Copyright © 1973 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 10, December, 1973
Printed in U.S.A.

Theory of Minimum Mean-Square-Error QAM Systems Employing Decision Feedback Equalization

By D. D. FALCONER and G. J. FOSCHINI

(Manuscript received June 7, 1973)

Decision feedback equalization is presently of interest as a technique for reducing intersymbol interference in high-rate PAM data communications systems. The basic principle is to cancel out intersymbol interference arising from previously decided data symbols at the receiver, leaving remaining intersymbol interference components to be handled by linear equalization. In this work we consider the application of decision feedback equalization to quadrature-amplitude modulation (QAM) transmission, in which two independent information streams modulate quadrature carriers. Extending Salz's treatment in a companion paper of decision feedback for a baseband channel, we derive the form of the optimum receiver filters via a matrix Wiener-Hopf analysis. We obtain explicit analytical expressions for minimum mean-square error and optimum transmitting filters. The optimization is subject to a constraint on the transmitted signal power and assumes no prior decision errors. The class of QAM transmitter and receiver structures treated here is actually much larger than the class usually considered for QAM systems. However, our results for decision feedback equalization show that, for nonexcess bandwidth systems, optimum performance is achievable without taking advantage of the most general structure. If the transmitter is required to have the conventional QAM structure, study of the time continuous system that gives rise to the sampled data system considered here demonstrates that under quite general assumptions a nonexcess bandwidth system is optimum. Finally, the explicit description of the optimum transmitting matrix filter follows from an information-theoretic "water-pouring" algorithm in conjunction with the determination of the form of the points of maxima of a determinant extremal problem.

I. INTRODUCTION

Interest has recently intensified in receiver structures which hopefully will permit higher data symbol rates than are possible with conventional demodulator/linear equalizer structures having the same error probability. The decision feedback equalizer is an example of a receiver component that can have important performance advantages over a linear equalizer operating over dispersive channels with additive noise.[1-7] The basic structure of a decision feedback equalizer (DFE) is shown in Fig. 1. The function of the filter in the feedback path is to cancel "postcursors" of the channel's impulse response; that is, intersymbol interference components arising from *previously* decided symbols. Thus, the job of the linear filter in the forward path is to minimize (according to some criterion) "precursors" of the channel's impulse response which cause intersymbol interference from future data symbols. Of course, there is a possibility of error propagation with this nonlinear feedback structure. We avoid this intractable problem by assuming that no erroneous decisions pass into the feedback filter. Thus, our results provide a performance lower bound. Earlier experimental studies indicated that error propagation is not a serious problem on some channels.[3,4]

Price[6] (whose bibliography on the subject is extensive) has derived asymptotic formulas (allowing for an infinite number of equalizer taps) for error probability, optimum transmitter pulse spectrum, and communication efficiency for the "zero-forcing" DFE, which minimizes the noise variance at the DFE output subject to the constraint that the intersymbol interference is zero at the receiver's sampling instants. As is the case for linear equalization, the mean-square-error (MSE) criterion is more general than the zero-forcing criterion. The MSE criterion minimizes the mean square of the total error (noise plus residual intersymbol interference) at the DFE output.[2,5] Asymptotic results and illuminating calculations of performance for MSE-minimizing DFE's are contained in a companion paper by Salz.[7]

All previous theoretical studies of decision feedback equalization have assumed a "baseband" linear PAM channel model depicted in

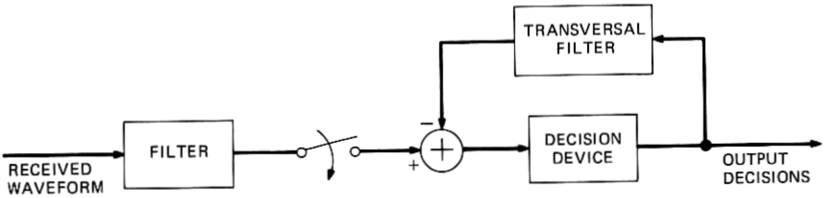


Fig. 1—Basic decision feedback equalizer structure.

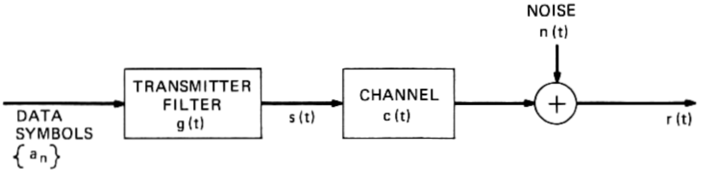


Fig. 2—Baseband channel model.

Fig. 2. The transmitted waveform $s(t)$ is

$$s^{(b)}(t) = \sum_n a_n g^{(b)}(t - nT),$$

where the data symbols $\{a_n\}$ are statistically independent discrete-valued random variables from a finite set and $g^{(b)}(t)$ is some suitable transmitted pulse waveform. The channel output waveform is then

$$r^{(b)}(t) = \sum_n a_n h^{(b)}(t - nT) + n^{(b)}(t),$$

where the overall impulse response is

$$h^{(b)}(t) = \int_{-\infty}^{\infty} c^{(b)}(\tau) g^{(b)}(t - \tau) d\tau$$

and $n^{(b)}(t)$ is additive noise. This model is certainly valid for a real linear channel accepting every T second a pulse of the form $a_n g^{(b)}(t)$. It is also valid for the important case where the linear channel $c^{(b)}(t)$ is actually the baseband equivalent of a passband channel when the modulation is either double-, vestigial-, or single-sideband. (See Ref. 8, Chapter 7.) Of course, $c^{(b)}(t)$ then depends on the carrier frequency and on any phase offset between the reference carriers at the modulator and demodulator.

In this paper we extend the asymptotic DFE theory to the case of QAM (quadrature amplitude modulation) signaling, for which the baseband model of Fig. 2 is not sufficient. We summarize our results at the end of Section II. The most general QAM transmitter structure is illustrated in Fig. 3. Two independent data sequences enter a lattice network comprising filters with impulse responses $g_{11}(t)$, $g_{21}(t)$, $g_{12}(t)$, and $g_{22}(t)$. Modulation is done with two quadrature carriers with frequency f_0 Hz. In practice, most QAM transmitters are specialized to the case $g_{11}(t) = g_{22}(t)$; $g_{21}(t) = -g_{12}(t)$.* We call the class of trans-

* Indeed, it is often assumed that $g_{12}(t)$ and $g_{21}(t)$ are zero.

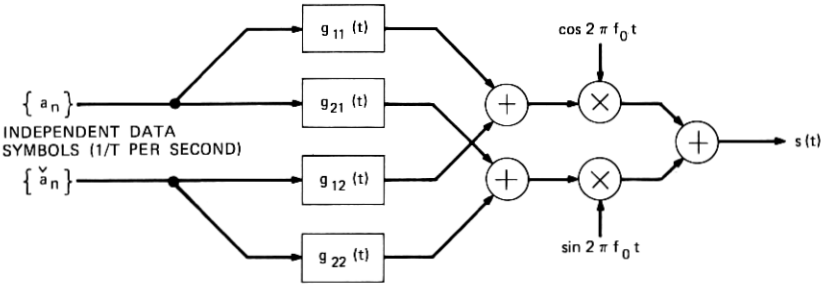


Fig. 3—General QAM transmitter structure.

mitters with this special structure the class of "passband" transmitters ($\mathcal{P}$). We show in later sections that optimum performance is in general achievable by restricting the transmitter to this class or a simple variant thereof. It is worth noting that QAM systems with passband transmitters are mathematically equivalent to baseband PAM systems, but with complex impulse responses and information symbols.[9,10]

For the most general QAM structure, the waveform $s(t)$ is expressed in terms of two-dimensional vectors and matrices. Define the vector $\mathbf{a}_n$ to be the nth pair of information symbols,

$$\mathbf{a}_n = \begin{pmatrix} a_n \\ \check{a}_n \end{pmatrix}. \tag{1}$$

The most general QAM transmitter is characterized by the matrix filter

$$g(t) = \begin{pmatrix} g_{11}(t) & g_{12}(t) \\ g_{21}(t) & g_{22}(t) \end{pmatrix}. \tag{2}$$

Then the structure of Fig. 3 yields

$$s(t) = (\cos 2\pi f_0 t, \sin 2\pi f_0 t) \sum_n g(t - nT)\mathbf{a}_n. \tag{3}$$

We assume that the data symbols are uncorrelated discrete-valued random variables with variance σ_a^2. Thus

$$\langle \mathbf{a}_n \mathbf{a}_m^\dagger \rangle = \sigma_a^2 \delta_{nm} I, \tag{4}$$

where $\langle\ \rangle$ denotes expectation, † denotes transpose,* δ_{nm} is the Kronecker delta, and I is the identity matrix. The transmitted power is

* The symbol † will denote conjugate transpose for complex vectors and matrices.

then given by

$$P \equiv \lim_{\tau\to\infty} \frac{1}{2\tau}\int_{-\tau}^{\tau} \langle s^2(t)\rangle dt = \frac{\sigma_a^2}{2T}\int_{-\infty}^{\infty} [g_{11}^2(t) + g_{12}^2(t) + g_{21}^2\,(t) + g_{22}^2(t)]dt$$

$$= \frac{\sigma_a^2}{4\pi T}\int_{-\infty}^{\infty} [|G_{11}(\omega)|^2 + |G_{12}(\omega)|^2 + |G_{21}(\omega)|^2 + |G_{22}(\omega)|^2]d\omega, (5)$$

where $G_{ij}(\omega)$ is the Fourier transform of $g_{ij}(t)$. For future reference note that we can also write P as

$$P = \frac{\sigma_a^2}{4\pi T}\int_{-\pi/T}^{\pi/T} \sum_n \left[\left|G_{11}\left(\omega + \frac{2\pi n}{T}\right)\right|^2 + \left|G_{12}\left(\omega + \frac{2\pi n}{T}\right)\right|^2 + \left|G_{21}\left(\omega + \frac{2\pi n}{T}\right)\right|^2 + \left|G_{22}\left(\omega + \frac{2\pi n}{T}\right)\right|^2\right]d\omega$$

$$= \frac{\sigma_a^2}{4\pi T}\int_{-\pi/T}^{\pi/T} \sum_n \operatorname{tr} G\left(\omega + \frac{2\pi n}{T}\right)^\dagger G\left(\omega + \frac{2\pi n}{T}\right)d\omega, \tag{6}$$

where

$$G(\omega) = \begin{pmatrix} G_{11}(\omega) & G_{12}(\omega) \\ G_{21}(\omega) & G_{22}(\omega) \end{pmatrix}$$

is the matrix frequency response of the transmitter. We use tr to denote the trace of a matrix.

Later sections will show that without an initial assumption of the special passband transmitter structure the treatment of decision feedback equalization for two- (and hence higher) dimensional signals is a nontrivial generalization of the baseband signal case.

II. THE CHANNEL MODEL AND SUMMARY OF RESULTS

The impulse response $q(t)$ of any linear channel can be resolved about a center frequency f_0:

$$q(t) = c_1(t)\cos 2\pi f_0 t - c_2(t)\sin 2\pi f_0 t. \tag{7}$$

It is easy to show that the channel model of Fig. 4 yields exactly the above impulse response, and thus any linear channel can be conveniently represented in terms of an arbitrary center frequency f_0 by the structure of Fig. 4. We note in passing that the so-called "in-phase" and "quadrature" impulse responses $c_1(t)$ and $c_2(t)$ are often interpreted as the real and imaginary parts, respectively, of the "complex envelope" of the impulse response $q(t)$ with respect to the frequency f_0.

We assume that the low-pass transmitter impulse responses $\{g_{ij}(t)\}$

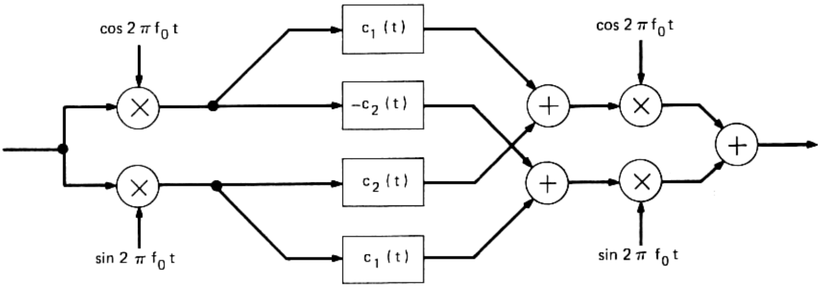


Fig. 4—A passband channel model.

are all strictly bandlimited to lie within the frequency interval $(-f_0, f_0)$; otherwise, the system would suffer distortion from aliasing effects. There is then no loss of generality in assuming that the channel's in-phase and quadrature impulse responses $c_1(t)$ and $c_2(t)$ are also strictly bandlimited to this interval.

With these assumptions, double-frequency terms disappear,[11] and it is easily shown that the noise-free channel output

$$\int_{-\infty}^{\infty} q(\tau)s(t-\tau)d\tau, \tag{8}$$

where $q(t)$ is given by (7) and $s(t)$ by (3), can be written

$$\tfrac{1}{2}(\cos 2\pi f_0 t, \sin 2\pi f_0 t) \sum_n \int_{-\infty}^{\infty} c(t-\tau)g(\tau-nT)d\tau \mathbf{a}_n, \tag{9}$$

where $c(t)$ is the matrix

$$c(t) = \begin{pmatrix} c_1(t) & c_2(t) \\ -c_2(t) & c_1(t) \end{pmatrix}, \tag{10}$$

the matrix $g(t)$ is given by (2), and integration of matrices and vectors means integration of each entry.

Consider receiver structures whose "front end" is the type shown in Fig. 5—sine and cosine demodulators followed by identical ideal low-pass filters that are strictly bandlimited to $(-f_0, f_0)$ and whose outputs are labelled $r(t)$ and $\check{r}(t)$, respectively. This structure causes no loss of information, since any bandlimited input signal can be reproduced exactly if the outputs $r(t)$ and $\check{r}(t)$ are multiplied by $\cos 2\pi f_0 t$ and $\sin 2\pi f_0 t$, respectively, and then added together. The function of the low-pass filters is to remove double frequency terms; it will turn out that the "front end" will be followed by a band-limiting matched filter, so that the low-pass filters are not necessary.

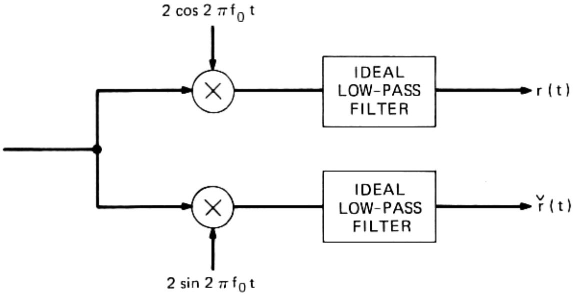


Fig. 5—Receiver "front end."

The low-pass outputs $r(t)$ and $\check{r}(t)$ can be written in vector form as

$$\mathbf{r}(t) = \begin{pmatrix} r(t) \\ \check{r}(t) \end{pmatrix} = \frac{1}{2} \sum_n h(t - nT)\mathbf{a}_n + \mathbf{v}(t), \tag{11}$$

where the matrix impulse response $h(t)$ is

$$h(t) = \int_{-\infty}^{\infty} c(\tau)g(\tau - \tau)d\tau, \tag{12}$$

and the components of the vector

$$\mathbf{v}(t) = \begin{pmatrix} n_c(t) \\ n_s(t) \end{pmatrix}$$

represent additive noise. Assuming that the additive noise in the channel is white with double-sided power spectral density $N_0/2$, it can be shown[11] that $n_c(t)$ and $n_s(t)$ are statistically independent stationary zero mean processes; each is the result of passing a stationary white noise with double-sided power spectral density N_0 through an ideal low-pass filter. Noise outside the signal bandwidth will be eliminated by a matched filter. Accordingly, we take the covariance matrix of the noise to be

$$\langle \mathbf{v}(t)\mathbf{v}^\dagger(t + \tau) \rangle = N_0 I \delta(\tau), \tag{13}$$

where I is the identity matrix and $\delta(t)$ is a "unit-area delta function." The mathematical model for the transmitter and channel is now complete and is summarized in Fig. 6a.

We remark in passing that linear modulation of a single stream of data symbols (e.g., single-sideband or vestigial-sideband modulation) constitutes a special case of this model. In that case, $g_{12}(t) = g_{22}(t) = 0$, and the receiver front end consists of a cosine demodulator with some phase shift θ, followed by an ideal low-pass filter. Then the overall

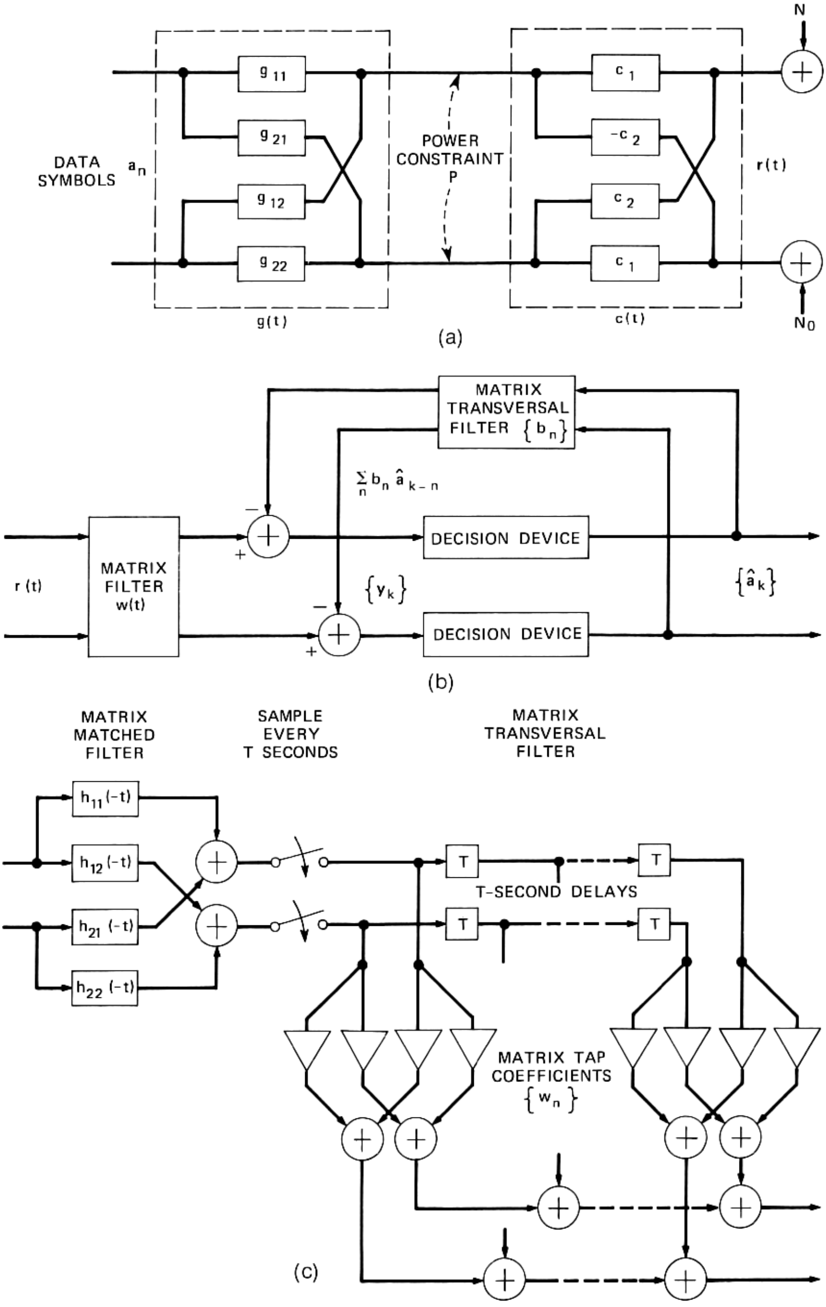


Fig. 6—(a) Canonical mathematical model of transmitter and channel. (b) QAM decision feedback equalizer structure. (c) Structure of the matrix filter $w(t)$.

impulse response is a scalar function of time (which depends on the receiver phase shift θ), and hence all two-dimensional matrices and vectors in the present treatment would be replaced by scalar quantities (see Ref. 7).

The following list summarizes our main results:

(*i*) The optimum linear forward filter at the receiver for a given transmitter channel cascade, $H(\omega) = C(\omega)G(\omega)$, is found to have the form

$$\text{Const} \times H(\omega)/\sqrt{\Phi(e^{-j\omega T}) + \frac{N_0}{\sigma_a^2}I}^{\wedge\phi},$$

where

$$\Phi(e^{-j\omega T}) = \frac{1}{T}\sum_n H^\dagger\left(\omega + \frac{2\pi n}{T}\right)H\left(\omega + \frac{2\pi n}{T}\right)$$

and $\sqrt{\ }^{\wedge\phi}$ denotes minimum-phase square root. This filter can be viewed as a matrix matched filter followed by an anticausal matrix tap delay line. (See Sections III and VI.)

(*ii*) For a given transmitter power spectral density: if a nonexcess bandwidth system (Section V) is required, an optimum transmitter is found and it is passband; conversely if the transmitter is taken to be passband, the optimum system is a nonexcess bandwidth one (Section III).

(*iii*) Given a passband transmitter, the MSE (the sum of the mean-square errors of the two unquantized receiver outputs) is

$$2\sigma_a^2 \exp\left\{-\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \log\left[\frac{\sigma_a^2}{N_0}X_{eq}(\omega) + 1\right]d\omega\right\},$$

where

$$X_{eq}(\omega) = \sum_n \left|G_1\left(\omega + \frac{2\pi n}{T}\right) + jG_2\left(\omega + \frac{2\pi n}{T}\right)\right|^2 \times\left|C_1\left(\omega + \frac{2\pi n}{T}\right) + jC_2\left(\omega + \frac{2\pi n}{T}\right)\right|^2$$

and $G_1(\omega) = G_{11}(\omega) = G_{22}(\omega)$ and $G_2(\omega) = G_{12}(\omega) = -G_{21}(\omega)$ (Section VI).

(*iv*) The optimum transmitter power spectral density is found for the class of passband transmitters meeting an output power constraint (Section VII). This optimal density has a water-pouring description. (Since the processing capability considered here represents an advancement over conventional linear equalization, this emergence of an information theoretic type density is perhaps not surprising.)

(*v*) Although we do not constrain the in-phase and quadrature mean-square errors to be equal, we show that for the above-mentioned optimal systems the errors on the two data streams have equal variances and are uncorrelated (Section VI).

In a nutshell, the system optimization proceeds as follows:

(*i*) Find the optimal receiver for each transmitter.
(*ii*) Find an optimal transmitter for each transmitter power spectral density.
(*iii*) Find the optimal transmitter power spectral density.

Then we reverse, using the solution of (*iii*) to specify an optimal transmitter and then using this optimal transmitter to specify the optimal receiver.

III. THE RECEIVER OPTIMIZATION PROBLEM

The DFE structure consists of a linear matrix filter $w(t)$, quantizer, and a transversal feedback filter with matrix tap coefficients $\{b_n\}$ which processes previously made decisions as shown in Fig. 6b. The kth sampled vector input to the quantizers is written

$$\mathbf{y}_k = \int_{-\infty}^{\infty} w(\tau)\mathbf{r}(kT - \tau)d\tau - \sum_{n=1}^{\infty} b_n\hat{\mathbf{a}}_{k-n}, \tag{14}$$

where $\hat{\mathbf{a}}_n$ is the receiver's decision on the nth data symbol-pair. Note that we allow the feedforward and feedback matrix filters to have infinite-duration impulse responses. We also replace $\hat{\mathbf{a}}_{k-n}$ in (14) by the true data symbol vector $\mathbf{a}_{k-n}$ for mathematical tractibility; thus, we in effect postulate a "magic genie" preceding the feedback filter who corrects any decision errors. The genie's existence is immaterial up to the time of the first decision error, and hence our expression for MSE is certainly valid up to that time.

The error vector $\boldsymbol{\varepsilon}_n$ is defined to be the difference between $\mathbf{y}_k$ and the correct symbol $\mathbf{a}_k$,

$$\boldsymbol{\varepsilon}_k = \mathbf{y}_k - \mathbf{a}_k, \tag{15}$$

and the MSE is defined to be the trace of the *error matrix* e_0, where

$$e_0 = \langle \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^{\dagger} \rangle, \tag{16}$$

the average being with respect to the noise and the data symbol sequence. Note that e_0 is positive semidefinite and symmetric.

Substituting (14) and (15) into (16) and using the noise correlation matrix (13) and the data symbol correlation matrix (4), we can write

$$e_0 = \sigma_a^2 \sum_{n \geqq k} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w(\tau_1) h[(k-n)T - \tau_1] h^\dagger[(k-n)T - \tau_2]$$
$$\times w^\dagger(\tau_2) d\tau_1 d\tau_2 + N_0 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w(\tau_1) w^\dagger(\tau_2) \delta(\tau_1 - \tau_2) d\tau_1 d\tau_2$$
$$+ \sigma_a^2 I + \sigma_a^2 \sum_{n=1}^{\infty} \left[b_n - \int_{-\infty}^{\infty} w(\tau) h(nT - \tau) d\tau \right]$$
$$\times \left[b_n - \int_{-\infty}^{\infty} w(\tau) h(nT - \tau) d\tau \right]^\dagger - \sigma_a^2 \int_{-\infty}^{\infty} w(\tau) h(-\tau) d\tau$$
$$- \sigma_a^2 \int_{-\infty}^{\infty} h^\dagger(-\tau) w^\dagger(\tau) d\tau. \quad (17)$$

We immediately observe that tr e_0 is minimized with respect to the matrices $\{b_n\}$ if and only if for all $n \geqq 1$, $b_n = s_n$, where

$$s_n \equiv \int_{-\infty}^{\infty} w(\tau) h(nT - \tau) d\tau \quad \text{all } n \quad (18)$$

represents the matrix samples for $n \geqq 1$ of the impulse response of the transmitter/channel-receiver filter combination. Then, once the $\{b_n\}$ are optimized in this way, the remaining terms comprising the matrix e_0 can be written

$$e_0 = \sigma_a^2 \sum_{n \leqq 0} \left[\delta_{n0} I - \int_{-\infty}^{\infty} w(\tau) h(nT - \tau) d\tau \right]$$
$$\times \left[\delta_{n0} I - \int_{-\infty}^{\infty} w(\tau) h(nT - \tau) d\tau \right]^\dagger + N_0 \int_{-\infty}^{\infty} w(\tau) w^\dagger(\tau) d\tau. \quad (19)$$

We wish to minimize tr e_0 with respect to the entries in the matrix $w(t)$. Notice from eq. (16) that tr e_0 is a positive quadratic form. Thus from Ref. 12 we set the gradient equal to zero to determine the stationary points which are necessarily points of global minima. We shall find that there is only one solution.

Proceeding with the calculus of variations method, we replace

$$w(t) = \begin{pmatrix} w_{11}(t) & w_{12}(t) \\ w_{21}(t) & w_{22}(t) \end{pmatrix}$$

by

$$w(t) + \begin{pmatrix} \epsilon_{11}\eta_{11}(t) & \epsilon_{12}\eta_{12}(t) \\ \epsilon_{21}\eta_{21}(t) & \epsilon_{22}\eta_{22}(t) \end{pmatrix},$$

where the $\eta_{ij}(t)$ are arbitrary. Setting

$$\begin{bmatrix} \dfrac{\partial \mathrm{tr}\, e_0}{\partial \epsilon_{11}} & \dfrac{\partial \mathrm{tr}\, e_0}{\partial \epsilon_{12}} \\ \dfrac{\partial \mathrm{tr}\, e_0}{\partial \epsilon_{21}} & \dfrac{\partial \mathrm{tr}\, e_0}{\partial \epsilon_{22}} \end{bmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{at } \epsilon_{11} = \epsilon_{12} = \epsilon_{21} = \epsilon_{22} = 0,$$

we get

$$-\sigma_a^2 h^\dagger(-\tau) + \sigma_a^2 \sum_{n \leqq 0} \int_{-\infty}^{\infty} w(\tau_1) h(nT - \tau_1) d\tau_1 h^\dagger(nT - \tau) + N_0 w(\tau) = [0]$$

or

$$w(\tau) = \sum_{n \leqq 0} w_n h^\dagger(nT - \tau), \tag{20}$$

where

$$w_n = \frac{\sigma_a^2}{N_0} (\delta_{n0} I - s_n) \quad n \leqq 0. \tag{21}$$

This means that the matrix filter $w(t)$ can be interpreted as a matrix matched filter with impulse response $h^\dagger(-t)$ followed by a sampler and matrix transversal filter with matrix tap coefficients $\{w_n\}$. Note that the transversal filter is "anticausal"—that is, $w_n = [0]$ for $n > 0$. The structure of $w(t)$ is illustrated in Fig. 6c.

Furthermore, substitution of the optimum filter (10) back into expression (19) for the error matrix e_0 results in

$$e_0 = \sigma_a^2 (I - s_0)^\dagger \tag{22a}$$

and from (21)

$$e_0 = N_0 w_0^\dagger. \tag{22b}$$

An explicit solution for the optimum tap coefficient matrices $\{w_n\}$ can be obtained by postmultiplying (20) by $h(mT - \tau)$ and integrating, using (21) and (18) and the definition

$$\phi_n \equiv \int_{-\infty}^{\infty} h^\dagger(-\tau) h(nT - \tau) d\tau$$

to yield

$$\sum_{m \leqq 0} w_m \left[\phi_{n-m} + \frac{N_0}{\sigma_a^2} \delta_{m-n,0} I \right] = \delta_{n0} I \quad \text{for } n \leqq 0. \tag{23}$$

We recognize eq. (23) as a classical Wiener-Hopf equation for which we are assured the existence of a unique solution.[13]

We attach a plus (minus) subscript to any matrix sequence whose value is the zero matrix on the strictly negative (positive) integers.*

* A matrix sequence u_+, zero on the strictly negative integers, is referred to as *causal*. A sequence u_-, zero on the strictly positive integers, is referred to as *anticausal*.

By 1 we mean the matrix sequence vanishing everywhere but zero, where the value is I. Then (23) is written

$$w_{-}*\mathscr{r} = 1 \quad (n \leqq 0), \tag{24}$$

where we have used $\mathscr{r}$ to denote the sequence sum which we observe is the Fourier coefficient sequence for a positive definite Hermitian matrix function whose determinant is uniformly above (N_0/σ_a^2). Hence $\mathscr{r}$ admits a causal, anticausal deconvolution of the kind provided by Wiener and Akutowicz[14] (generalizing a result of Szegö). Based on Ref. 14, we can say

$$\mathscr{r} = u_{-}*u_{+}, \quad \{(u_{-})_n = [(u_{+})_{-n}]^{\dagger}\}_{n=0}^{-\infty}, \tag{25}$$

where we have used $(u_{-})_n$ to denote the nth entry in the u_{-} sequence (similarly for u_{+}). Corresponding to u_{-}, we have its convolution inverse, $[u_{-}]^{-1}$, which is also anticausal.

From what we have just said,

$$[(u_{+})_0]^{-1}[u_{-}]^{-1}*u_{-}*u_{+} = 1, \tag{26}$$

and so $w_{-} = [(u_{+})_0]^{-1}(u_{-})^{-1}$ and, in particular,

$$w_0 = (u_0 u_0^{\dagger})^{-1}, \tag{27}$$

where in the last equation the negative subscripts have been suppressed. Thus the anticausal transversal filter is found from eqs. (23) through (26) to have a frequency response inversely proportional to

$$\sqrt{\Phi(e^{-j\omega T}) + \frac{N_0}{\sigma_a^2} I}^{\,\wedge\,\phi},$$

where the notation $\sqrt{}^{\,\wedge\,\phi}$ means minimum-phase square root and $\Phi(e^{-j\omega T})$ is the discrete Fourier transform of the matrix sequence $\{\phi_n\}$. Recalling eq. (22b), we have the following expression for the error matrix:

$$e_0 = N_0 w_0^{\dagger} = N_0[u_0 u_0^{\dagger}]^{-1}. \tag{28}$$

We remark that the development so far is analogous to that of the baseband decision feedback equalizer.[5] Further progress toward achieving a closed-form expression for $\operatorname{tr} e_0$ thus depends on obtaining a closed-form expression for the matrix $[u_0 u_0^{\dagger}]^{-1}$ or for its trace, corresponding to the result recently developed for the baseband case.[7] It has not been possible to do this directly for the QAM case when the most general transmitter matrix is allowed. We shall prove under quite general conditions that the minimum of $\operatorname{tr} w_0$ is achieved with a transmitter of passband structure, and that, given this transmitter structure,

$$\text{tr } w_0 = \tfrac{1}{2}\sqrt{\det w_0}$$

and

$$\det w_0 = \exp\left\{-\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \log \det \left[\Phi(e^{-j\omega T}) + \frac{N_0}{\sigma_a^2} I\right] d\omega\right\},$$

where "det" denotes the determinant.

IV. CLOSED-FORM EXPRESSION FOR DET e_0

For the most general matrix filter, attainment of a closed-form MSE expression for tr e_0 in terms of the matrix $\Phi(e^{-j\omega T})$ has so far proved intractable. However, we shall see that such a general expression is unnecessary to describe the behavior of optimum systems. Our approach is to employ the following easily proven lower bound for 2 × 2 positive semi-definite symmetric matrices

$$\text{tr } w_0 \geqq 2|\det{}^{\frac{1}{2}} w_0|, \tag{29a}$$

which holds with equality if and only if w_0 is a scalar matrix (i.e., multiple of the identity). In this section we develop a closed-form expression for det w_0. In the following sections where we deal with optimum systems, we can always perform the analysis in a context where eq. (29) holds the equality.

We begin the analysis of $\sqrt{\det w_0}$ by recalling (25a), from which follows

$$\det\left[\Phi(e^{-j\omega T}) + \frac{N_0}{\sigma_a^2} I\right] = \det U_-(e^{-j\omega T}) \det U_-(e^{j\omega T}).$$

Then from the one-dimensional theory we have[15]

$$\det\,(u_0 u_0^\dagger) = \exp\left\{\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \log \det \left[\Phi(e^{-j\omega T}) + \frac{N_0}{\sigma_a^2} I\right] d\omega\right\},$$

and from (28) and (29a)

$$\frac{1}{N_0}\text{tr } e_0 = \text{tr } w_0 \geqq 2\sqrt{\det w_0}, \tag{29b}$$

where

$$\det w_0 = \exp\left\{-\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \log \det \left[\Phi(e^{-j\omega T}) + \frac{N_0}{\sigma_a^2} I\right] d\omega\right\}. \tag{29c}$$

V. FOR A NONEXCESS BANDWIDTH SYSTEM, PASSBAND TRANSMITTERS CANNOT BE OUTPERFORMED

In this section we begin by expressing Φ explicitly in terms of the transmitter and channel matrices. Then we define the notion of a nonexcess bandwidth system. The primary result of this section is that, for a nonexcess bandwidth system, if the transmitter power density function

$$f(\omega) = \operatorname{tr} G^{\dagger}G \qquad \left(f(\omega) \geqq 0, \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} f(\omega) d\omega = \frac{2TP}{\sigma_a^2} \right)$$

is specified, then there exists a passband transmitter in the class of all matrix transmitters optimal under the constraint that $f(\omega)$ is the power density function.

To display the dependence of our results so far on the transmitter frequency response $G(\omega)$, we first rewrite the matrix $\Phi(e^{-j\omega T})$ using the definition of ϕ_n as

$$\Phi(e^{-j\omega T}) \equiv \int_{-\infty}^{\infty} h^{\dagger}(-\tau) \mathcal{H}(\omega, \tau) d\tau, \tag{30a}$$

where

$$\mathcal{H}(\omega, \tau) \equiv \sum_n h(nT - \tau) e^{-j\omega nT}. \tag{30b}$$

Expression (30b) is a Fourier series. Thus,

$$h(nT - \tau) = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \mathcal{H}(\omega, \tau) e^{j\omega nT} d\omega. \tag{31}$$

But the matrix impulse response $h(nT - \tau)$ can also be written as the inverse Fourier transform of a matrix frequency response $H(\omega)$,

$$h(nT - \tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\omega) e^{j\omega(nT-\tau)} d\omega,$$

which, upon splitting up the range of integration and changing the variable of integration, can be written

$$h(nT - \tau) = \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} e^{j\omega nT} \times \left[\sum_m H\left(\omega + \frac{2\pi m}{T}\right) \exp\left[-j\left(\omega + \frac{2\pi m}{T}\right)\tau \right] \right] d\omega. \tag{32}$$

Equating the integrands in (31) and (32), we obtain an explicit expression for $\mathcal{H}(\omega, \tau)$ which when substituted into (30a) yields

$$\Phi(e^{-j\omega T}) = \frac{1}{T}\sum_n H\left(\omega + \frac{2\pi n}{T}\right)^{\dagger} H\left(\omega + \frac{2\pi n}{T}\right). \tag{33}$$

Furthermore, denoting the Fourier transforms of $c(t)$ and $g(t)$ by the channel matrix $C(\omega)$ and the transmitter matrix $G(\omega)$ respectively, we can write $H(\omega) = C(\omega)G(\omega)$ and

$$\Phi(e^{-j\omega T}) = \frac{1}{T}\sum_n G\left(\omega + \frac{2\pi n}{T}\right)^{\dagger} \times C\left(\omega + \frac{2\pi n}{T}\right)^{\dagger} C\left(\omega + \frac{2\pi n}{T}\right) G\left(\omega + \frac{2\pi n}{T}\right). \tag{34}$$

A *nonexcess bandwidth system* is defined by the property that for any radian frequency ω there is no more than one nonzero term in the above sum. It can be taken to be the $n = 0$ term by making a trivial frequency translation where necessary. Hence for a nonexcess bandwidth system

$$\Phi(e^{-j\omega T}) = \frac{1}{T} G(\omega)^{\dagger} C(\omega)^{\dagger} C(\omega) G(\omega) \qquad \left(|\omega| \leqq \frac{\pi}{T}\right). \tag{35}$$

In this section we deal exclusively with nonexcess bandwidth systems. In Section VIII we refer to a recent theorem of H. Witsenhausen which enables us to do a complete analysis of excess bandwidth systems by transforming them to a canonical nonexcess bandwidth "equivalent" and then transforming back.

To model the class of transmitter frequency responses $G(\omega)$, we introduce $\mathcal{G}$ to denote the (Hilbert) space of all 2×2 matrices whose entries are Hermitian symmetric $\{G(\omega) = [G(-\omega)]^*\}$ finite energy functions on $(-\pi/T, \pi/T)$. The Hermitian symmetry of the entries is required so that each entry represents the Fourier transform of a real-time function. As in Section I, we use $\mathcal{P}$ to denote the passband subspace of $\mathcal{G}$ consisting of matrices of the form

$$\begin{pmatrix} G_{11}(\omega) & G_{12}(\omega) \\ -G_{12}(\omega) & G_{11}(\omega) \end{pmatrix}.$$

We shall be dealing only with matrix filters G of finite power P, given by (6). Thus we use $\mathcal{G}_P$ and $\mathcal{P}_P$ to denote

$$\left\{ G \,\middle|\, \frac{1}{2\pi}\int_{-\pi/T}^{\pi/T} \operatorname{tr} G(\omega)G(\omega)^{\dagger} d\omega = \frac{2TP}{\sigma_a^2} \right\}$$

in $\mathcal{G}$ and $\mathcal{P}$, respectively. In the sequel, all transmitter filters will be assumed to have power P.

We now optimize det $(G^\dagger C^\dagger CG + (N_0/\sigma_a^2)\,I)$ at each radian frequency ω for a fixed amount of power transmitted at ω.

Fix $N_0 > 0,\ f > 0$ and $C = \begin{pmatrix} C_1 & C_2 \\ -C_2 & C_1 \end{pmatrix}$ (C_i's complex functions of frequency). Explicitly, we shall show that

$$\max \det \left\{ G^\dagger C^\dagger CG + \frac{N_0}{\sigma_a^2}\, I \right\}$$

over all complex G such that tr $G^\dagger G = f(\omega)$ is achieved for a G of the passband form $\begin{pmatrix} G_{11} & G_{12} \\ -G_{12} & G_{11} \end{pmatrix}$. (For linear QAM systems, the same determinant extremal problem arises in the optimum selection of a transmitter with a specified power spectral density function. To our knowledge, this aspect of linear QAM systems has escaped the literature.)

Notice the unitary transformation $\Psi = 2^{-\frac{1}{2}}\begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix}$ diagonalizes matrices of the form $\begin{pmatrix} a & jb \\ -jb & a \end{pmatrix}$ in that

$$\Psi^\dagger \begin{pmatrix} a & jb \\ -jb & a \end{pmatrix} \Psi = \begin{pmatrix} a+b & 0 \\ 0 & a-b \end{pmatrix}. \tag{36}$$

Since $C^\dagger C$ is of the form $\begin{pmatrix} a & jb \\ -jb & a \end{pmatrix}$ (a, b, real, $a > b$), if we let $G = \Psi B$ the problem becomes

$$\max \det \left\{ B^\dagger \begin{pmatrix} a+b & 0 \\ 0 & a-b \end{pmatrix} B + N_0 I \right\}, \quad \text{tr } B^\dagger B = f(\omega).$$

Let $D = \begin{pmatrix} a+b & 0 \\ 0 & a-b \end{pmatrix}$ and rewrite the problem as

$$\max \{\det(B^\dagger DB) + N_0 \operatorname{tr}(B^\dagger DB) + N_0^2\}, \quad \text{tr } B^\dagger B = f(\omega).$$

At this stage we denote the Hermitian matrix $BB^\dagger$ by Q and write

$$\max \{\det QD + N_0 \operatorname{tr} QD + N_0^2\}, \quad \text{tr } Q = f(\omega).$$

Of course, an optimum Q exists, since we are maximizing a continuous function over a compact set. A nonzero off-diagonal entry in Q would only affect the determinant and not the traces. Since Q is Hermitian, the optimal Q is diagonal. Retracking, $Q = BB^\dagger = \Psi^\dagger G(\Psi\Psi^\dagger)G^\dagger\Psi$. Now Q is positive definite and so has a positive definite square root $Q^{\frac{1}{2}}$. From the definition of B,

$$G = \Psi Q^{\frac{1}{2}} \Psi^{\dagger}$$

which shows that the optimal G has the form

$$G = \begin{pmatrix} G_{11} & G_{12} \\ -G_{12} & G_{11} \end{pmatrix}. \tag{37}$$

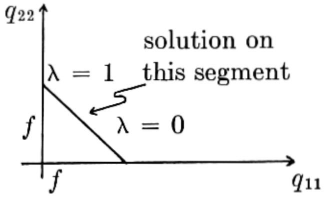
Although the proof is now complete, we go further and find G_{11} and G_{12}, as this will be used in the sequel. To find G_{11} and G_{12} we have seen that we must first find the entries q_{11} and q_{22} of Q so as to maximize

$$\{q_{11}q_{22}(a+b)(a-b) + N_0[q_{11}(a+b) + q_{22}(a-b)] + N_0^2\}$$

on the triangle in the (q_{11}, q_{22}) plane described by

$$q_{11} + q_{22} \leqq f, \quad q_{11} \geqq 0, q_{22} \geqq 0.$$

Since $a > 0$ and $a \geqq b$, the optimum (q_{11}, q_{22}) is achieved with $q_{11} + q_{22} = f$. Let λ linearly parametrize the segment joining $(f, 0)$ and $(0, f)$ as shown below



so $(q_{11}, q_{22}) = [(1-\lambda)f, \lambda f]$ where $(0 \leqq \lambda \leqq 1)$.

The criterion becomes: Maximize

$$f\{\lambda(1-\lambda)f(a^2-b^2) + N_0[(1-\lambda)(a+b) + \lambda(a-b)]\} + N_0^2, \tag{38}$$

which is a parabola concave in λ. Our problem is to determine $\lambda_{\text{opt}}(0 \leqq \lambda_{\text{opt}} \leqq 1)$. Now the parabola is maximized at

$$\bar{\lambda} = \frac{1}{2} - \frac{N_0}{f}\left(\frac{b}{a^2-b^2}\right).$$

If $\bar{\lambda}$ satisfies $0 \leq \bar{\lambda} < 1$, then $\lambda_{\text{opt}} = \bar{\lambda}$. If $\bar{\lambda} < 0$, $\lambda_{\text{opt}} = 0$ and if $\bar{\lambda} > 1$, $\lambda_{\text{opt}} = 1$.

So from $G = \Psi Q^{\frac{1}{2}} \Psi^{\dagger}$, we obtain

$$G_{11} = \frac{[|\sqrt{(1-\lambda_{\text{opt}})}| + |\sqrt{\lambda_{\text{opt}}}|]|f^{\frac{1}{2}}|}{|2^{\frac{1}{2}}|} \tag{39a}$$

$$G_{12} = \mathrm{j}\frac{(|\sqrt{1-\lambda_{\text{opt}}}| - |\sqrt{\lambda_{\text{opt}}}|)|f^{\frac{1}{2}}|}{|2^{\frac{1}{2}}|} \text{ signum } b. \tag{39b}$$

The determination of the signs attached to G_{11} and G_{12} was made by noticing that at each frequency

$$\det\left(G^{\dagger}C^{\dagger}CG + \frac{N_0}{\sigma^2}\,I\right) \tag{40}$$

is invariant to the sign of G_{11}, while (40) is maximized if signum b is used for G_{12}.

VI. CLOSED FORM EXPRESSION FOR ALL PASSBAND G

We have seen that, for nonexcess bandwidth systems, an extremal G for

$$\det w_0 = \exp\left\{-\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \log\det\left[\Phi(e^{-j\omega T}) + \frac{N_0}{\sigma_a^2}\,I\right]d\omega\right\}$$

exists in the space $\mathcal{P}$. Next we show that for each $G\in\mathcal{P}$, whether or not it has excess bandwidth, tr $w_0 = 2\sqrt{\det w_0}$. To do this we must show that w_0 is a scalar matrix. First observe that the matrices G and C are in $\mathcal{P}$, and their entries, being Fourier transforms of real-time functions, are Hermitian symmetric. The matrix $\Phi(e^{-j\omega T}) + (N_0/\sigma_a^2)I$, which is designated by $\mathcal{R}$ and is the Fourier transform of the matrix sequence $\mathscr{r}$ in (24), can be expressed in terms of the channel matrix $C(\omega)$ and a passband transmitter matrix

$$G(\omega) = \begin{pmatrix} G_1(\omega) & G_2(\omega) \\ -G_2(\omega) & G_1(\omega) \end{pmatrix}$$

as in (34) to yield

$$\mathcal{R} = \begin{pmatrix} \mathcal{R}_1(\omega) & j\mathcal{R}_2(\omega) \\ -j\mathcal{R}_2(\omega) & \mathcal{R}_1(\omega) \end{pmatrix},$$

where

$$\begin{aligned}\mathcal{R}_1(\omega) = \frac{1}{T}\sum_n \Bigg[\Bigg|&G_1\left(\omega+\frac{2\pi n}{T}\right)C_1\left(\omega+\frac{2\pi n}{T}\right)\\ &- G_2\left(\omega+\frac{2\pi n}{T}\right)C_2\left(\omega+\frac{2\pi n}{T}\right)\Bigg|^2\\ &+\Bigg|G_1\left(\omega+\frac{2\pi n}{T}\right)C_2\left(\omega+\frac{2\pi n}{T}\right)\\ &+ G_2\left(\omega+\frac{2\pi n}{T}\right)C_1\left(\omega+\frac{2\pi n}{T}\right)\Bigg|^2 + \frac{N_0}{\sigma_a^2}\Bigg]\end{aligned} \tag{41}$$

and

$$\begin{aligned}\mathcal{R}_2(\omega) = -\frac{2}{T}\sum_n \mathrm{Im}\, \Bigg\{ \Bigg[& G_2\left(\omega + \frac{2\pi n}{T}\right) C_1\left(\omega + \frac{2\pi n}{T}\right) \\ & + G_1\left(\omega + \frac{2\pi n}{T}\right) C_2\left(\omega + \frac{2\pi n}{T}\right)\Bigg]^* \\ & \times \Bigg[G_1\left(\omega + \frac{2\pi n}{T}\right) C_1\left(\omega + \frac{2\pi n}{T}\right) \\ & - G_2\left(\omega + \frac{2\pi n}{T}\right) C_2\left(\omega + \frac{2\pi n}{T}\right)\Bigg]\Bigg\}. \qquad (42)\end{aligned}$$

The entries $\mathcal{R}_1$ and $\mathcal{R}_2$ are real functions of ω. $\mathcal{R}_1$ is a positive even function and $\mathcal{R}_2$ is an odd function. The matrix $\mathcal{R}$ is positive definite; i.e., $\mathcal{R}_1^2 > \mathcal{R}_2^2$. It is also Hermitian and passband.

We have previously noted in eq. (25) that the matrix $\mathcal{R}$ can be factored into the anticausal and causal matrices $U_-(e^{-j\omega T})$ and $[U_-(e^{-j\omega T})]^\dagger$, respectively. The matrix $u_0 u_0^\dagger$, which is proportional to the error matrix inverse, is unique and the factor $U_-(e^{-j\omega T})$ is unique up to an arbitrary unitary matrix post-multiplicative factor Ψ. We now pick a particular unitary matrix.

The matrix $\mathcal{R}$ is diagonalized by the unitary matrix

$$\Psi = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix}; \quad \text{i.e., } \Psi^\dagger \mathcal{R} \Psi = \begin{pmatrix} \mathcal{R}_1 - \mathcal{R}_2 & 0 \\ 0 & \mathcal{R}_1 + \mathcal{R}_2 \end{pmatrix}. \qquad (43)$$

Now the entries $\mathcal{R}_1 - \mathcal{R}_2$ and $\mathcal{R}_1 + \mathcal{R}_2$ are nonnegative real functions on $-(\pi/T) \leqq \omega \leqq (\pi/T)$. Since

$$-\infty < \int_{-\pi/T}^{\pi/T} \log\, (\mathcal{R}_1 \pm \mathcal{R}_2) d\omega,$$

we have from Szegö's theorem[15] that

$$\mathcal{R}_1 - \mathcal{R}_2 = |\alpha_-|^2 \qquad (44a)$$

and

$$\mathcal{R}_1 + \mathcal{R}_2 = |\beta_-|^2, \qquad (44b)$$

where α_- and β_- are anticausal functions of ω, i.e.,

$$\alpha_- = \sum_{m \leqq 0} \alpha_m e^{-j\omega m T} \qquad (45a)$$

and

$$\beta_- = \sum_{m \leqq 0} \beta_m e^{-j\omega m T}; \qquad (45b)$$

the $\{a_m\}$ and $\{\beta_m\}$ being sequences of complex numbers. We can assume

α_0 and β_0 are real and positive without loss of generality. Therefore

$$\Psi^\dagger \mathcal{R} \Psi = VV^\dagger, \tag{46}$$

where

$$V = \begin{pmatrix} \alpha_- & 0 \\ 0 & \beta_- \end{pmatrix},$$

and since Ψ is a unitary matrix,

$$\begin{aligned} \mathcal{R} &= (\Psi V \Psi^\dagger)(\Psi V^\dagger \Psi^\dagger) \\ &= U_- U_-{}^\dagger, \end{aligned} \tag{47a}$$

where

$$U_- = \Psi V \Psi^\dagger = \begin{pmatrix} \alpha_- + \beta_- & j(\alpha_- - \beta_-) \\ -j(\alpha_- - \beta_-) & \alpha_- + \beta_- \end{pmatrix}. \tag{47b}$$

Thus in this factorization, U_- can be taken to be passband. Furthermore, since α_0 and β_0 are real,

$$u_0 = \begin{pmatrix} \alpha_0 + \beta_0 & j(\alpha_0 - \beta_0) \\ -j(\alpha_0 - \beta_0) & \alpha_0 + \beta_0 \end{pmatrix} \tag{48}$$

is both Hermitian and passband, and so therefore is the error matrix $e_0 = [u_0 u_0^\dagger]^{-1}$; i.e., its off-diagonal terms are purely imaginary.[*] But we know that e_0, defined by (15), must have real equal off-diagonal terms, and therefore e_0 must be a scalar matrix. Thus $w_0 = (1/N_0)e_0$ is also scalar and

$$\operatorname{tr} w_0 = 2\sqrt{\det w_0}. \tag{49}$$

Summarizing the development so far, we have shown that, for nonexcess bandwidth systems, if the transmitted power spectrum is specified, the passband transmitter structure is optimum. We then showed that if the transmitter has the passband structure, the MSE is given by eqs. (29b) and (29c), (29b) holding with equality. Incidentally, using the results of the last paragraph it can be shown that $\alpha_0 = \beta_0 = \sqrt{N_0/2\text{MMSE}}$; hence u_0 is known. In Section III the optimum linear receiver filter $w(t)$ was found up to the constant (matrix) multiplier u_0^{-1}. For nonexcess bandwidth passband systems, we can now make the more complete statement that the matrix Fourier transform of $w(t)$ is

$$\sqrt{\frac{\text{MMSE}}{2N_0}}\, G^\dagger(\omega) C^\dagger(\omega) \bigg/ \sqrt{G^\dagger(\omega) C^\dagger(\omega) C(\omega) G(\omega) + \frac{N_0 I}{\sigma_a^2}}^{\,\wedge \phi}$$

(where $\sqrt{}^{\,\wedge\phi}$ means minimum-phase square root).

[*] It is important to notice that, although u_0 is unique only up to a postmultiplicative unitary factor, the matrix $u_0 u_0^\dagger$ is unique.

We shall now express (29c) in a somewhat different form, which avoids the use of the determinant. As pointed out earlier, $\mathcal{R}_1(\omega)$ and $\mathcal{R}_2(\omega)$ are real even and odd functions, respectively. Thus

$$\mathcal{R}_1(\omega) - \mathcal{R}_2(\omega) = \mathcal{R}_1(-\omega) + \mathcal{R}_2(-\omega)$$

and

$$\int_{-\pi/T}^{\pi/T} \log\left[\mathcal{R}_1(\omega) - \mathcal{R}_2(\omega)\right] d\omega = \int_{-\pi/T}^{\pi/T} \log\left[\mathcal{R}_1(\omega) + \mathcal{R}_2(\omega)\right] d\omega, \quad (50)$$

from which it follows that

$$\begin{aligned} \operatorname{tr} e_0 &= 2N_0 \exp\left[-\frac{T}{4\pi}\int_{-\pi/T}^{\pi/T} \log \det \mathcal{R}(\omega) d\omega\right] \\ &= 2N_0 \exp\left(-\frac{T}{4\pi}\left\{\int_{-\pi/T}^{\pi/T} \log\left[\mathcal{R}_1(\omega) - \mathcal{R}_2(\omega)\right] d\omega \right.\right. \\ &\qquad \left.\left. + \int_{-\pi/T}^{\pi/T} \log\left[\mathcal{R}_1(\omega) + \mathcal{R}_2(\omega)\right] d\omega\right\}\right) \\ &= 2N_0 \exp\left\{-\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \log\left[\mathcal{R}_1(\omega) + \mathcal{R}_2(\omega)\right] d\omega\right\}. \end{aligned} \quad (51)$$

Substituting (41) and (42) into (51) gives the following expression for MSE:

$$\operatorname{tr} e_0 = 2\sigma_a^2 \exp\left\{-\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \log\left[\frac{\sigma_a^2}{N_0} X_{eq}(\omega) + 1\right] d\omega\right\}, \quad (52)$$

where

$$\begin{aligned} X_{eq}(\omega) = \frac{1}{T}\sum_n &\left| G_1\left(\omega + \frac{2\pi n}{T}\right) + jG_2\left(\omega + \frac{2\pi n}{T}\right)\right|^2 \\ &\times \left| C_1\left(\omega + \frac{2\pi n}{T}\right) + jC_2\left(\omega + \frac{2\pi n}{T}\right)\right|^2. \end{aligned}$$

This expression is valid for any passband transmitter and, as shown in the previous section, it is valid for optimum general QAM transmitters with no excess bandwidth.[*] We show in Appendix A that, under very general assumptions, optimum *passband* transmitters will have no excess bandwidth.

[*] We remark at this point that if we had restricted attention to passband transmitter structures from the outset, we could have derived the MSE expression (52) more directly by using the complex envelope notation referred to in Section II instead of the matrix formulation.

VII. OPTIMUM TRANSMITTER

Here we continue under the assumption of a nonexcess bandwidth system. So far, we know that, if $f = \text{tr}\, G^\dagger G$ is specified, an optimal passband G exists yielding a minimum for MSE and possessing power spectral density f. Our next step is to free $\text{tr}\, G^\dagger G$ and to find G, which minimizes MSE subject only to the constraint that

$$\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} f = \left(\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \text{tr}\, G^\dagger G =\right) = P. \tag{53}$$

Notice

$$\int_{-\pi/T}^{\pi/T} |G_1 + jG_2|^2 = \int_{-\pi/T}^{\pi/T} (|G_1|^2 + |G_2|^2) \tag{54}$$

since $G_1^* G_2 - G_1 G_2^*$ is odd. Thus our problem becomes to find $|G_1 + jG_2|^2$, minimizing

$$2 \exp - \left\{\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \log\left(\sigma_a^2 \frac{|G_1 + jG_2|^2 |C_1 + jC_2|^2}{TN_0} + 1\right)\right.$$

subject to $\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} |G_1 + jG_2|^2 = P.$

It is shown in Appendix B that the solution to this problem is given uniquely by

$$|G_1 + jG_2|^2 = \left(\mathcal{C} - \frac{N_0 T^2}{\sigma_a^2}|C_1 + jC_2|^{-2}\right)_+$$

(where $(\xi)_+ \triangleq \max[(\xi, 0)]$),

where $\mathcal{C}$ is a constant set at a value so that

$$\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} (|G_1|^2 + |G_2|^2) = P.$$

This solution also occurs in a related context in information theory, where it is dubbed "the water-pouring solution."[16]

Since $(G_1^*(\omega)G_2(\omega) - G_1(\omega)G_2^*(\omega))$ is odd and $f(\omega)$ is even, we average $|G_1^*(\omega) + jG_2(\omega)|^2$ and $|G_1^*(-\omega) + jG_2(-\omega)|^2$ to get*

$$f(\omega) = \frac{1}{2}\left\{\left[\mathcal{C} - \frac{N_0T^2}{\sigma_a^2}|C_1(\omega) + jC_2(\omega)|^{-2}\right]_+ + \left[\mathcal{C} - \frac{N_0T^2}{\sigma_a^2}|C_1(-\omega) + jC_2(-\omega)|^{-2}\right]_+\right\}.$$

To find $G_1(\omega)$ and $G_2(\omega)$, use the above $f(\omega)$ in Section V.

* Note that for $N_0 \to 0$, the optimum $f(\omega)$ tends to a constant.

VIII. THE ROLE OF NONEXCESS BANDWIDTH SYSTEMS

In the previous sections we have determined the optimum transmitter under the hypothesis that the system is nonexcess bandwidth. Here we point out that this hypothesis is not very restrictive.

In systems in which the transmitter is required to be passband, it follows, under very mild assumptions on the channel characteristics, that the optimum transmitter (subject to an output power constraint) is a nonexcess bandwidth system. The mathematical proof of the optimality of the nonexcess bandwidth system is considered in detail in Appendix A. For an example, if for each $\omega(|\omega| < \pi/T)$

$$|C_1(\omega) + jC_2(\omega)| > \left|C_1\left(\omega + \frac{2\pi k}{T}\right) + jC_2\left(\omega + \frac{2\pi k}{T}\right)\right| \qquad (k \neq 0),$$

then the optimal transmitter has no energy outside

$$\left\{\omega \middle| |\omega - 2\pi f_0| \leqq \frac{\pi}{T}\right\}.$$

For systems allowing any matrix transmitter, the question arises whether or not the optimal transmitter is passband. If the answer to the question is negative, the next question is whether or not the optimal transmitter is nonexcess bandwidth. The answers to these questions depend on the system parameters, and there are channels for which the answers to both questions are negative. It is beyond the scope of this paper to give a detailed mathematical discussion of these more complex systems. Such systems are still under investigation, and so we shall limit ourselves to mentioning without proof some important facts concerning the analysis of such systems.

The analysis begins by returning to Section IV fixing ω and posing the extremal problem of

$$\max \det \left(\sum G_k^\dagger C_k^\dagger C_k G_k + N_0 I\right),$$

subject to $\operatorname{tr} \sum G_k^\dagger G_k = f$. If for each ω it is optimal to expend all of f on one of the G_k's, then we are in the line pursued in the previous sections. However, to achieve optimality one may need to use more than one G. Indeed, H. Witsenhausen has solved this determinant extremal problem showing that at most two G_k's are required to achieve optimality, and there are instances where two G_k's are necessary. Even when two G_k's are needed, the w_0 matrix remains a scalar matrix and once again the trace and the determinant optimization are equivalent. The fact that two G_k's are required means the transmitter is excess bandwidth.

Witsenhausen has shown that when two G_k's are needed, one can be taken to be a multiple of $\begin{pmatrix} 1 & j \\ -j & 1 \end{pmatrix}$ and the other a multiple of $\begin{pmatrix} 1 & j \\ j & -1 \end{pmatrix}$. Although both G_k's cannot have the passband form, the $\begin{pmatrix} 1 & j \\ j & -1 \end{pmatrix}$ matrix corresponds to a very simple structural variation of a passband filter.

We mention in closing that systems whose optimization takes us outside the realm of passband structures can be analyzed via equivalent canonical nonexcess bandwidth passband systems. The equivalence is in the sense that MMSE versus P curves for the two systems are identical, and optimum design can be carried out in the canonical system and then transformed to the more complicated system.

IX. ACKNOWLEDGMENTS

We gratefully acknowledge very helpful discussions with R. D. Gitlin, S. Halfin, J. Salz, and H. S. Witsenhausen.

APPENDIX A

Optimality of a Nonexcess Bandwidth System

Fix $P > 0$ and $q(\omega)$ a positive continuous real function on $(-\infty, +\infty)$. In the text we are confronted with the optimization

$$\sup \int_{-\pi/T}^{\pi/T} \log \left[\sum_{k=-\infty}^{\infty} r\left(\omega + \frac{2\pi k}{T}\right) q\left(\omega + \frac{2\pi k}{T}\right) + 1 \right] d\omega,$$

where the sup is over all nonnegative Lebesque integrable $r(\omega)$ for which

$$\int_{-\infty}^{+\infty} r(\omega) d\omega \leqq P > 0.$$

We show here that, under weak conditions on $q(\omega)$, the optimization problem can be replaced by an equivalent "nonexcess bandwidth problem," namely, find

$$\sup \int_{-\pi/T}^{+\pi/T} \log [\tilde{r}(\omega)\tilde{q}(\omega) + 1] d\omega,$$

where $\tilde{q}(\omega)$ is a given continuous function and $\tilde{r}(\omega)$ is any nonnegative integrable function satisfying

$$\int_{-\pi/T}^{\pi/T} \tilde{r}(\omega) d\omega = P > 0.$$

Define $\tilde{q}(\omega)$ on $[-\pi/T, \pi/T]$ to be the envelope sup $q(\omega + 2\pi k/T)$. To avoid annoying pathologies, assume $q(\omega)$ is such that for each ω

$$\left\{k \mid q\left(\omega + \frac{2\pi k}{T}\right) = \tilde{q}(\omega)\right\}$$

is not empty. Moreover, assume that $(-\pi/T, \pi/T)$ can be expressed as a disjoint union of subsets $\{V_m\}_1^m$ of total measure $2\pi/T$ such that on each V_m there exists a k_m so

$$q\left(\omega + \frac{2\pi k_m}{T}\right) = \tilde{q}(\omega)$$

holds uniformly in ω on V_m. So $\tilde{q}(\omega)$ is continuous on $(-\pi/T, \pi/T)$. Define

$$V = \bigcup_1^m \left(V_m + \frac{2\pi k_m}{T}\right).$$

Given any $r(\omega) \geqq 0$ satisfying $\|r\|_1 = P$, define ρ on $(-\infty, \infty)$ by

$$\rho\left(\omega + \frac{2\pi k_m}{T}\right) = \begin{cases} \sum_{-\infty}^{\infty} r\left(\omega + \frac{2\pi k}{T}\right) & \text{for } \omega \in V_m \quad m = 1, 2, \cdots, M \\ 0 & \omega \notin V. \end{cases}$$

So

$$\int_{-\infty}^{\infty} \rho d\omega = \sum_{-\infty}^{\infty} \int_{-\pi/T}^{\pi/T} \rho\left(\omega + \frac{2\pi k}{T}\right) d\omega$$
$$= \int_{-\pi/T}^{\pi/T} \sum_h r\left(\omega + \frac{2\pi k}{T}\right) = \sum_k \int_{-\pi/T}^{\pi/T} r\left(\omega + \frac{2\pi k}{T}\right) d\omega = P,$$

where the second equality results from the definition of ρ and the third equality is from the Lebesque Dominated Convergence Theorem. Now for $|\omega| < \pi/T$

$$\sum_k r\left(\omega + \frac{2\pi k}{T}\right) q\left(\omega + \frac{2\pi k}{T}\right) \leqq \sum_k r\left(\omega + \frac{2\pi k}{T}\right) \tilde{q}(\omega)$$
$$= \tilde{q}(\omega) \sum_k \rho\left(\omega + \frac{2\pi k}{T}\right) = \sum_k \rho\left(\omega + \frac{2\pi k}{T}\right) q\left(\omega + \frac{2\pi k}{T}\right),$$

where the very last equality follows from the fact that ρ vanishes off V. Since in $L[-\pi/T, \pi/T]$ $\rho(\omega)$ always fares at least as well as $r(\omega)$, we have the fact that the supremum can be taken over the class of nonnegative functions vanishing off V.

In the applications it often occurs that $V \subseteq (-\pi/T, \pi/T)$, in which case $q(\omega) = \tilde{q}(\omega)$ on V and the optimization problem becomes

$$\sup_{\|r\| = P} \int_{-\pi/T}^{\pi/T} \log \left[r(\omega) q(\omega) + 1 \right] d\omega.$$

Even when $V \not\subset (-\pi/T, \pi/T)$, we need only solve

$$\sup_{\|\tilde{r}\| = P} \int_{-\pi/T}^{\pi/T} \log \left[\tilde{r}(\omega) \tilde{q}(\omega) + 1 \right] d\omega$$

with the optimand "rearranged" to produce the desired $r(\omega)$. The rearrangement procedure is simply that, for each $\omega \in (-\pi/T, \pi/T)$, we define $r(\omega + 2\pi k_m/T) = \tilde{r}(\omega)$. Elsewhere $r(\omega)$ is defined to be zero. In dealing with even $q(\omega)$, if $V \not\subset (-\pi/T, \pi/T)$, the rearrangement produces an uneven $r(\omega)$. When this occurs, $[r(\omega) + r(-\omega)]/2$ provides an even optimand.

APPENDIX B

Maximization of the Exponent Functional

Let $q(\omega)$ be a continuous positive function on an interval $[a, b]$. Fixing a real number $P > 0$, let Γ be the convex set of nonnegative continuous functions with integral less than or equal to P. We seek γ to maximize the nonlinear functional

$$I(\gamma) \triangleq \int_a^b \log \left(1 + \gamma q\right).$$

This same problem occurs in classical information theory where, for reasons we shall see, it is dubbed "the water-pouring problem." Although the solution is correctly described in the literature, the supporting arguments are formal (for example, see Ref. 16 or 17). We give a rigorous proof here, although our argument is not constructive in that the extremal function is "pulled out of the air." To motivate the extremal function, the reader can turn to the references or supply for himself a variational derivation.

Now $I(\gamma)$ is concave on Γ, as we see by employing the Liebnitz rule to confirm the strict negativity of $I''[\lambda\gamma_1 + (1 - \lambda)\gamma_2]$ on $0 \leqq \lambda \leqq 1$ with γ_1 and γ_2 in Γ (differentiation is with respect to λ). It is clear that if the extremal function exists it has integral equal to P and so we can redefine Γ to require equality of the integrals.

For each constant $\mathcal{C}$, the function $(\mathcal{C} - q^{-1})_+$ denotes the function equal to $\mathcal{C} - q^{-1}$ when $\mathcal{C} - q^{-1} > 0$ and equal to zero otherwise. Now $\int (\mathcal{C} - q^{-1})$ is a continuous strictly increasing function of $\mathcal{C}$ with range

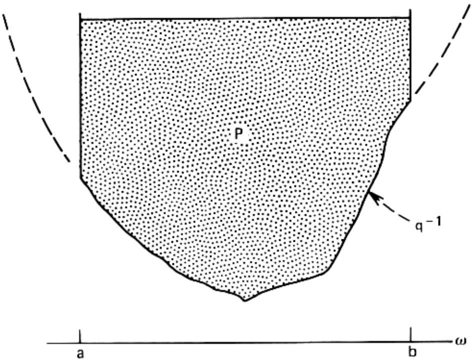


Fig. 7—Optimal power spectral density.

$[0, \infty]$. Fix $\mathcal{C}$ so $\int (\mathcal{C} - q^{-1})_+ = P$ and call the resulting function $\bar{\gamma}$. To show $I(\bar{\gamma})$ is the global maximum of $I(\gamma)$ over Γ, let γ_1 denote any other function in Γ and let us investigate the segment $\{\lambda\bar{\gamma} + (1 - \lambda)\gamma_1, 0 \leqq \lambda \leqq 1\}$. Now $I[\lambda\bar{\gamma} + (1 - \lambda)\gamma_1]$ is concave in λ and straightforwardly

$$I'[\lambda\bar{\gamma} + (1 - \lambda)\gamma_1]|_{\lambda=1} = \mathcal{C}^{-1}\left\{P - \int_{\bar{\gamma}=0} \mathcal{C} q \gamma_1 - \int_{\bar{\gamma}>0} \gamma_1\right\}$$

which is nonnegative as $\mathcal{C}q \leqq 1$. By definition, for a concave function the graph lies above any chord joining two points on the graph. So $\lambda = 1$ must be a point of global maxima of the segment.

Also, $\bar{\gamma}$ is the unique point of maxima since, if there were another point of maxima $\bar{\bar{\gamma}}$, we would have $I(\gamma)$ constant on the line segment joining $\bar{\gamma}$ and $\bar{\bar{\gamma}}$ contradicting the strict negativity of I''.

To understand the water-pouring terminology, look at Fig. 7 where we consider the graph of q^{-1} with vertical walls based at $[a, q^{-1}(a)]$ and $[b, q^{-1}(b)]$ to be a vessel into which water of amount (area) P is poured. Relocate the ω axis to the water level line. Then reflecting the water accumulation about the level line gives the shape of $\bar{\gamma}$.

We mention in closing that $\bar{\gamma}$ is optimal in a larger set than Γ obtained by requiring integrability rather than continuity in the definition of the constraint set. The optimality over the larger set follows from a function space continuity argument.

REFERENCES

1. MacColl, L. A., "Signaling Method and Apparatus," U. S. Patent No. 2,056,284, October 6, 1936.

2. Austin, M. E., "Decision—Feedback Equalization for Digital Communication over Dispersive Channels," Technical Report 461, Research Laboratory of Electronics, M.I.T., August 11, 1967.
3. Keeler, R. J., "Construction and Evaluation of a Decision Feedback Equalizer," Record of IEEE Int. Conf. on Comm. 1971, Montreal, Canada, June 1971, pp. 21–8 to 21–18.
4. George, D. A., Bowen, R. R., and Storey, J. R., "An Adaptive Decision Feedback Equalizer," IEEE Trans. Comm. Tech., *COM-19*, June 1971, pp. 281–293.
5. Monsen, P., "Feedback Equalization for Fading Dispersive Channels," IEEE Trans. on Info. Theory, *IT-17*, January 1971, pp. 56–64.
6. Price, R., "Nonlinearly Feedback-Equalized PAM vs. Capacity for Noisy Filter Channels," Record of IEEE Int. Conf. on Comm. 1972, Philadelphia, Pa., June 1972, pp. 22–12 to 22–17.
7. Salz, J., "Optimum Mean-Square Decision Feedback Equalization," B.S.T.J., *52*, No. 8 (October 1973), pp. 1341–1373.
8. Lucky, R. W., Salz, J., and Weldon, E. S., Jr., *Principles of Data Communication*, New York: McGraw-Hill, 1968.
9. Kobayashi, H., "Simultaneous Adaptive Estimation and Decision Algorithm for Carrier-Modultated Data Transmission Systems," IEEE Trans. Comm. Tech., *COM-19*, October 1971, pp. 835–848.
10. Dugundji, J., "Envelopes and Pre-Envelopes of Real Waveforms," IRE Trans. Info. Theory, *IT-4*, March 1958, pp. 53–57.
11. Wozencraft, J. M., and Jacobs, I. M., *Principles of Communication Engineering*, New York: John Wiley and Sons, 1965, pp. 492–504.
12. Vainberg, M. M., *Variational Methods for the Study of Nonlinear Operators*, San Francisco: Holden-Day, Chapter III.
13. Dunford, N., and Schwartz, J., *Linear Operators, Part III*, New York: Wiley-Interscience, 1971, Chapter XV.
14. Wiener, N., and Akutowicz, E. J., "A Factorization of Positive Hermitian Matrices," J. Math. and Mech., *8*, 1959, pp. 111–120.
15. Doob, J. L., *Stochastic Processes*, New York: John Wiley and Sons, 1967, pp. 160–164.
16. Fano, R. M., *Transmission of Information*, New York: John Wiley and Sons, 1961, pp. 168–178.
17. Walvick, E. A., "On the Capacity of an Ensemble of Channels with Differing Parameters," B.S.T.J., *49*, No. 3 (March 1970), pp. 415–429.