shown in Fig. 6 by that shown in Fig. 8, which has the same pertinent statistical characteristics. Now, the pertinent statistics are even more restricted, and a further change in the Gaussian model is useful. In particular, the

frequency functions Y_F, Y_{FS} and the single random quantity v_s may be modified in any way which leaves the following covariances unchanged (in which σ is a nonnegative integer): the auto-covariance of $f(t)$ at the discrete lag times σT, the cross-covariance of $f(t)$ and $s(t)$ at the discrete lag times $(\alpha + \sigma T)$ and the auto-covariance of $s(t)$ at the lag time zero.

Our present problem may be solved by changing Y_F, Y_{FS}, N_S (without changing the pertinent statistics), in such a way that the optimum $Y_M(p)$, computed on a continuous-data basis, will be the transform of a $K_M(t)$ which is zero except when $t = \sigma T$. The output of a corresponding device, at any sampling time, will depend only on the values of the input at sampling times. A suitable mechanization for the sampled-data system, derived from this Y_M or K_M, may be either a fixed network (or equivalent device) with sampling at input or output, or digital computations carried out once each computing cycle.

When N and S are rational functions of frequency, Y_F and Y_{FS} may be changed to rational functions of z, and Section 3.5.1 may be applied to find Y_M as a rational function of z. When $\alpha = mT$, with m an integer, it is sufficient to change N and S to suitable rational functions of z, preserving auto-covariances of $s(t)$ and $n(t)$ at lag times σT, and then to apply Section 3.5.1. When $\alpha \neq mT$, however, Y_{FS} must be modified further (or the related function $Se^{\alpha p}$). As in the previous section, many details remain to be filled in.

3.6 *Nonstationary Systems*

In this section, we consider signal and noise ensembles which are statistically nonstationary. Some years ago, Booton[21] described integral equations which determine the corresponding linear least-squares smoothing and prediction operators. The general integral equations, however, are very much more difficult to solve than are the equations corresponding to stationary systems. On the other hand, the techniques described above, for stationary systems, may be paralleled, in principle at least, for analogous nonstationary systems. More specifically, a one-to-one correspondence may be established between the individual relationships and operations used in the techniques and a new set of relationships and operations which are appropriate for at least a large class of nonstationary systems.

While it is clear that the new relationships and operations are appropriate for many nonstationary systems, the exact range of their applicability has not been fully established. In addition, the manipulations required in numerical problems are very much more complicated than

the stationary counterparts and they may be feasible only for quite simple systems, even when electronic computers are available. Accordingly, a very brief outline may be sufficient here, leaving a fuller exposition for a later paper.

3.6.1 *A Class of Nonstationary Systems*

In previous sections, we considered finite networks of lumped elements, driven by white noise, as physical models which generate stationary Gaussian ensembles with rational spectra (Figs. 2, 3, 6 of Section II and Fig. 8 of Section 3.2.4). We now change the picture only by permitting the network components to be time variable; we thereby define a class of nonstationary Gaussian ensembles, analogous to the stationary ensembles which have rational spectra. Miller and Zadeh[22] have studied essentially the same class of nonstationary ensembles, in somewhat different terms.

Let V be the input and E the output of a finite network of lumped, linear, time-variable elements. Then

$$\begin{aligned} PV &= QE, \\ P &= a_0 + a_1 \frac{d}{dt} + \cdots + \frac{d^n}{dt^n}, \\ Q &= H\left(b_0 + b_1 \frac{d}{dt} + \cdots + \frac{d^m}{dt^m}\right). \end{aligned} \tag{93}$$

The coefficients a_σ, b_σ, H are now generally time-variable.

Consider the functions U_σ, defined by $PU_\sigma = 0$, and the functions L_σ, defined by $QL_\sigma = 0$. There will always exist n linearly independent U_σ's and m linearly independent L_σ's. The U_σ's may be referred to as the basis functions, or bf's, and the L_σ's as the zero-response functions, or zrf's. In the theory of nonstationary systems, the bf's correspond to the poles of the admittances of stationary systems, and the zrf's correspond to the zeros. In general, the bf's and the zrf's may be chosen in any of various ways, since linear transformations of legitimate choices are also legitimate choices. Frequently, however, a specific choice is particularly useful, as in the situation described below.

Existence theorems are more easily established if the coefficients a_σ, b_σ, H in P and Q, when viewed as functions of time t, are required to be analytic at $t = -\infty$. Then, except in degenerate cases which need not be considered here, the bf's, U_σ, and the zrf's, L_σ, may be so chosen that

they behave as follows, in accordance with Bellman:[23]

$$\begin{aligned} &a.\ \text{When } a_\sigma,\ b_\sigma,\ H \text{ are analytic at } t \to -\infty, \\ &b.\ \text{Then } U_\sigma \to t^{\gamma_\sigma'} e^{p_\sigma' t} \text{ and } L_\sigma \to t^{\gamma_\sigma''} e^{p_\sigma'' t}. \end{aligned} \tag{94}$$

The coefficients γ_σ and p_σ are *constants*. The p_σ's may be described further in terms of P and Q of (93). Let $P_\infty(p)$ and $Q_\infty(p)$ be the polynomials derived from P and Q in the following way: Replace d/dt by p. For coefficients, use the values assumed by the a_σ's and b_σ's at $t = -\infty$. Then the p_σ''s are the zeros of $P_\infty(p)$ and the p_σ'''s are the zeros of $Q_\infty(p)$. The same sort of conditions for "physical" networks may now be applied to the p_σ''s and p_σ'''s as to the poles and zeros of stationary admittances. (The coefficients must also behave "reasonably" in some sense, although not necessarily analytically, at times other than $t \to -\infty$.)

When the p_σ''s have negative real parts, as required for "physical" networks, the integration of (93) gives (for $m < n$):

$$\begin{aligned} &a.\ V(t) = \int_{-\infty}^{+\infty} K(t, \tau) E(\tau)\, d\tau, \\ &b.\ K(t, \tau) = \sum_{\sigma=1}^{n} U_\sigma(t) X_\sigma(\tau) \quad \text{when} \quad \tau < t, \\ &\qquad\quad = 0 \qquad\qquad\qquad\quad \text{when } \tau > t. \end{aligned} \tag{95}$$

Here, K is the impulse response, as in previous sections, but it is no longer a function of the single variable $t - \tau$. The U_σ's are again the bf's of the differential equation (93) and the X_σ's are new functions. Their calculation is at least straightforward, provided the U_σ's are known, as well as the coefficients of Q in the differential equation. The X_σ's are roughly analogous to the residues at the poles of the transfer admittance of a stationary network. (The functions X_σ/U_σ are a closer analog.)

3.6.2 *Manipulations of Differential Equations*

When a time-variable network is made up of networks in tandem or networks in parallel the differential equation for the complete network may be found from the differential equations which correspond to the different parts. The processes are analogous to, but not the same as, the algebraic products and sums which are used to combine the rational admittance functions of stationary partial networks. The differences may be said to stem from the noncommutability of time-variable coefficients and the derivative operator, d/dt.

$$\left(C \frac{d}{dt} V \neq \frac{d}{dt} CV, \text{ when } C \text{ is time-variable.} \right)$$

Conversely, differential equations such as (93) may be decomposed into sets of simpler differential equations, corresponding to partial networks connected either in tandem or parallel fashion. For this, however, one needs to know the bf's of the differential equation (and also the zrf's, for tandem circuit decompositions), just as one needs to know the poles (and also the zeros, for tandem circuit decompositions) when decomposing the admittances of stationary circuits. The operations are of course analogous to, but not the same as, the factoring and partial fraction expansion of admittance functions.

Manipulations of the sort noted above are described in more detail in a previous paper by the author.[24]

3.6.3 *Auto-Covariances*

When the input signal V is (unit level) white noise, the auto-covariance, Φ, of the output signal, E, may be calculated from the impulse response, K. When K is as in (95b), Φ may be expressed as a somewhat similar finite sum. More specifically,

$$\begin{aligned} a.\ \Phi(t_2\,,\,t_1) &= \int_{-\infty}^{+\infty} K(t_2\,,\,\tau)K(t_1\,,\,\tau)\,d\tau, \\ b.\ \Phi(t_2\,,\,t_1) &= \sum_{\sigma=1}^{n} U_\sigma(t_2)W_\sigma(t_1) \quad \text{when} \quad t_1 < t_2 \\ &= \sum_{\sigma=1}^{n} W_\sigma(t_2)U_\sigma(t_1) \quad \text{when} \quad t_1 > t_2\,. \end{aligned} \tag{96}$$

The U_σ's are again the basis functions or bf's of the differential equation (93). The W_σ's are new functions, which may be calculated in any of various ways.

A differential equation, corresponding to Φ, may be defined as the equation connecting two time functions, say G_1 and G_2, in such a way that

$$G_2\,(t_2) = \int_{-\infty}^{+\infty} \Phi\,(t_2,\,t_1)\,G_1\,(t_1)\,dt_1\,.$$

The differential equation is of order $2n$. Its bf's comprise both the U_σ's and W_σ's of (96), but its zrf's can be found only by solving a homogeneous differential equation with time-variable coefficients.

In (96a), $K(t_1\,,\,\tau)$ may be replaced by $K^a(\tau,\,t_1)$, where K^a is the "adjoint" of K, defined as the function obtained from $K(t,\,\tau)$ by interchanging, within the function, the input time, τ, and the output time, t. The integral of the product $K(t_2\,,\,\tau)K^a(\tau,\,t_1)$ is a convolution, represent-

ing the impulse response of the tandem combination of two networks. The two networks are the original network preceded by a (nonphysical) network with the adjoint response, as in Fig. 14.

We may now relate G_1 and G_2 through an intermediate variable, say G_m, by means of

$$\begin{aligned} a.\ & P^a G_m = Q^a G_1 , \\ b.\ & P G_2 = Q G_m . \end{aligned} \tag{97}$$

Here, P and Q are as in (93), and P^a and Q^a are similar operators, which lead to the adjoint $K^a(t, \tau)$ of the impulse response $K(t, \tau)$.

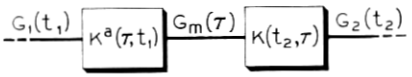
When systems are stationary, the transform of Φ is the product $Y(p)\tilde{Y}(p)$. When systems are nonstationary, the bf's and zrf's of (97*a*) are analogous to the poles and zeros of $\tilde{Y}(p)$, just as the bf's and zrf's of (97*b*) are analogous to the poles and zeros of $Y(p)$. The simple product of Y and $\tilde{Y}$ is replaced by the construction of a single differential equation, from the two equations of (97), through the elimination of the intermediate variable G_m. One half of the bf's of Φ are exactly the bf's of K, and one half of the zrf's of Φ are the zrf's of K^a. The other halves of both the bf's and zrf's of Φ are *not* exactly the bf's of K^a and the zrf's of K, except when systems are stationary.

The function $\Phi(t_2 , t_1)$ is symmetrical in t_2 and t_1. As a result, Φ is its own adjoint. The self-adjoint property of $\Phi(t_2 , t_1)$ corresponds to the evenness of the frequency function $Y(p)\tilde{Y}(p)$, which was noted in the analysis of stationary systems.

When (uncorrelated) signal and noise ensembles each have auto-covariances of the general form (96*b*), the auto-covariance of the signal-plus-noise is $\Phi_S + \Phi_N$, and it has the same general form. The U_σ's and W_σ's of Φ_F simply comprise all the U_σ's and W_σ's of Φ_S and Φ_N.

3.6.4. *An Analog of the Bode-Shannon Method*

Now consider the calculation of linear least-squares smoothing and prediction operators when auto-covariances, Φ_S and Φ_N, of signal and



$$G_m(\tau) = \int_{-\infty}^{+\infty} K^a(\tau, t_1) G_1(t_1)\, dt_1 \qquad G_2(t_2) = \int_{-\infty}^{+\infty} K(t_2, \tau) G_m(\tau)\, d\tau$$

Fig. 14 — A (nonphysical) network whose impulse response is $\Phi(t_2 , t_1)$.

noise are known, have the form (96*b*) and satisfy condition (94a). The method described below is analogous to the method of Bode and Shannon, as interpreted in Section 3.2.4.

We begin with the physical model illustrated in Fig. 15, which is exactly like Fig. 8, of Section 3.2.4, except that the (now nonstationary) impulse responses of the networks have the form $K(t, \tau)$, instead of $K(t - \tau)$ [the inverse transform of $Y(p)$]. The method described in Section 3.2.4 can be adapted to the model of Fig. 15, in principle at least, provided an impulse response $K_F(t, \tau)$ of the form (96*b*) does, in fact, exist which has the following properties: it must turn white noise into an $f(t) = s(t) + n(t)$ with the required auto-covariance, $\Phi_F(t_2, t_1)$; it must be "physical"; there must exist a "physical" impulse response, $K_F^{-1}(t, \tau)$, which turns $f(t)$ back into the white noise.

The manipulations described in Section 3.6.2 may be used to decompose the differential equation corresponding to Φ_F into a pair of differential equations which are at least superficially like (97*a*) and (97*b*). Under condition (94*a*), a particular decomposition will always have the following properties: the orders of the two equations are the same; the two equations have identical coefficients H of (93); the bf's and zrf's of (97*a*) are all "nonphysical", while those of (97*b*) are all "physical" [as determined by the real parts of the p_σ's of (94*b*)]; the corresponding impulses will be respectively K_F^a and K_F provided they are, in fact, adjoints, each of the other.

While a rigorous proof has not been completed, there is strong evidence that the two impulse responses will, in fact, be adjoints when derived in the manner described from auto-covariances of the form (96*b*) subject to the condition (94*a*). The same probably holds true in many situations where (94*a*) is not satisfied. When (97*a*) and (97*b*) have been determined, K_F^{-1} may also be found, by merely interchanging P and Q in (97*b*) and

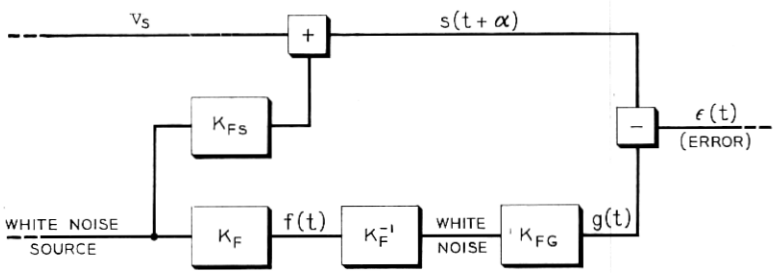


Fig. 15 — A physical model of a nonstationary system.

then calculating the corresponding impulse response. It will be physical because the zrf's of K have been made physical.

For the most part, the calculations described above are straightforward but laborious. The bf's of the differential equation for Φ_F are the basis functions U_σ and W_σ of Φ_S and Φ_N, and are known when Φ_S and Φ_N are known in the form (96*b*). The zrf's of Φ_F, however, must be calculated as solutions of a homogeneous linear differential equation with time-variable coefficients. This is analogous to finding the zeros of the spectrum $N + S$ when given rational spectra N and S. The computational difficulties, however, are very much greater, and they are likely to limit applications to quite simple problems.

3.6.5 *An Analog of the Zero and Pole Method*

Some of the laborious calculations (but not the calculation of zrf's of Φ_F) may be avoided by techniques analogous to the zero and pole techniques described in Section 3.2, etc. For definiteness, let the optimum impulse response, $K_M(t, \tau)$, be chosen from the class of all stationary and nonstationary physical impulse responses (infinite memories included) and let the prediction interval α be greater than 0. Methods similar to those used in Sections 2.6 and 3.2.3 yield a Wiener-Hopf equation of the following sort [a generalization of (45)]:

$$\int_{-\infty}^{t} K_M(t, t_1)\Phi_F(t_2, t_1)\, dt_1 = \Phi_S(t + \alpha, t_2), \quad t_2 < t. \tag{98}$$

When $\Phi_F = \Phi_S + \Phi_N$, and Φ_S, Φ_N have the form (96*b*), (98) becomes:

$$\left.\begin{aligned} &\sum_\sigma U_\sigma(t_2) \int_{-\infty}^{t_2} K_M(t, t_1)W_\sigma(t_1)\, dt_1 \\ &+ \sum_\sigma W_\sigma(t_2) \int_{t_2}^{t} K_M(t, t_1)U_\sigma(t_1)\, dt_1 \end{aligned}\right\} = \sum_\rho W_\rho(t_2)U_\rho(t + \alpha), t_2 < t. \tag{99}$$

The sum $\sum_\rho$ includes those terms in $\sum_\sigma$ which are contributed by Φ_S, but not those from Φ_N.

A differential equation may be derived from (99) which is exactly the differential equation corresponding to Φ_F, with K_M taking the role of the "input" function, $G_1(t_1)$, and the right-hand side of (99) taking the role of the "output" function $G_2(t_2)$. Time t is a constant in the derivation and analysis of this differential equation. Then $G_2(t_2)$ is a sum of some of the bf's of the differential equation. As a result, K_M must be a linear combination of zrf's of Φ_F, except that δ functions may be permitted at the limit of the integration, $t_1 = t$. More exactly, K_M is a linear combination of the

physical zrf's of Φ_F, plus possible δ functions. This is analogous to condition (36*a*) of Section 3.2.1, which makes the poles of Y_M the lhp zeros of $N + S$.

The coefficients of the various terms in the linear combination may be determined by using the linear combination with general coefficients, in place of K_M in (99). Evaluating the integrals yields a linear combination of the U_σ's, which are the physical bf's of Φ_F. Since the combination must be identically zero [when the right-hand side of (99) is included], the net coefficient of each U_σ must be zero. The result is a set of linear equations which determine the coefficients of the various terms in K_M. The equations are analogous to the conditions (36*b*) on the behavior of Y_M at lhp poles of N and S.

IV. FURTHER SPECIFIC PROBLEMS AND APPLICATIONS

The central problems described in Section III may be adjusted to fit various engineering problems. The adjustments, however, may require changes in various details. The examples described in this section illustrate both engineering usefulness and ways in which details may be changed. Some of the examples represent existing engineering applications. Some are merely potentially useful. Others are of interest primarily for theoretical reasons. The specific changes in the central problems are reviewed in more general terms in Section 4.6.

4.1 *Problems Related to Anti-Aircraft Fire Control*

The correct aiming of anti-aircraft artillery depends, fundamentally, on data smoothing and prediction. It also illustrates several of the ways in which the central problems may be modified to meet practical requirements. The anti-aircraft problem, as such, will not be developed in more detail than is needed for purposes of illustration.

An anti-aircraft projectile must be aimed at a predicted future position of the target, for which the prediction is based on positions of the target observed at present and past times. The observations (optical or radar) are contaminated by fluctuating observational errors, corresponding to some statistical ensemble. The true motion of the target also corresponds to a statistical ensemble, for the direction of flight and the speed will change with time, in ways which are not entirely predictable. Then the observational errors correspond to our noise, $n(t)$, and the true position of the target corresponds to our true signal, $s(t)$. While the problem is really three-dimensional, it will be sufficient for our purposes to consider a single one-dimensional component.

It is generally reasonable to use a Gaussian model for the observational errors. The observations, however, are likely to be available only for a limited interval. Thus the finite-memory form of the general problem should be assumed.

The true-signal (target) statistics are *not* well represented by a Gaussian model. Furthermore, the average square error is not a reasonable criterion of accuracy, without a careful interpretation. (The average square error gives most weight to large errors, while the "kill probability" depends on the frequency of small errors.) A somewhat nonoptimal solution is generally accepted, as described below.

Most anti-aircraft fire control systems are designed around the following assumption:

In the absence of observational errors, the future position of a nonaccelerating target is to be predicted perfectly. (100)

A nonaccelerating target flies a straight-line, constant speed course. Under (100), the actual errors in prediction will depend on the actual observational errors and on the actual accelerations of the target during the combined observation and prediction intervals ($t - T$ to $t + \alpha$).

The future position of a nonaccelerating target may be determined from its present position and present velocity. Hence, physical linear operators *can* satisfy (100). Generally, the truly optimum linear operator cannot be very different. Anti-aircraft systems must have a high *percentage* accuracy. That is, position errors must be small compared with typical distances from tracker to target or from present target position to predicted future position. Furthermore, the signal ensemble may generally be regarded as centered about the nonaccelerating target courses. Therefore, the truly optimum linear operator must sum the observations in a way which comes very close to giving perfect prediction under the special condition stated in (100). Then small changes will make these particular predictions perfect.

Let x be any one component of position. Then

$$x(t + \alpha) = x(t) + \alpha\bar{\dot{x}}\,. \tag{101}$$

Here, $\bar{\dot{x}}$ is the average rate of change of x, averaged over the prediction interval, t to $t + \alpha$. Under (100), separately optimum estimates of $x(t)$ and $\bar{\dot{x}}$, used in (101), give the optimum estimate of $x(t + \alpha)$. (A proof is not needed for our present purposes.) In anti-aircraft systems, errors in $x(t)$ are usually less significant than errors in $\alpha\bar{\dot{x}}$, and they are generally less subject to reduction by data smoothing. Then attention is centered on the optimum estimation of $\bar{\dot{x}}$.

The optimum estimation of $\bar{\dot{x}}$ is described in Sections 4.1.1 through 4.1.4. In these sections, our signal, $s(t)$, becomes the true volocity $\dot{x}(t)$. Instead of predicting $\dot{x}(t + \alpha)$, we are to predict $\bar{\dot{x}}$, which is the average of $\dot{x}(t)$ over the interval t to $t + \alpha$ ($\bar{\dot{x}}$ is a functional of $\dot{x}$). The prediction is to be obtained by linear operations applied to observed positions $x(t)$. Since a factor p in a frequency function corresponds to differentiation, we may write

$$g(t) = \bar{\dot{x}} + \epsilon(t) = x(t) \text{ modified by a linear operator } Y_G(p)p. \quad (102)$$

Here, $Y_G(p)$ represents the data smoothing applied to the apparent rate of change, $f(t)$ = observed $\dot{x}(t)$. The error part of $f(t)$ is described by the spectrum

$$N = Y_N \tilde{Y}_N = \omega^2 N_x, \quad (103)$$

where N_x is the error spectrum for $x(t)$ itself.

4.1.1 *Optimum Measurement of a Constant Velocity**

In this section we assume the following conditions:

$$\begin{aligned} &a. \text{ Positions } x \text{ are observed from } t - T \text{ to } t, \\ &b. \text{ Conditions (100) are to be satisfied,} \\ &c. \text{ The true } \dot{x}(t) \text{ is constant from } t - T \text{ to } t + \alpha. \end{aligned} \quad (104)$$

When condition (100) is satisfied and the actual $\dot{x}(t)$ is constant, the entire error $\epsilon(t)$ must be due to errors in observation. Then, by (103) and the noise part of (19) or of Fig. 2,

$$\sigma^2 = \int_{-\infty}^{+\infty} |Y_G|^2 N_x \omega^2 \, d\omega. \quad (105)$$

When $\dot{x}(t)$ is constant, present $\dot{x}(t) = \bar{\dot{x}}$, and prediction time α need not appear at all. Then (104b) requires that $Y_G(p)$ applied to a constant must yield the same constant. Since the response to a constant signal C is $CY_G(0)$, we now have

$$C_Y \text{ is the subclass of the class (47) such that } Y_G(0) = 1. \quad (106)$$

If $Y_{Gj}(0) = Y_{Gk}(0) = 1$, the difference $\Delta_Y(0) = 0$. Then

$$C_\Delta \text{ is the subclass of the class (48) such that } \Delta_Y(0) = 0. \quad (107)$$

Note that (107) is consistent with (29).

* An early treatment of this problem, yielding solutions of special cases is included in a wartime report by Phillips and Weiss.[25]

The optimum operator Y_M is now the Y_G of the class (106) which minimizes σ^2 in (105). As in Section 3.3, use

$$\begin{aligned} &Y_M = A + Be^{-Tp}, \\ &Y_M,\ \tilde{Y}_M e^{-Tp} \text{ are rrhp.} \end{aligned} \tag{108}$$

Equation (51) may now be adapted to the present problem as follows: Omit terms proportional to S. Replace N by $N_x\omega^2$, or, in terms of $p^2 = -\omega^2$, by $-N_x p^2$. To take account of (107), rearrange the integrands, so that $\tilde{\Delta}_Y/p$ becomes the variation factor rather than $\tilde{\Delta}_Y$ [(107) excludes poles of $\tilde{\Delta}_Y/p$ at $p = 0$]. This changes (51) into

$$\int_{-i\infty}^{+i\infty} (AN_x p^2 p)\frac{\tilde{\Delta}Y}{p}\,dp + \int_{-\infty}^{+\infty} (BN_x p^2 p)e^{-Tp}\frac{\tilde{\Delta}Y}{p}\,dp = 0. \tag{109}$$

Section 3.3 may now be paralleled further, to assemble conditions corresponding to (52). Since there are no terms in $e^{\alpha p}$ in (109), a rational N_x leads to rational A and B. Then (52a) and (52b) may be replaced by (49). The resulting list of conditions is as follows:

$$\begin{aligned} &a.\ A + Be^{-Tp} = 1 \text{ at } p = 0, \\ &b.\ A + Be^{-Tp} \text{ is rfpp}, \\ &c.\ AN_x p^3 \text{ is rlhp}, \\ &d.\ BN_x p^3 \text{ is rrhp}, \\ &e.\ \text{When } \omega \text{ is real and} \to \infty, \\ &|A|^2 N_x\omega^2 \text{ and } |B|^2 N_x\omega^2 = O\omega^{-2}. \end{aligned} \tag{110}$$

When N_x is rational, these conditions are just sufficient to determine a rational A and B.

As an example, suppose N_x is*

$$N_x = \frac{1}{\pi}\frac{\sigma_x^2\omega_0}{\omega_0^2 + \omega^2} = \frac{1}{\pi}\frac{\sigma_x^2\omega_0}{\omega_0^2 - p^2}. \tag{111}$$

Then (110c), (110d) and (110e) restrict A and B to

$$A = \frac{a_1 + a_2 p + a_3 p^2}{p^3}, \qquad B = \frac{b_1 + b_2 p + b_3 p^2}{p^3}. \tag{112}$$

* The scale factor has been designated in such a way that σ_x^2 is in fact the average squared position error which is related to N_x by

$$\sigma_x^2 = \int_{-\infty}^{+\infty} N_x\,d\omega$$

Given (112), one can use (110*a*) through (110*d*) to find a_1, a_2, a_3, b_1, b_2, b_3. The resulting Y_M may be arranged as follows:

$$Y_M = -12J\left[\frac{\left(1+\frac{p}{\omega_0}\right)(1-cp)}{T^3p^3} - \frac{\left(1-\frac{p}{\omega_0}\right)(1+cp)}{T^3p^3}e^{-Tp}\right], \quad (113)$$

in which the constants J, c are related to T, ω_0, by

$$c = \frac{T}{2}\left(1+\frac{2}{T\omega_0}\right),$$
$$J = \frac{1}{1+\frac{6}{T\omega_0}+\frac{12}{T^2\omega_0^2}}. \quad (114)$$

The corresponding impulse response $K_M(t)$ may be arranged as follows:

$$K_M(t) = JK_1(t) + (1-J)K_2(t),$$
$$K_1(t) = \frac{6}{T^3}t(T-t) \quad \text{when} \quad 0 < t < T,$$
$$K_2(t) = \frac{1}{T} \quad \text{when} \quad 0 < t < T, \quad (115)$$
$$K_1(t) = 0,\ K_2(t) = 0 \quad \text{when} \quad t < 0 \text{ or } > T.$$

The two functions K_1 and K_2 are the unit-area parabola and unit-area step shown in Figs. 16(a) and 16(b). Then K_M is the combination shown in Fig. 16(c).

The minimum σ^2 determined by (105) may be found by evaluating

$$\sigma^2 = \int_{-\infty}^{+\infty}(A + Be^{-Ti\omega})(\tilde{A} + \tilde{B}e^{Ti\omega})N_x\omega^2\,d\omega. \quad (116)$$

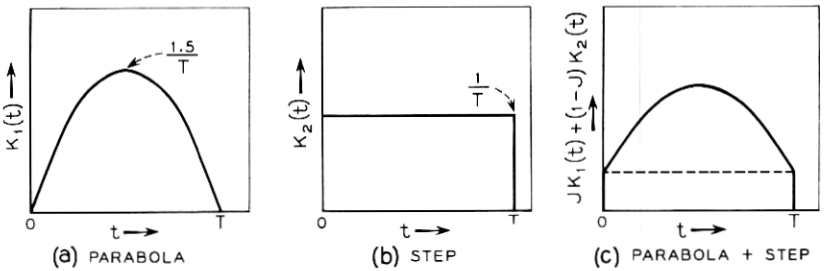


Fig. 16 — Parabolic smoothing.

The evaluation is complicated by the presence of the two exponentials in the integrand, and the multiple poles of A and B at $p = 0$, as determined by (112). The complications may be resolved by splitting the integral into the following two integrals:

$$\sigma^2 = -\int_{-i\infty}^{+i\infty} \left(A + Be^{-Tp} - \frac{1}{1+cp}\right) \tilde{A}N_x p^2\, dp - \int_{-i\infty}^{+i\infty} \left[\frac{\tilde{A}N_x p^2}{1+cp} + (Ae^{Tp} + B)\tilde{B}N_x p^2\right] dp. \tag{117}$$

Each integrand $= O\omega^{-2}$ when ω is real and $\to \infty$; it is regular at $p = 0$ [by (110*a*) and (110*b*)]; and it involves only a single exponential. The first integrand is rrhp; the second is rlhp except for a single pole of $\tilde{A}N_x p^2/(1 + cp)$ at $p = -\omega_0$. Closing the contours of integration by suitable arcs at ∞, and applying the residue theorem gives

$$\sigma^2 = \frac{24J}{T^3}\frac{\sigma_x^2}{\omega_0}. \tag{118}$$

4.1.2 *A White Noise Approximation*

The constant J in (113) and (114) is a function of $T\omega_0$. Generally, $T\omega_0$ is quite large. In the limit, as $T\omega_0 \to \infty$, $J \to 1$ and (113), (115) and (118) become

$$Y_M = 12\,\frac{-\left(1 - \frac{Tp}{2}\right) + \left(1 + \frac{Tp}{2}\right)e^{-Tp}}{T^3p^3},$$

$$K_M = \frac{6}{T^3}\,t(T - t), \tag{119}$$

$$\sigma^2 = \frac{24}{T^3}\frac{\sigma_x^2}{\omega_0} = \frac{12\pi}{T^3}\,N_x(0).$$

The parabolic smoothing represented by (119) was derived by Bode[26] in a quite different way.

Note that $N_x(0)$ is the spectral density of the position errors at low frequencies. The result, (119), may be obtained more simply by substituting the constant $N_x(0)$ for $N_x(p)$ in (110).

The nature of the approximation is explained further by Fig. 17. Curve (*a*) represents a more general N_x, including a "spike" at $\omega = 0$ to allow for drift errors, etc. Curve (*b*) indicates the (qualitative) nature of the corresponding function $|Y_M|^2\omega^2$. The variance, σ^2, is the integral of the product of these two functions. Therefore, the value of N_x is unim-

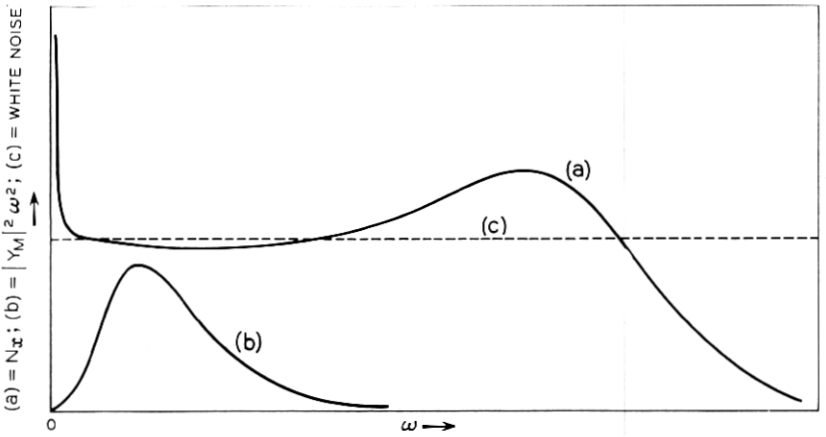


Fig. 17 — Explanation of a white noise approximation.

portant in regions where $| Y_M |^2\omega^2$ is very small. Thus, the white noise spectrum c gives about the same result as the actual spectrum $N_{\dot{x}}$.

4.1.3 *Constant Target Accelerations*

According to (118) and (119) increasing T always decreases σ^2. The σ^2 of (118) and (119), however, is not the total variance $\sigma_{\dot{\bar{x}}}^2$ of the predicted $\dot{\bar{x}}$, except when the target acceleration may be neglected. The effect of a given target acceleration increases as T increases, and this limits the values of T for which (118) and (119) are good approximations to $\sigma_{\dot{\bar{x}}}^2$.

A higher order of approximation takes account of target accelerations, but assumes that they are constant over the interval $t - T$ to $t + \alpha$. If the Y_M of (119) is retained unchanged, $\sigma_{\dot{\bar{x}}}^2$ becomes

$$\sigma_{\dot{\bar{x}}}^2 = \frac{12\pi}{T^3} N_x(0) + \frac{(T + \alpha)^2}{4} \sigma_{\ddot{x}}^2. \tag{120}$$

Here, $\sigma_{\ddot{x}}^2$ is the (ensemble) average squared acceleration of the target.

A smaller $\sigma_{\dot{x}}^2$ may be obtained by modifying the conditions which determine Y_M in a way which takes account of target accelerations. When the expected accelerations are large, it may be reasonable to strengthen (100) to the following:

In the absence of observational errors, the future position of a target with constant acceleration is to be predicted perfectly. (121)

The stronger condition reduces the function classes C_Y and C_Δ to the

subclasses of (47) and (48), such that

$$\begin{aligned} &a.\ Y_G(p) \to 1 + \tfrac{1}{2}\alpha p \quad \text{as} \quad p \to 0, \\ &b.\ \Delta_Y(p) \underset{\to}{\leq} cp^2 \quad \text{as} \quad p \to 0. \end{aligned} \tag{122}$$

These changes in C_Y and C_Δ lead to the following changes in (110): condition (110a) is replaced by (122a); the factor p^3 in (110c) and (110d) is replaced by p^4. The white noise assumption leads to the following, in place of (119):

$$\begin{aligned} K_M(t) &= \frac{6}{T^3}\, t(T-t)\left[1 + \frac{5(T+\alpha)}{T^2}\,(T-2t)\right], \\ \sigma^2 &= \frac{12\pi}{T^3}\left[1 + 15\left(1 + \frac{\alpha}{T}\right)^2\right] N_x(0). \end{aligned} \tag{123}$$

If the expected acceleration does not have a large effect in (120) with the largest T for which observations are available, full compensation may not be justified. (The compensation increases the sensitivity to observational errors unless T can be increased.) A possible, though more complicated procedure is then as follows: Assume that the target acceleration is nonzero but time-invariant, and that the ensemble average of the squared acceleration is $\sigma_{\ddot{x}}^2$. Then minimize the combined effects of the target acceleration and the observational errors.

The combined effects now give

$$\sigma_{\ddot{x}}^2 = \int_{-\infty}^{+\infty} |\, Y_G \,|^2 N_x \omega^2\, d\omega + \sigma_{\ddot{x}}^2 \left[\frac{Y_G - 1}{p} - \frac{\alpha}{2}\right]_{p=0}^2. \tag{124}$$

To keep the second term finite, (106) and (107) are again in order. For a minimum $\sigma_{\ddot{x}}^2$,

$$\int_{-i\infty}^{+i\infty} [\ ]\, dp + \int_{i\infty}^{+i\infty} [\ ]\, dp + 2\left[\frac{\tilde{\Delta}Y}{p}\left(\frac{Y_M - 1}{p} - \frac{\alpha}{2}\right)\right]_{p=0} \sigma_{\ddot{x}}^2 = 0, \tag{125}$$

in which the integrals are the same as in (109). The conditions (110) are now modified to permit simple poles of the two integrands at $p = 0$, which cancel out in their sum. The two contours of integration may be indented at $p = 0$, provided they are indented in the same way. Conditions (110c) and (110d) permit one contour to be closed at infinity around the lhp, and the other around the rhp. Then the residue theorem makes one integral proportional to $\tilde{\Delta}_Y/p$ at $p = 0$, and parameters can be adjusted to cancel the last term in (125).* The result is

* The contour may be indented to pass to either side of $p = 0$, provided the indentation is the same for both integrals. The final results are the same because the poles of the integrands at $p = 0$ cancel out in their sum.

$$K_M = \frac{6}{T^3} t(T - t)\left\{1 + Q\left[\frac{5(T + \alpha)}{T^2}(T - 2t)\right]\right\},$$

$$\sigma^2 = \frac{12\pi}{T^3}\left\{1 + Q\left[15\left(1 + \frac{\alpha}{T}\right)^2\right] N_x(0)\right\}, \tag{126}$$

$$Q = \frac{\sigma_{\ddot{x}}^2}{\sigma_{\ddot{x}}^2 + \frac{720}{T^5} N_x(0)}.$$

4.1.4 *More General Target Accelerations*

Comparing (123) with (120) shows that compensation for constant target acceleration increases the sensitivity to observational errors, unless T is increased. (This is because the observation of acceleration, as well as velocity, is included implicitly in the new Y_M.) Sensitivity to fluctuations in the target acceleration is also increased, and becomes greater as T increases. In principle, Y_M can be modified further, to give perfect prediction in the absence of observational errors, whenever the acceleration is matched perfectly by, say, an nth degree polynomial with random coefficients.* The sensitivity to observational errors will be further increased, unless T is increased. However, any reasonable acceleration ensemble which involves only a finite number of random parameters will lead to a Y_M such that $\sigma^2 \to 0$ as $T \to \infty$.

The infinite memory problem may be handled more reasonably by assigning a *spectrum*, $N_{\ddot{x}}$, to the target accelerations. Then (105) may be replaced by

$$\sigma_{\ddot{x}}^2 = \int_{-\infty}^{+\infty}\left[|Y_G|^2 N_{\dot{x}}\omega^2 + \left|Y_G - \frac{e^{\alpha i\omega} - 1}{\alpha i\omega}\right|^2\left(\frac{N_{\ddot{x}}}{\omega^2}\right)\right] d\omega. \tag{127}$$

If the optimum impulse response $K_M(t)$ corresponding to $Y_M(p)$ is negligible when $t > t_m$, there may be no advantage in extending the interval of observation to times older than $t - t_m$. The practical significance of a quantitative limit may be weakened, however, by the non-Gaussian character of actual target statistics.

4.2 *Measurements with Multiple Instruments*†

This section describes further the two-instrument problem noted in Section 3.4. It will be assumed that $s(t)$ is to be estimated for present

* Blackman[26] has related the corresponding $K_M(t)$ to Legendre polynomials.

† The use of multiple instruments is described in more detail in reports[4, 5, 6] relating to specific applications. Principles are described in papers by Bendat[14] and Stewart and Parks.[15]

time, t, and need not be predicted for a future time, $t + \alpha$. When $\alpha = 0$, (73) becomes

$$\sigma^2 = \int_{-\infty}^{+\infty} (|\, Y_{G1}\,|^2 N_1 + |\, Y_{G2}\,|^2 N_2 + |\, Y_{G1} + Y_{G2} - 1\,|^2 S)\, d\omega. \quad (128)$$

In (72) $f_1(t)$ and $f_2(t)$ are now the results of observing a physical variable, $s(t)$, with two different instruments. Then N_1 and N_2 are the spectra of the instrumental errors $n_1(t)$ and $n_2(t)$.

4.2.1 *Elimination of Errors Proportional to True Signal*

When the two instruments have reasonably high percentage accuracies, $S \gg N_1$ and N_2. Then the Y_{G1} and Y_{G2} which minimize σ^2 in (128) will make $Y_{G1} + Y_{G2} - 1$ very small. Under these conditions, it is reasonable to approximate the true optimum by making the factor exactly zero. Then Y_{G2} is related to Y_{G1} by

$$Y_{G2} = 1 - Y_{G1}. \quad (129)$$

Under (129), the S term in (128) drops out. Using Y_G to represent Y_{G1} and $1 - Y_G$ to represent Y_{G2} gives

$$\sigma^2 = \int_{-\infty}^{+\infty} (|\, Y_G\,|^2 N_1 + |\, 1 - Y_G\,|^2 N_2)\, d\omega. \quad (130)$$

Now, σ^2 is to be a minimum with respect to the single-frequency function Y_G.

Formally, (130) is exactly the same as (19) (with $\alpha = 0$). Rational spectra N and S have merely been replaced by rational spectra N_1 and N_2. Thus the whole of Sections 3.2 and 3.3 may be applied.

4.2.2 *Determination of Position from Position and Velocity Measurements*

As an example, suppose an aircraft's position is measured with one instrument and its velocity with another, and that the two measurements are to be combined to determine the present position to high accuracy. Let N_x and N_v be the error spectra of the position and velocity observations. Then, if all errors are referred to positions, $N_1 = N_x$, $N_2 = N_v/\omega^2$, and (130) becomes

$$\sigma^2 = \int_{-\infty}^{+\infty} \left(|\, Y_G\,|^2 N_x + |\, 1 - Y_G\,|^2 \frac{N_v}{\omega^2} \right) d\omega. \quad (131)$$

In (131), σ^2 will be bounded only if $Y_G(0) = 1$ [assuming that $N_v(0) \neq 0$]. When $Y_G(0) = 1$, applying Y_G to a constant position leaves

the position unchanged. This situation may be interpreted as follows: By (13), any Y_G applied to the true position yields a weighted integral of the true positions existing during the "smoothing interval". When $Y_G(0) = 1$, the weighted integral is a weighted *average*, and the position measurements are used to determine a weighted average position, averaged over the smoothing interval. When the position is not constant, the present position will generally be different from the average, and the difference may be calculated from the velocities observed during the averaging or smoothing interval. The correction may be said to "update" the average. The first term in the integrand in (131) corresponds to errors in the weighted average position, determined from the position measurements, and the second term corresponds to errors in the updating, determined from the velocity measurements.

If the position measurements are used alone and if guessed velocities are, in fact, quite uncertain, an adequate smoothing interval may lead to large updating errors. On the other hand, if the velocity measurements are used alone, they must be integrated over the entire time of flight, and velocity errors may accumulate into large position errors. Thus, the two instruments together may give a very much higher accuracy than is possible with either instrument alone. This may be explained further by citing differences between the spectra N_x and N_v/ω^2, and comparing Y_G and $1 - Y_G$, in (131), to the transfer functions of a pair of separating filters.

4.2.3 *Precalculated Information*

Sometimes, the second measurements may be either replaced by, or augmented by previous nonstatistical information concerning the physical variable. The "biased statistics" may be taken account of as follows: Let $s_0(t)$ be a precalculated "nominal" $s(t)$, which may be regarded as the ensemble average of $s(t)$. Then let

$$\begin{aligned} s(t) &= s_r(t) + s_0(t), \\ f(t) &= f_r(t) + s_0(t), \\ g(t) &= g_r(t) + s_0(t), \\ S &= \text{spectrum of } s_r(t) \text{ ensemble.} \end{aligned} \tag{132}$$

The time series $s_r(t)$ with spectrum S may be regarded as the error in the prediction of $s(t)$ without measurements, by precalculation alone. The time function $g_r(t)$ is to be obtained by applying operator $Y_G(p)$ to

$f_r(t) = f(t) - s_0(t)$. Then $g(t)$ may be found by adding $s_0(t)$ to $g_r(t)$. The error integral (128) is unchanged, provided S is redefined as in (132).

If $Y_{G2} = 0$, in (128), the second measurements are replaced entirely by the precalculation of $s_0(t)$, and the error in the second measurements is replaced by the error in the precalculation. If Y_{G2} is neither 0 nor $1 - Y_{G1}$, the estimate of $s(t)$ is based on three sources of information: the two kinds of measurements and the precalculation of $s_0(t)$. Whether the full generality is justified depends on the relative magnitudes and spectra of the three corresponding errors.

4.3 *A Signal Detection Problem.*

This section describes a simple problem related to signal detection. The time function $f(t) = s(t) + n(t)$ is again observed and $g(t)$ is again produced by applying, to $f(t)$, a *physical* linear operator $Y_G(p)$. Now, however, the true signal $s(t)$ has the following properties: it is a time function which has finite duration and a known shape, but which starts at an unknown time. It may be represented as follows:

$$\begin{aligned} s(t) &= r(t - t_1), \quad t_1 < t < t_1 + w \\ &= 0, \quad t < t_1 \quad \text{or} \quad > t_1 + w \\ t_1 &= \text{a random variable.} \end{aligned} \tag{133}$$

In the absence of noise, the response $g(t)$ will have a maximum value at some value of $(t - t_1)$. The contribution to $g(t)$ from the noise $n(t)$ will have an rms value. We are to find a particular linear operator Y_M which minimizes the ratio of rms noise response, to maximum response to true signal.

Since only ratios are of interest, the scale of Y_M may be chosen to give a unit maximum true response. Then the problem is to minimize the rms noise response within this constraint. Given any valid solution for Y_M producing a maximum true response at, say, $t_1 + t_m$, there will be equally valid solutions producing maximum responses at later times. The operator Y_M can always be multiplied by $e^{-\beta p}$, representing an ideal delay β. Thus, if $t_1 + t_m$ is the time of maximum true response, t_m may be treated as an arbitrary parameter, provided the final results are examined, to determine what values of t_m are, in fact, valid.

When $t = t_1 + t_m$ (133) makes $s(t - \tau)$ become $r(t_m - \tau)$. Then Y_M is the physical Y_G which minimizes the following σ^2, subject to the following constraint:

$$\sigma^2 = \int_{-\infty}^{+\infty} |\, Y_G \,|^2 N \, d\omega,$$
$$\int_{-\infty}^{+\infty} r(t_m - \tau) K_G(\tau) \, d\tau = 1. \tag{134}$$

Let $Y_r(p)$ be the transform of $r(t_m - \tau)$, regarded as a function of τ. Then, by Parseval's equation (T-14), the constraint may be written

$$\int_{-\infty}^{+\infty} Y_r \tilde{Y}_G \, d\omega = 1,$$
$$Y_r(p) = \text{transform of } r(t_m - \tau). \tag{135}$$

The isoperimetric method of the calculus of variations may now be applied in the following way: When Y_G is replaced by $Y_M + \Delta_Y$, (134) and (135) each yield an integral which must vanish. The two integrals may be summed in arbitrary proportion to get

$$\int_{-i\infty}^{+i\infty} (Y_M N - k Y_r) \tilde{\Delta}_Y \, dp = 0, \tag{136}$$

in which k is an initially undetermined constant. The methods of Section 3.2 then give†

$$Y_M Y_N - k \frac{Y_r}{\tilde{Y}_N} \text{ is rlhp.} \tag{137}$$

These can be solved for the physical Y_M, in either frequency-domain or time-domain terms. The resulting Y_M is proportional to k, and k may then be determined by (135).

After further manipulation, the corresponding noise ratio turns out to be

$$\sigma_m^2 = k = \frac{1}{\int_0^{\infty} [K_{rN}(\tau)]^2 \, d\tau},$$
$$K_{rN}(\tau) = \text{inverse transform of } \frac{Y_r(p)}{\tilde{Y}_N(p)}, \tag{138}$$
$$= r(t_m - \tau) * \left(\text{inverse transform of } \frac{1}{\tilde{Y}_N(p)} \right).$$

In the formula for σ^2, the integration runs from 0 to ∞. In general, $K_{rN}(\tau) \neq 0$ when $\tau < 0$, but increasing τ_m shifts $K_{rN}(\tau)$ along the time

† Recall the definition of Y_N by $N = Y_N \tilde{Y}_N$, in Section 2.3.

axis in the direction of $\tau > 0$. Then the corresponding σ_m^2 must decrease. It approaches a minimum value asymptotically, as t_m becomes so large that the "tail" of $[K_{rN}(\tau)]^2$ at $\tau < 0$ becomes negligibly small.

When N is either a constant or the reciprocal of a polynomial $1/\tilde{Y}_N$ is at most a polynomial in p and $K_{rN}(\tau) = 0$ when $\tau > w$, which is the length of the signal $r(t - t_1)$ as defined in (133). Then σ^2 reaches its asymptotic value as soon as $t_m \geqq w$. When N has zeros at finite values of p, σ_m^2 approaches its asymptotic value only when t_m exceeds w by the effective correlation time of the spectrum $1/N$.

When N is a constant, corresponding to white noise, the optimum impulse response is $K_M(\tau) = r(t_m - \tau)$. Then K_M may be described as a mirror image of the given signal form, $r(t - t_1)$, as illustrated in Fig. 18. This is an old principle described, for example, by North.[27] The more general solution, for $N \neq$ constant, has been described by Zadeh and Ragazzini.[28]

A variation of the present problem restricts the class C_Y, of permitted frequency functions, Y_G, to the finite memory class considered in Section 3.3. The problem may be solved by combining the methods of this section and of Section 3.3.4.

When $N \to 0$ as $\omega \to \infty$, it is easy to find pulse shapes such that the values of Section 2.2 and 2.3 regarding behavior at ∞ are violated unless $k = 0$ in (136). The corresponding σ_m^2 does, in fact, $= 0$. An explanation is as follows:

Suppose the $(m - 1)$th derivative of $r(t - t_1)$ is discontinuous and consider a $Y_G(p)$ which approaches cp_m as $p \to \infty$. The corresponding response to $r(t - t_1)$ will include a δ function and its maximum value will be ∞. If, at the same time, the noise spectrum $N = O\omega^{-2(m+1)}$ as $\omega \to \infty$, the rms response to the noise will be bounded and the ratio of rms response to noise to maximum response to true signal will be 0. If $N \to c\omega^{-2m}$, the rms noise itself will be ∞, but it can be shown that the ratio of noise to maximum signal response will still be 0. Compare these

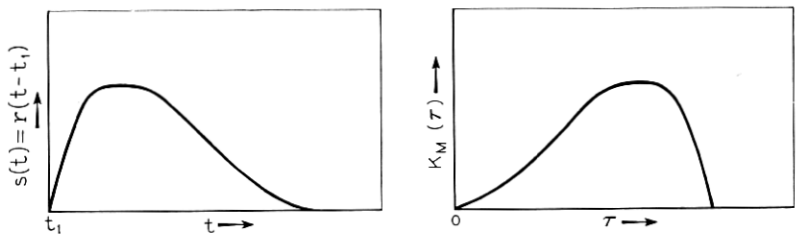


Fig. 18 — Time functions which are mirror images.

conditions with the continuity conditions described in Section 3.3.4. *Actual* pulse shapes and noise spectra are not likely to behave in this way, but what appear to be good approximations may. Care must be used to choose approximations which behave properly as $p \to \infty$.

4.4 *A Principle Relating to Diversity Systems*

This section adapts the method of Section 3.4 to the following problem: A single signal $s(t)$, with known spectrum $S(\omega)$, is observed by m different devices. (These may be, for example, the receivers of a communication system using the diversity principle.) The signal actually observed by device k is $f_k(t) = s(t) + n_k(t)$. The m different $n_k(t)$ are uncorrelated and have known spectra $N_k(\omega)$. Finally, the various noise spectra all have the same shape, but they differ in amplitude. Thus,

$$N_k(\omega) = J_k N(\omega), \qquad k = 1, \cdots, m, \tag{139}$$

where the J_k are known constants and $N(\omega)$ is a known function. The optimum physical filters are to be determined for estimating $s(t + \alpha)$ by summing linear operations on the various $f_k(t)$.

Paralleling (78) gives

$$\begin{aligned}
&a.\ Y_{Mk} \text{ is rrhp,} \\
&b.\ \sum_j Y_{Mk} F_{kj} - e^{\alpha p} S \quad \text{is} \quad \text{rlhp,} \\
&\qquad F_{kk} = S + J_k N, \\
&\qquad F_{kj} = S, \qquad j \neq k, \\
&c.\ \text{When } \omega \text{ is real and} \to \infty, \\
&\qquad |\, Y_{Mk} \,|^2 N \text{ and } |\, \sum_k Y_{Mk} - e^{\alpha p} \,|^2 S = O\omega^{-2}.
\end{aligned} \tag{140}$$

The indices k and j run through the integers 1 to m.

A set of m equations similar to the pair in (79) may now be solved for the Y_{Mk}, in terms of (unknown) rlhp functions U_k. Under the special conditions which we are now assuming, however, a solution is obtained more simply by examining linear combinations of the conditions (140*b*), which give

$$\begin{aligned}
&\left(\frac{N}{\sum_k \frac{1}{J_k}} + S \right) \sum_k Y_{Mk} - e^{\alpha p} S \text{ is rlhp,} \\
&(J_k Y_{Mk} - J_j Y_{Mj}) N \text{ is rlhp.}
\end{aligned} \tag{141}$$

The second of these conditions can be satisfied, within the convergence

condition (140*c*), only if

$$J_k Y_{Mk} = J_j Y_{Mj} . \tag{142}$$

Because of (142), the various $f_k(t)$ may first be summed with weights proportional to $1/J_k$ and then a single frequency-dependent operator may be applied to the sum. In other words, only a single "filtering" device is needed.* This confirms a result which might reasonably be expected without formal analysis. The optimum filter characteristic is proportional to $\sum Y_{Mk}$, and may be found by applying the methods of Section 3.2 to the first condition of (141).

The use of a single filter for all the channels may actually be dictated by cost considerations. The above analysis confirms its use on a performance basis.

4.5 *Nonstatistical Network Synthesis Applications*

Nonstatistical problems in network synthesis sometimes may be formulated in terms of the mathematics of data smoothing, even though no data smoothing is involved. In particular, reasonable solutions sometimes may be found by minimizing integrals similar to those which represent our σ^2. This has been pointed out by Chang,[11] with illustrations in terms of a frequency-domain theory of optimum infinite-memory networks. The possibilities appear to be much greater when a frequency-domain form of the finite-memory problem is available. Possible uses, however, have not been explored in detail. The example described below will illustrate how problems may be formulated.

It will be simplest to develop the example in two stages. We will begin by seeking a physical network function $Y_G(p)$ with the following properties: The "step response" is to have a "rise time" T and is to be exactly 1 thereafter, as in Fig. 19(a). At the same time, $|Y_G|^2$ is to be small at real frequencies above a cutoff frequency, $\omega = \omega_c$, as in Fig. 19(b). More exactly, $|Y_G|^2$ is to be as small as possible, within the rise-time restriction. In order to apply the mathematics of data-smoothing, we will use an average square criterion σ^2 to judge the effective smallness.

The impulse response $K_G(t)$ is the derivative of the step response. It will have to be zero when $t > T$, as in Fig. 19(a). Also, $Y(0)$ is equal to the step response at $t = \infty$, and it will have to be exactly 1:

$$\begin{aligned} K_G(t) &= 0 \quad \text{when} \quad t > T, \\ Y_G(0) &= 1. \end{aligned} \tag{143}$$

* When the N_k are not related as in (139), the Y_{Mk} generally share a common set of poles, but differ in regard to the residues at the poles.

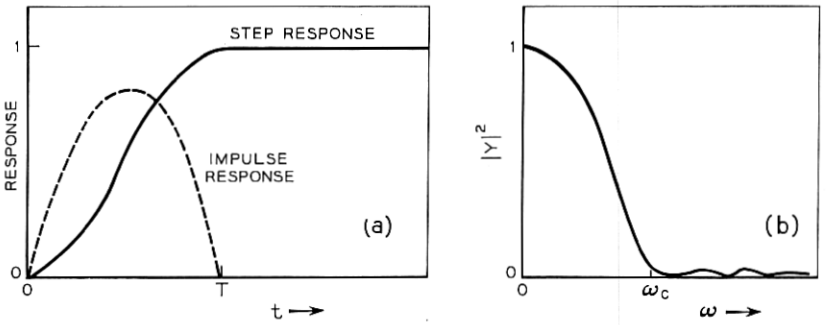


Fig. 19 — Requirements imposed on a network response.

If the average-square criterion σ^2, applied to $|\ Y_G\ |^2$, is unweighted,

$$\sigma^2 = \int_{\omega_c}^{\infty} |\ Y_G\ |^2\, d\omega. \tag{144}$$

The limited range of integration, however, makes the minimization problem very difficult. A more tractable compromise is the following weighted average-square:

$$\sigma^2 = \int_{-\infty}^{+\infty} |\ Y_G\ |^2\, R\, d\omega, \tag{145}$$

where

$$\begin{aligned} R &= \text{a rational function of } \omega, \\ &= \text{an even function} \geqq 0, \\ &= \text{small when } \omega \text{ is small.} \end{aligned}$$

The problem now is as follows: find the particular "physical" function $Y_M(p)$ which makes σ^2 of (145) a minimum, within the constraints (143). Mathematically, this is exactly the problem described in Section 4.1.1, except that the weight factor R replaces spectrum $N_x\omega^2$.

Details of the weight factor R are arbitrary. They influence the distribution of the loss, due to the network, in the high-loss region $\omega > \omega_c$. An efficient choice of R may place zeros and poles on the axis of real ω, with zeros $< \omega_c$ and poles $> \omega_c$. Real zeros and poles must be in identical pairs to preserve $R \geqq 0$. One of each pair is interpreted as in each half plane, and $|\ Y_M\ |^2$ will always have zeros at real ω poles of R. Further analysis indicates the following:

When all zeros and poles of R are at real ω,

$$R = W^2,$$
$$Y_M = \frac{1}{W}(u - \tilde{u}e^{-Tp}). \tag{146}$$

Poles of $1/W$ are canceled in Y_M by zeros of the other factor. The function u may be determined by adapting the method of Section 4.1.1.

The problem may now be modified in the following way: the step response need only approximate 1 when $t > T$. The approximation is to be judged on an average-square basis. There are now two average square criteria, referring respectively to the step response and the frequency suppression. The two criteria are to be added to obtain a measure of over-all performance. Since the step response is the integral of the impulse response

$$\sigma^2 = \int_{-\infty}^{+\infty} |Y_G|^2 R\, d\omega + \int_T^{\infty} \left[\int_0^T K_G(\tau_1)\, d\tau_1 - 1\right]^2 d\tau, \tag{147}$$

the relative importance of the frequency suppression and the step response may be adjusted by adjusting the scale of the weight factor R.

The problem may be solved by splitting Y_G into the following two parts:

$$\begin{aligned} Y_G(p) &= Y_1(p) + pY_2(p), \\ K_1(t) &= 0 \quad \text{when} \quad t > T, \\ Y_1(0) &= 1, \\ K_2(t) &= 0 \quad \text{when} \quad t < T. \end{aligned} \tag{148}$$

The impulse response of Y_2 is the integral of the impulse response of pY_2 (and is therefore equal to the step response of pY_2). Then the constraints on $K_1(\tau)$ and $K_2(\tau)$ are such that

$$\int_T^{\infty} \left[\int_0^T K_G(\tau_1)\, d\tau_1 - 1\right]^2 d\tau = \int_{-\infty}^{\infty} [K_2(\tau)]^2\, d\tau. \tag{149}$$

Applying Parseval's equation (T-15) now gives:

$$\sigma^2 = \int_{-\infty}^{+\infty} [|Y_1 + pY_2|^2 R + |Y_2|^2]\, d\omega. \tag{150}$$

The problem now is as follows: Find the "physical" Y_1 and Y_2 which make σ^2 a minimum, within the constraints (148). The problem may be solved by combining the method of Sections 3.3, 3.4 and 4.1.1.

The optimum network function $Y_M(p)$ is not realizable with a finite

network in either form of the problem, but it may be approximated arbitrarily closely. It should also furnish a reference for judging the performance of finite networks designed in other ways. The method has not yet been tested by detailed calculations.

4.6 *More General Modifications of the Central Problems*

The specific problems described in Sections 4.1 to 4.5 illustrate various more general ways in which the central data smoothing and prediction problems may be modified. These include the following:

i. The restriction of the function class C_Y, by constraints which specify the response to certain frequencies (Section 4.1) or to certain more general time functions (Section 4.3).

ii. The substitution of simplified spectra, which approximate true spectra only at frequencies which are actually utilized (Section 4.1.2).

iii. The addition of signal or noise functions which involve a finite number of random variables (Sections 4.1.3, 4.1.4).

iv. The estimation or prediction of a functional of the true signal, rather than the signal itself (Sections 4.1.3, 4.1.4).

v. The substitution of random variables other than signal and noise, and the treatment of "biased" statistics (Section 4.2).

vi. The use of more than two simultaneous observations (Section 4.4).

vii. The application of the mathematics of data smoothing to non-statistical problems (Section 4.5).

Other modifications are possible, which have not been illustrated. For example, correlations between signal and noise may be handled very simply, by modifying the physical model illustrated in Fig. 8. It is only necessary to change the associated frequency functions, so as to generate the correct pertinent covariances listed in Section 3.2.4. The methods of Section 3.3 and 3.4 may be combined, to find two optimum operators Y_{G1}, Y_{G2} restricted to finite memories. The single operator problem may be solved for signal and noise situations which are different in different segments of past time. In a special case, $n(\tau)$ is observed when $-\infty < \tau < t - T$, and $s(\tau) + n(\tau)$ is observed when $t - T < \tau < t$. Added complications, however, are likely to increase very drastically the number of simultaneous linear equations which must be solved to find $Y_M(p)$.

A general solution of the following problem would be of interest to engineers, particularly in connection with preliminary system studies: Suppose the signal or noise spectrum is not known in detail but is known to lie within some sort of limits. What Y_G will give the best protection against a large σ^2? If the most unfavorable permitted spectrum is asso-

ciated with each Y_a, what Y_a will make the corresponding σ^2 a minimum, and how large will the minimum be? It appears that no general solution of this problem has been achieved.

V. ACKNOWLEDGMENTS

The author is indebted to numerous colleagues for valuable discussions, suggestions and criticisms which contributed to the development of the theory. The preparation of the manuscript was aided particularly by suggestions and criticisms from T. M. Burford, L. A. MacColl and H. O. Pollak.

REFERENCES

1. Wiener, N., *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*, John Wiley & Son, New York, 1949.
2. Kolmogoroff, A., Interpolation und Extrapolation von Stationären Zufälligen Folgen, Bull. Acad. Sci. (URSS) Ser. Math., **5,** 1941, pp. 3–14.
3 Zadeh, L. A. and Ragazzini, J. R., An Extension of Wiener's Theory of Prediction, J. Appl. Phys., **21,** July 1950, pp. 645–655.
4. Darlington, S., Vertical Velocity Computer Data Smoothing Possibilities, technical memorandum written for Sandia Corp., March 31, 1950.
5. *Command Inertial Guidance System for a Ballistic Missile*, report prepared by Bell Telephone Laboratories for Western Electric Co., April 1955; particularly, Section IV, Vol. II, Data Smoothing in a Ballistic Missile Guidance System, March 1955.
6. *Transactions of the First Technical Symposium on Ballistic Missiles*, The Ramo-Wooldridge Corp., Los Angeles, 1956.
7. Stumper, F. L., A Bibliography of Information Theory Communication — Cybernetics, Trans. I.R.E., **PGIT,** November 1953, Part IV, pp. 12–18; **IT-1,** September 1955, Part IV, pp. 35–37.
8. Doob, J. L., *Stochastic Processes*, John Wiley & Son, New York, 1953.
9. Bode, H. W. and Shannon, C. E., A Simplified Derivation of Linear Least-Squares Smoothing and Prediction Theory, Proc. I.R.E., **38,** April 1950, pp. 417–425.
10. Lloyd, S. P. and McMillan, B., Linear Least-Squares Filtering and Prediction of Sampled Signals, Proc. Symp. on Mod. Network Synthesis, Microwave Research Institutie Symposia Series, **5,** 1956, pp. 221–247.
11. Chang, S. S. L., Two Network Theorems for Analytical Determination of optimum-Response Physically Realizable Network Characteristics, Proc. I.R.E., **43,** September 1955, pp. 1128–1135.
12. Laning, J. H. and Battin, R. H., *Random Processes in Automatic Controls*, McGraw-Hill, New York, 1956.
13. Crooks, J. W., Jr., Guidance System for the MX-774 Missile, Report No. ZM7-011, Consolidated Vultee Aircraft Corp., August 18, 1948.
14. Bendat, J. S., Optimum Filters for Independent Measurements of Two Related Perturbed Messages, Trans. I.R.E., **CT-4,** March 1957, pp. 14–19.
15. Stewart, R. M. and Parks, R. J., Degenerate Solutions and an Algebraic Approach to the Multiple Input Filter Design Problem, Trans. I.R.E., **CT-4,** March 1957, pp. 10–14.
16. Bode, H. W., U. S. Patent 2,123,178, issued July 12, 1938; also, *Network Analysis and Feedback Amplifier Design*, D. Van Nostrand, New York, 1945
17. Beurling, A., On Two Problems Concerning Linear Transformations in Hilbert Space, Acta Math., 1948.
18. Nyman, B., On the One-Dimensional Translation Group and Semigroup in

Certain Function Spaces, Thesis of Upsala Univ., 1950, rev. in Math. Reviews, **2,** February 1951, p. 108.
19. Youla, D. C., Castriota, L. J. and Carlin, H. J., Scattering Matrices and the Foundations of Linear, Passive Network Theory, Report R-594-57, PIB-522 for Air Force Office of Scientific Research, Polytechnic Inst. of Brooklyn, September 1957.
20. Rice, S. O., The Mathematical Analysis of Random Noise, B.S.T.J., **23,** July 1944, pp. 282–332, and **24,** January 1945, pp. 46–156.
21. Booton, R. C., Jr., An Optimization Theory for Time-Varying Linear Systems with Nonstationary Statistical Inputs, Proc. I.R.E., **40,** August 1952, pp. 977–981.
22. Miller, K. S. and Zadeh, L. A., Solution of an Integral Equation Occurring in the Theories of Prediction and Detection, Trans. I.R.E., **IT-2,** June 1956, pp. 72–75.
23. Bellman, R., A Survey of the Theory of Boundedness, Stability and Asymptotic Behavior of Solutions of Linear and Nonlinear Differential and Difference Equations, for Office of Naval Research, Princeton Univ., January 1949.
24. Darlington, S., An Introduction to Time-Variable Networks, Proc. of Symp. on Circuit Analysis, Univ. of Illinois, 1955.
25. Phillips, R. S. and Weiss, P. R., Theoretical Calculation of Best Smoothing of Position Data for Gunnery Prediction, Radiation Lab. Report No. 532, February 1944.
26. Blackman, R. B., Bode, H. W. and Shannon, C. E., Data Smoothing and Prediction in Fire-Control Systems, Summary Technical Report of Div. 7, NDRC Vol. 1, Report Series #13, MGC 12/2, National Military Establishment Research and Development Board.
27. North, D. O., An Analysis of the Factors Which Determine Signal-Noise Discrimination in Pulsed Carrier Systems, RCA Lab. Report No. PTR-6C, 1943.
28. Zadeh, L. A. and Ragazzini, J. R., Optimum Filters for the Detection of Signals in Noise, Proc. I.R.E., **40,** October 1952, pp. 1223–1231