4

GENERALIZED DECISION-FEEDBACK EQUALIZATION FOR PACKET TRANSMISSION WITH ISI AND GAUSSIAN NOISE

J.M. Cioffi* and G.D. Forney, Jr.**

**Information Systems Lab.*
Stanford, CA 94305
email: cioffi@isl.stanford.edu

***Motorola*
Mansfield, MA
email: LUSE27@email.mot.com

Dedicated with respect and deepest regards to Professor Tom Kailath on his sixtieth birthday.

ABSTRACT

A general theory for transmission of finite-length packets over channels with intersymbol interference and additive Gaussian noise is developed. The theory is based on general principles of maximum-likelihood (ML) and linear minimum-mean-squared error (MMSE) estimation, innovations and modal representations of random vectors via Cholesky factorizations, eigendecompositions, and information theory. Using these principles, equivalent forward and backward channel models with desirable properties are developed. Fundamental relations between these theories are presented; for example, the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ between the input $\boldsymbol{X}$ and output $\boldsymbol{Y}$, when $\boldsymbol{X}$ is a Gaussian vector, is equal to $\log\{\|R_{\boldsymbol{x}'\boldsymbol{x}'}\|/\|R_{\boldsymbol{e}'\boldsymbol{e}'}\|\}$, where $\|R_{\boldsymbol{x}'\boldsymbol{x}'}\|$ and $\|R_{\boldsymbol{e}'\boldsymbol{e}'}\|$ are the effective determinants of the covariance matrices of the effective input and of the input linear MMSE estimation error, respectively. A Generalized Decision Feedback Equalization (GDFE) receiver structure is developed and is shown to be canonical for arbitrary linear Gaussian channels- i.e., a reliably transmitted data rate of $I(\boldsymbol{X};\boldsymbol{Y})$ can be approached arbitrarily closely with this receiver structure on any linear Gaussian channel with any input covariance matrix $R_{\boldsymbol{xx}}$. For optimal $R_{\boldsymbol{xx}}$, the performance of this receiver is in aggregate the same as the well-known vector

coding (VC) structure, but in detail the structure is quite different from VC or other previously proposed block DFE receiver structures.

1 INTRODUCTION

In [1], canonical minimum-mean-squared-error decision-feedback equalization (MMSE-DFE) receiver structures for infinite-length sequence transmission have been developed. That paper illustrated an intimate relationship between MMSE-DFE equalization performance and the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ in bits per complex symbol between channel input sequence $\boldsymbol{X}$ and output sequence $\boldsymbol{Y}$, given by the formula

$$I(\boldsymbol{X};\boldsymbol{Y}) = \log_2 \mathrm{SNR}_{\text{MMSE-DFE}} \;, \tag{4.1}$$

where $\mathrm{SNR}_{\text{MMSE-DFE}}$ is the signal-to-noise ratio at the decision point of an MMSE-DFE receiver. From (4.1), it follows that the capacity-achieving transmit power spectrum is the same spectrum that optimizes $\mathrm{SNR}_{\text{MMSE-DFE}}$. Thus, the performance of a MMSE-DFE transmission system[1], with optimized-spectrum transmit signals and powerful coding, can approach the channel capacity of an arbitrary stationary linear-ISI Gaussian sequence channel as closely as capacity can be approached on an ideal Gaussian channel with that same coding – a situation called "canonical" in [1].

In many applications, however, the number of input symbols and output samples is finite; e.g., in point-to-point packet transmission when finite complexity or delay constraints dictate a block structure, or in multi-user packet transmission.

In these applications an appropriate channel model is a finite-dimensional matrix model $\boldsymbol{Y} = H\boldsymbol{X} + \boldsymbol{N}$, where $\boldsymbol{X}$ is a random input m-tuple, H is an $n \times m$ channel-response matrix, and $\boldsymbol{N}$ is an additive Gaussian noise n-tuple. (All quantities are complex.) This paper shows that on such channels the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ in bits per block is

$$I(\boldsymbol{X};\boldsymbol{Y}) = \log_2 |\boldsymbol{SNR}_{\text{GDFE}}| \;, \tag{4.2}$$

where $\boldsymbol{SNR}_{\text{GDFE}}$ is an appropriately defined matrix. Furthermore, it shows that a certain Generalized DFE (GDFE) receiver structure is canonical for such channels.

This paper continues to call a receiver **canonical** if in combination with the same sufficiently powerful coding that approaches capacity on the ISI-free channel, this canonical receiver can achieve arbitrarily low error rates for data rates approaching the value of the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ between channel input and output on the ISI-channel. The mutual information that measures a canonical receiver is computed under the assumption that the input statistics are Gaussian. It should be emphasized that a canonical receiver is not necessarily an optimum receiver, and indeed with no coding or with only moderately powerful coding it may be distinctly inferior to an optimum receiver. The new MMSE-DFE receiver structure of this paper, like that of [1], is constructed using principles of optimum estimation theory, not optimum detection theory, and therefore may be suboptimum when the input sequence is a discrete digital sequence, as it always is in practice. As in [1], the point is that a receiver does not need to do optimum detection to approach channel capacity, when it is used in conjunction with sufficiently powerful codes.

[1] This MMSE-DFE actually can become several parallel MMSE-DFE's, one for each disconnected band of frequencies in the capacity-achieving power spectrum.

1.1 Parallel Channels - a simple illustration of canonical transmission

Suppose H is a square nonsingular $n \times n$ diagonal matrix and $R_{nn} = N_0 I$, then the channel is equivalent to n independent "parallel" subchannels, each with input/output relation $Y_i = H_i X_i + N_i$. The signal to noise ratio on the i^{th} subchannel is $\mathrm{SNR}_i = S_{x,i}|H_i|^2/N_0$ with $S_{x,i}$ the mean-square value for the i^{th} element of the input vector $\boldsymbol{X}$. For each of these parallel subchannels, the mutual information is $log_2(1 + \mathrm{SNR}_i)$ bits per subchannel and for the set of channels, the mutual information is easily determined as [2]

$$I(\boldsymbol{X};\boldsymbol{Y}) = \log_2 \prod_{i=1}^{n} (1 + \mathrm{SNR}_i) \;. \tag{4.3}$$

Each of the subchannels can be independently coded with a powerful code for the ideal additive white Gaussian noise channel so that the data rate achieved is arbitrarily close to the mutual information. The set of such codes and channels then has an aggregate data rate that is the mutual information for the aggregate channel. Figure 4.1 illustrates a set of parallel channels.

The energy (values of $S_{x,i}$ allocated to each subchannel can be determined by a "water-filling" solution [2] and capacity for this block-diagonal-H channel can then be achieved with the same powerful codes that would be used on an ISI-free white-Gaussian noise channel.

While the parallel channels example is trivial, it is also very important in the study of canonical transmission because all the structures that this paper derives for more general H eventually reduce to a set of parallel channels for which the mutual information is the same as the original channel and the same powerful codes that would be used on an AWGN channel can be applied to achieve the highest possible data rates. This paper often uses the example of a one-dimensional channel to illustrate various properties, which can tacitly be inferred to be equivalent to the set of parallel channels.

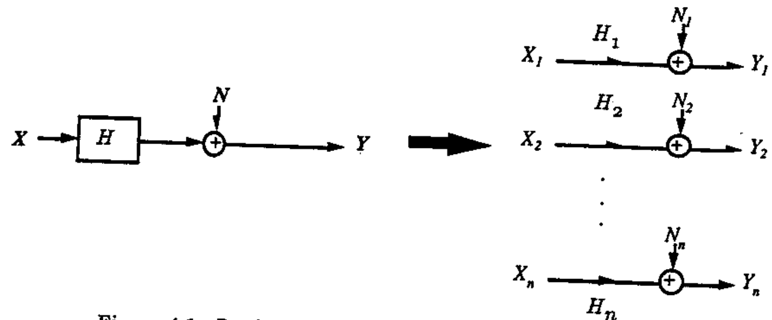

Figure 4.1 Parallel independent channels and equivalent channel.

1.2 More general canonical transmission

Like the well-known vector coding (VC) structure (which is shown to be a special case of the GDFE) [3] -[11], the canonical GDFE structure developed in this paper (which is not the same as the DFE receiver structures of [12] -[17]) effectively decomposes a matrix (block) channel into a number r_y of decoupled one-dimensional Gaussian subchannels with signal-to-noise ratios SNR_j, $1 \leq j \leq r_y$, such that the sum of the component mutual informations $I_j = \log_2(1 + \text{SNR}_j)$ is equal to $I(\boldsymbol{X};\boldsymbol{Y})$. It then follows from the channel coding theorem that for any rate $R_j < I_j$ there exists a discrete (non-Gaussian) code of rate R_j that is capable of achieving arbitrarily low error rates on the j^{th} subchannel. By using such a code on each subchannel, an aggregate rate arbitrarily close to $I(\boldsymbol{X};\boldsymbol{Y})$ bits per block can be transmitted with arbitrarily low probability of error.

More generally than the VC special case of the GDFE, the subchannels in the GDFE receiver are not completely independent, but rather are decoupled by use of the "ideal DFE assumption," which is that the inputs to "past" subchannels are available to the receiver when decoding the current subchannel.

It is shown that the GDFE receiver is canonical even in the general case in which the input covariance matrix $R_{\boldsymbol{xx}}$ does not commute with the channel covariance matrix $H^*R_{\boldsymbol{nn}}^{-1}H$, in which case the vector coding special case is not defined. However, the optimum $R_{\boldsymbol{xx}}$, which is the same for all cases of the GDFE, always commutes with $H^*R_{\boldsymbol{nn}}^{-1}H$.

The set of $\{\text{SNR}_j\}$ also differ. In the limit of large blocks (long packet lengths) and stationary channels, one special case known as the "packet" GDFE receiver approaches the MMSE-DFE receiver structure (or structures for disconnected transmission bands) of [1], Cholesky factorization becomes spectral factorization, and all SNR_j tend to become equal, provided that all subchannels are used. With the vector coding special case, the $\{\text{SNR}_j\}$ are distributed in water-pouring fashion as a function of frequency and vector coding tends to what is

known as multitone transmission [4]. However, the products of the $(1+\text{SNR}_j)$ for the set of subchannels in both cases are the same and equal to $2^{I(\boldsymbol{X};\boldsymbol{Y})}$, as is always the case with any GDFE.

After introducing the general linear Gaussian block channel model, Section 2 discusses modal representations of random vectors based on eigendecompositions and innovations representations (or "Cholesky" factorizations), which are the basic tools used to develop our canonical receivers. It then reviews general principles of linear MMSE estimation. Finally, it discusses additional information-theoretic properties that hold when $\boldsymbol{X}$ is Gaussian.

Section 3 begins by reducing the general channel model without loss of optimality to equivalent forward and backward matrix channel models that have many nice properties: unnecessary dimensions are eliminated, all matrices are square and nonsingular, and the channel-response matrix is equal to the noise covariance matrix. The operation of elimination of unnecessary dimensions is crucial in canonical receivers and asymptotically corresponds as the packet length increases to "symbol rate" optimization and carrier (center) frequency optimization in each used band for the MMSE-DFE. Elimination of unnecessary dimensions also corresponds to selecting good frequency bands for transmission in a vector coding (or multitone) transmission system as packet length increases. The optimum ML and linear MMSE estimators are developed from these models. When the input $\boldsymbol{X}$ is Gaussian, some interesting connections are developed between mutual information and optimal estimation. For example,

$$I(\boldsymbol{X};\boldsymbol{Y}) = \log \|R_{\boldsymbol{x}'\boldsymbol{x}'}\| / \|R_{\boldsymbol{e}'\boldsymbol{e}'}\| , \qquad (4.4)$$

where $\|R_{\boldsymbol{x}'\boldsymbol{x}'}\|$ and $\|R_{\boldsymbol{e}'\boldsymbol{e}'}\|$ are the effective determinants of the covariance matrices of the effective input and of the error in the linear MMSE estimate of the input, respectively. Also,

$$I(\boldsymbol{X};\boldsymbol{Y}) = \log |\boldsymbol{SNR}_{\text{GDFE}}| = \log |I + \boldsymbol{SNR}_{\text{ML}}| , \qquad (4.5)$$

where $\boldsymbol{SNR}_{\text{GDFE}}$ and $\boldsymbol{SNR}_{\text{ML}}$ are matrix generalizations of the usual one-dimensional SNRs for optimum linear MMSE and ML estimation, respectively.

Using an equivalent backward channel model and the "ideal DFE assumption," Section 4 then develops the GDFE receiver structure and shows that it is canonical.

Section 5 addresses the problem of choosing of the input covariance matrix $R_{\boldsymbol{xx}}$ for the GDFE to maximize $I(\boldsymbol{X};\boldsymbol{Y})$, which as is well known is solved by discrete water-pouring. The optimum $R_{\boldsymbol{xx}}$ is shown to commute with the channel covariance matrix $H^*R_{\boldsymbol{nn}}^{-1}H$. Vector coding is well defined in this situation, is also canonical, and uses the same $R_{\boldsymbol{xx}}$ and is a special case of the GDFE where the feedback section disappears.

Section 6 considers the passage to the limit of large blocks (long packets) for stationary Gaussian ISI channels and illustrates that the results of this paper converge to the results in [1] in the limit of infinite-length packets. Further Section 6 illustrates expanded interpretations of the results in [1] where while the MMSE-DFE converges to a stationary structure, there could be several such structures covering only those frequency bands that would also be used by water-pouring transmit optimization – this clearly shows that conventional MMSE-DFE structures such as those considered by Price [18], Salz [19] and others [20] are too generally claimed to be optimum as proposed. However, the necessary modifications (often not understood nor used) to restore optimality are illustrated generally by this paper and in the limit in Section 6.

2 THE BLOCK OR "PACKET" GAUSSIAN ISI CHANNEL

A block (or packet) transmission channel has a finite number of input samples and output samples. Such a channel model is appropriate when a finite-length information packet is transmitted, and detection is based on a finite number of samples of the received signal. Usually, the term packet refers to the situation where the samples are successively indexed in time within a block.

This section begins with a general block Gaussian ISI channel model. Two representations of random vectors are then discussed; in particular, modal representations based on eigendecompositions, and innovations representations based on Cholesky factorizations. These two types of representations are the basis of the canonical receivers to be discussed in this paper. This section progress to discussions of MMSE linear estimation, innovations recursions, and Gaussian random vectors.

2.1 Channel model

On a block Gaussian ISI channel, the received vector of sequence samples $\boldsymbol{Y}$ may be expressed by the matrix equation

$$\boldsymbol{Y} = H\boldsymbol{X} + \boldsymbol{N} \,, \tag{4.6}$$

where $\boldsymbol{X} = \{X_j, 1 \le j \le m\}$ is a complex random input m-vector, $\boldsymbol{Y} = \{Y_k, 1 \le k \le n\}$ is a complex random output n-vector, H is an $n \times m$ complex channel-response matrix, and $\boldsymbol{N}$ is a complex random Gaussian noise n-vector independent of $\boldsymbol{X}$. If $n = m$, the channel is square. All vectors are written as column vectors.

All random vectors, whether discrete, continuous or Gaussian, will be characterized solely by their second-order statistics. The mean of all unconditioned random variables is assumed to be zero, since a nonzero mean costs energy but carries no information. A random vector such as $\boldsymbol{X}$ is then characterized by its covariance matrix

$$R_{\boldsymbol{xx}} = E[\boldsymbol{X}\boldsymbol{X}^*] \,, \tag{4.7}$$

where the asterisk denotes conjugate transpose. The rank of $\boldsymbol{X}$ is the rank r_x of its covariance matrix $R_{\boldsymbol{xx}}$, which is the dimension of the complex vector space $\mathcal{S}_{\boldsymbol{X}}$ in which $\boldsymbol{X}$ takes its values. If $R_{\boldsymbol{xx}}$ is nonsingular, then $\boldsymbol{X}$ has full rank and $r_x = m$, otherwise $r_x < m$.

No restrictions are placed on the input covariance matrix $R_{\boldsymbol{xx}}$ or on the noise covariance matrix $R_{\boldsymbol{nn}}$, except that $\boldsymbol{N}$ is assumed to have full rank, $r_n = n$, so as to avoid noiseless channels of infinite capacity. Similarly, the channel-response matrix H is an arbitrary $n \times m$ complex matrix. The signal component of the output, namely the n-tuple

$$\hat{Y}(\boldsymbol{X}) = H\boldsymbol{X} \,, \tag{4.8}$$

then has covariance matrix $HR_{\boldsymbol{xx}}H^*$. The notation $\hat{Y}(\boldsymbol{X})$ indicates that $\hat{Y}(\boldsymbol{X})$ is the conditional mean of $\boldsymbol{Y}$ given $\boldsymbol{X}$ (see Section 2.3). The vector space $\mathcal{S}_{\hat{\boldsymbol{Y}}}$ of $\hat{Y}(\boldsymbol{X})$ is the image of the input space $\mathcal{S}_{\boldsymbol{X}}$ under the linear transformation H, and therefore the rank $r_{\hat{y}}$ of $\hat{Y}(\boldsymbol{X})$ is not greater than r_x, with equality if and only if the map H acting on $\mathcal{S}_{\boldsymbol{X}}$ is one-to-one. Since $\boldsymbol{X}$ and $\boldsymbol{N}$ are independent, the output covariance matrix is

$$R_{\boldsymbol{yy}} = HR_{\boldsymbol{xx}}H^* + R_{\boldsymbol{nn}} \,. \tag{4.9}$$

Since $\boldsymbol{N}$ has full rank and $HR_{\boldsymbol{xx}}H^*$ is non-negative definite, $\boldsymbol{Y}$ has full rank, $r_y = n$ – however, $r_{\hat{y}} \le \min(n, r_x)$.

2.2 Random vectors and covariance matrix factorizations

This section develops two characteristic representations of random vectors on which our canonical receiver structures will be based. A few preliminary remarks on the geometry of signal spaces may be helpful.

Geometries of vector spaces

There are two kinds of geometry that characterize a random vector such as $\boldsymbol{X}$, and two corresponding inner products:

1. First, there is the **ordinary Euclidean geometry** of the complex vector space $\mathcal{S}_{\boldsymbol{X}}$ in which $\boldsymbol{X}$ takes values. In $\mathcal{S}_{\boldsymbol{X}}$. the inner product of two ordinary ("deterministic") complex column vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is the ordinary Hermitian dot product:

$$\boldsymbol{x}^*\boldsymbol{y} = \sum_i x_i^* y_i \,, \tag{4.10}$$

where, as always in this paper, the asterisk denotes conjugate transpose. In ordinary Euclidean geometry the squared norm of a vector $\boldsymbol{x}$ is the usual Euclidean squared norm $\|\boldsymbol{x}\|^2$, namely the sum of the squared magnitudes $|x_i|^2$ of the components x_i, and two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal if their dot product $\boldsymbol{x}^*\boldsymbol{y}$ is zero.

2. Second, there is the **geometry of Hilbert spaces** of complex random variables, in which the inner product of two complex random variables X and Y is defined by their Hermitian cross-correlation

$$< X, Y >= E[XY^*] \,. \tag{4.11}$$

In Hilbert-space geometry the squared norm of a zero-mean random variable X is its variance $E[|X|^2]$, and two random variables X and Y are orthogonal if they are uncorrelated, $E[XY^*] = 0$.

The set $\{X_i\}$ of components of a random vector $\boldsymbol{X}$ generate a Hilbert space $V(\boldsymbol{X})$ consisting of all complex linear combinations

$$\sum_i a_i^* X_i = \boldsymbol{a}^* \boldsymbol{X} \tag{4.12}$$

of elements of $\boldsymbol{X}$. The inner product of two elements $\boldsymbol{a}^*\boldsymbol{X}$, $\boldsymbol{b}^*\boldsymbol{X} \in V(\boldsymbol{X})$ is

$$< \boldsymbol{a}^*\boldsymbol{X},\ \boldsymbol{b}^*\boldsymbol{X} >= E[\boldsymbol{a}^*\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{b}] = \boldsymbol{a}^* R_{\boldsymbol{xx}} \boldsymbol{b} \,. \tag{4.13}$$

Thus the geometry of $V(\boldsymbol{X})$, which is characterized by the set of inner products between any two of its vectors, is entirely determined by the covariance matrix $R_{\boldsymbol{xx}}$, which is the matrix of inner products (Gram matrix) of elements of $\boldsymbol{X}$.

Charateristic Representations

Characteristic representations will enable the design of canonical receivers:

Definition 1 (Characteristic representation of a random vector) *A* **characteristic representation** *of a random m-tuple* $\boldsymbol{X}$ *is a linear combination*

$$\boldsymbol{X} = F\boldsymbol{V} = \sum_j V_j \boldsymbol{f}_j \,, \tag{4.14}$$

where $\{V_j\}$ is a set of uncorrelated random variables–i.e., the covariance matrix $R_{\boldsymbol{vv}}$ is diagonal- and F is a square matrix with determinant $|F| = 1$.

The covariance matrix of $\boldsymbol{X}$ is then

$$R_{\boldsymbol{xx}} = E[F\boldsymbol{V}\boldsymbol{V}^*F^*] = FR_{\boldsymbol{vv}}F^* \,. \tag{4.15}$$

Thus characteristic representations of the form $\boldsymbol{X} = F\boldsymbol{V}$ are closely related to covariance matrix factorizations of the form $R_{\boldsymbol{xx}} = FR_{\boldsymbol{vv}}F^*$, where $|F| = 1$ and $R_{\boldsymbol{vv}}$ is diagonal. Indeed, given such a factorization, define $\boldsymbol{V} = F^{-1}\boldsymbol{X}$; then $\boldsymbol{V}$ has the diagonal covariance matrix $R_{\boldsymbol{vv}}$ that occurs in the factorization and $\boldsymbol{X} = F\boldsymbol{V}$.

Since F is nonsingular, it is rank-preserving; i.e., the rank of $\boldsymbol{V}$ is equal to the rank of $\boldsymbol{X}$, $r_v = r_x$, which implies that precisely r_x of the random variables V_j are not identically zero. Since every element $\boldsymbol{a}^*\boldsymbol{X}$ of $V(\boldsymbol{X})$ is a linear combination of these rx nonzero random variables V_j via

$$\boldsymbol{a}^*\boldsymbol{X} = \boldsymbol{a}^*F\boldsymbol{V} \,, \tag{4.16}$$

it follows that $V(\boldsymbol{X}) = V(\boldsymbol{V})$ and that these r_x nonzero random variables V_j form an orthogonal basis for $V(\boldsymbol{X})$, whose dimension is thus also equal to r_x. The r_x corresponding complex vectors $\boldsymbol{f}_j$ generate the deterministic r_x-dimensional Euclidean space $\mathcal{S}_{\boldsymbol{X}}$, although they are not necessarily orthogonal in $\mathcal{S}_{\boldsymbol{X}}$.

Finally, the unimodular condition $|F| = 1$ implies that F and its inverse F^{-1} are volume-preserving transformations, provided that $\boldsymbol{X}$ has full rank. In other words, F is determinant-preserving:

$$|R_{\boldsymbol{xx}}| = |F||R_{\boldsymbol{vv}}||F^*| = |R_{\boldsymbol{vv}}| \,. \tag{4.17}$$

However, if $\boldsymbol{X}$ does not have full rank, then F is not necessarily a volume-preserving transformation from the r_x-dimensional subspace $\mathcal{S}_{\boldsymbol{V}}$ that supports the r_x nonzero random variables $\{V_j\}$ to the r_x-dimensional subspace $\mathcal{S}_{\boldsymbol{X}}$ that supports the random vector $\boldsymbol{X}$. There are two types of characteristic representations of interest:

Modal representations

A covariance matrix $R_{\boldsymbol{xx}}$ is square, Hermitian-symmetric, and nonnegative definite. Such a matrix has a (nonunique, in general) eigendecomposition

$$R_{\boldsymbol{xx}} = U\Lambda_x^2 U^* = (U\Lambda_x)(U\Lambda_x)^* \,, \tag{4.18}$$

where U is a unitary matrix ($UU^* = U^*U = I$, so $U^{-1} = U^*$ and $|U| = 1$) and Λ_x^2 is a nonnegative real diagonal matrix whose diagonal elements are the eigenvalues of $R_{\boldsymbol{xx}}$. The set of eigenvalues is invariant in any eigendecomposition. The last expression shows that $U\Lambda_x$ may be regarded as a square root of $R_{\boldsymbol{xx}}$.

Correspondingly, if the modal variables $\boldsymbol{M}$ are defined by

$$\boldsymbol{M} = U^{-1}\boldsymbol{X} = U^*\boldsymbol{X} \,, \tag{4.19}$$

then $R_{\boldsymbol{mm}} = U^*R_{\boldsymbol{xx}}U = \Lambda_x^2$ and $\boldsymbol{X} = U\boldsymbol{M}$, where $|U| = 1$. Thus any eigendecomposition of $R_{\boldsymbol{xx}}$ leads to a characteristic representation of $\boldsymbol{X}$, called a **modal representation.**

Since the columns u_j of a unitary matrix U are orthonormal, a modal representation

$$\boldsymbol{X} = U\boldsymbol{M} = \sum_j M_j u_j \ , \tag{4.20}$$

has the desirable property that the r_x vectors u_j corresponding to the r_x nonzero modal variables M_j form an orthonormal basis for $\mathcal{S}_{\boldsymbol{X}}$; i.e., both kinds of orthogonality occur in a modal decomposition.

Consequently, a unitary transformation is length-preserving; that is,

$$\|U\boldsymbol{m}\|^2 = \boldsymbol{m}^* U^* U \boldsymbol{m} = \boldsymbol{m}^* \boldsymbol{m} = \|\boldsymbol{m}\|^2 \ . \tag{4.21}$$

A fortiori, U is volume-preserving regardless of whether $\boldsymbol{X}$ has full rank.

Example 4.1 (Modal Representation Example) Let $\boldsymbol{X}$ be a random vector $[X_1, X_2]^*$ with covariance matrix

$$R_{\boldsymbol{xx}} = \begin{bmatrix} a & b \\ b & a \end{bmatrix} , \tag{4.22}$$

where a and b are real and $0 \le |b| \le a$. Then, $|R_{\boldsymbol{xx}}| = a^2 - b^2$, the eigenvalues of $R_{\boldsymbol{xx}}$ are $a+b$ and $a-b$, its eigenvectors are $\frac{1}{\sqrt{2}}[11]^*$ and $\frac{1}{\sqrt{2}}[-11]^*$, and its rank r_x is 2 unless $|b| = a$, when $r_x = 1$. An eigendecomposition of $R_{\boldsymbol{xx}}$ is thus

$$R_{\boldsymbol{xx}} = \begin{bmatrix} 2^{-1/2} & -2^{-1/2} \\ 2^{-1/2} & 2^{-1/2} \end{bmatrix} \begin{bmatrix} a+b & 0 \\ 0 & a-b \end{bmatrix} \begin{bmatrix} 2^{-1/2} & 2^{-1/2} \\ -2^{-1/2} & 2^{-1/2} \end{bmatrix} \tag{4.23}$$

and a modal representation of $\boldsymbol{X}$ is

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 2^{-1/2} & -2^{-1/2} \\ 2^{-1/2} & 2^{-1/2} \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} , \tag{4.24}$$

where $M_1 = \frac{1}{\sqrt{2}}(X_1 + X_2)$ has variance $a+b$, $M_2 = \frac{1}{\sqrt{2}}(X_2 - X_1)$ has variance $a-b$ and M_1 and M_2 are uncorrelated. If $b = a$, then $X_1 = X_2$ and $M_1 = \sqrt{2}X_1$, $M_2 = 0$, whereas if $b = -a$, then $X_1 = -X_2$ and $M_1 = 0$, $M_2 = \sqrt{2}X_1$. Note that if $b = 0$ then $R_{\boldsymbol{xx}} = U(aI)U^*$ for any 2×2 unitary matrix U, so there is a family of eigendecompositions of which the one given above is only one member.

Innovations representations

Alternatively, a covariance matrix $R_{\boldsymbol{xx}}$ has a unique factorization of the form

$$R_{\boldsymbol{xx}} = LD_x^2 L^* = (LD_x)(LD_x)^* \ , \tag{4.25}$$

where L is a lower triangular matrix that is monic (i.e., which has ones on the diagonal, so $|L| = 1$) and D_x^2 is diagonal. This factorization is called the

Cholesky factorization of $R_{\boldsymbol{xx}}$, and the diagonal elements of D_x^2 (which must be real and nonnegative, with r_x of them nonzero) are called the **Cholesky factors** of $R_{\boldsymbol{xx}}$. The matrix LD_x is another square root of $R_{\boldsymbol{xx}}$.

Correspondingly, the innovations variables $\boldsymbol{W}$ are

$$\boldsymbol{W} = L^{-1}\boldsymbol{X} \ , \tag{4.26}$$

and $R_{\boldsymbol{ww}} = L^{-1} R_{\boldsymbol{xx}} L^{-*} = D_x^2$ and $\boldsymbol{X} = L\boldsymbol{W}$, where $|L| = 1$. (Here L^{-*} denotes $(L^{-1})^* = (L^*)^{-1}$.) Thus, the Cholesky factorization of $R_{\boldsymbol{xx}}$ leads to a unique characteristic representation of $\boldsymbol{X}$, called the **innovations representation.**

Since L is lower triangular, the innovations representation

$$\boldsymbol{X} = L\boldsymbol{W} = \sum_j W_j l_j \ , \tag{4.27}$$

has the desirable property that, for any k, the first k components of $\boldsymbol{X}$ depend only on the first k components of $\boldsymbol{W}$ (and, since L^{-1} is also lower triangular, vice versa). From a dynamical point of view, an innovations representation thus has a kind of causality property, which is important when the sequential ordering of the components of $\boldsymbol{X}$ is important. Also, in matrix terms, this property implies that a Cholesky factorization has a nesting property that leads to recursive implementations. Again, the r_x columns l_j corresponding to the r_x nonzero innovations variables W_j span $\mathcal{S}_{\boldsymbol{X}}$, although they are not in general orthogonal.

The Cholesky factorization of $R_{\boldsymbol{xx}}$ and corresponding innovations representation of $\boldsymbol{X}$ depend very much on the ordering of the components of $\boldsymbol{X}$. If $\boldsymbol{X}'$ is a permutation of $\boldsymbol{X}$, then the innovations representation of $\boldsymbol{X}'$ and its Cholesky factors will be different (although because of the invariance of the effective determinant, the product of the nonzero Cholesky factors will be unchanged). In particular, if $\boldsymbol{X}'$ is the reversal of $\boldsymbol{X}$, then the Cholesky factorization of $R_{\boldsymbol{x}'\boldsymbol{x}'}$ can be permuted to give an upper-diagonal-lower factorization of $R_{\boldsymbol{xx}}$ of the form

$$R_{\boldsymbol{xx}} = (L')^*(D_x')^2 L' \ , \tag{4.28}$$

where $(L')^*$ is upper triangular, and a corresponding reverse innovations representation of $\boldsymbol{X}$ is then obtained:

$$\boldsymbol{X} = (L')^* \boldsymbol{W}' \ . \tag{4.29}$$

Example 4.2 (Innovations Representation Example) Again let $\boldsymbol{X}$ be a random vector $[X_1, X_2]^*$ with covariance matrix

$$R_{\boldsymbol{xx}} = \begin{bmatrix} a & b \\ b & a \end{bmatrix} , \tag{4.30}$$

with a, b real and $0 \le |b| \le a$. Then the unique Cholesky decomposition of $R_{\boldsymbol{xx}}$ is

$$R_{\boldsymbol{xx}} = \begin{bmatrix} 1 & 0 \\ b/a & 1 \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & (a^2 - b^2)/a \end{bmatrix} \begin{bmatrix} 1 & b/a \\ 0 & 1 \end{bmatrix} , \quad (4.31)$$

and an innovations representation of $\boldsymbol{X}$ is

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ b/a & 1 \end{bmatrix} \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} , \quad (4.32)$$

where $W_1 = X_1$ has variance a, $W_2 = X_2 - (b/a)X_1$ has variance $(a^2 - b^2)/a$, and W_1 and W_2 are uncorrelated. If $b = a$, then $X_1 = X_2$ and $W_1 = X_1$, $W_2 = 0$, whereas if $b = -a$, then $X_1 = -X_2$ and again $W_1 = X_1$, $W_2 = 0$. Note that when $\boldsymbol{X}$ does not have full rank, this map between the one-dimensional spaces $\mathcal{S}_{\boldsymbol{W}}$ and $\mathcal{S}_{\boldsymbol{X}}$ is not volume- (length-) preserving. But note that even when $b = 0$, the Cholesky decomposition is unique.

2.3 MMSE linear estimation

Suppose that the Hilbert space $V(\boldsymbol{X})$ generated by the elements of $\boldsymbol{X}$ is a subspace of a larger Hilbert space $V(\boldsymbol{X})^+$, and that the complex scalar Y is a random variable in $V(\boldsymbol{X})^+$. Then by the projection theorem, the closest variable to Y in $V(\boldsymbol{X})$ is the projection of Y onto $V(\boldsymbol{X})$, denoted by $Y_{|\boldsymbol{x}}$.

By the orthogonality principle, the projection $Y_{|\boldsymbol{x}}$ is the unique element of $V(\boldsymbol{X})$ such that the estimation error

$$E = Y - Y_{|\boldsymbol{x}} \quad (4.33)$$

is orthogonal to (uncorrelated with) all elements of $V(\boldsymbol{X})$, or equivalently to all elements X_i of $\boldsymbol{X}$. Since $Y_{|\boldsymbol{x}}$ is some linear combination of elements of $\boldsymbol{X}$, $Y_{|\boldsymbol{x}} = \boldsymbol{a}^*\boldsymbol{X}$, this implies that for all i

$$< X_i, E > = < X_i, Y > - < X_i, \boldsymbol{a}^*\boldsymbol{X} > = < X_i, Y > - < X_i, \boldsymbol{X} > \boldsymbol{a} = 0 . \quad (4.34)$$

Equation (4.34) for all i may be written as a matrix equation

$$< \boldsymbol{X}, E > = < \boldsymbol{X}, Y > - < \boldsymbol{X}, \boldsymbol{X} > \boldsymbol{a} = \boldsymbol{r}_{\boldsymbol{x}Y} - R_{\boldsymbol{xx}}\boldsymbol{a} = 0 , \quad (4.35)$$

where $\boldsymbol{r}_{\boldsymbol{x}Y} = < \boldsymbol{X}, Y >$ is the column vector with components $< X_i, Y > = E[X_i Y^*]$, and $R_{\boldsymbol{xx}}$ is the covariance matrix $< \boldsymbol{X}, \boldsymbol{X} > = E[\boldsymbol{X}\boldsymbol{X}^*]$. When $R_{\boldsymbol{xx}}$ is nonsingular, this determines a unique solution for $\boldsymbol{a}$:

$$\boldsymbol{a} = R_{\boldsymbol{xx}}^{-1}\boldsymbol{r}_{\boldsymbol{x}Y} . \quad (4.36)$$

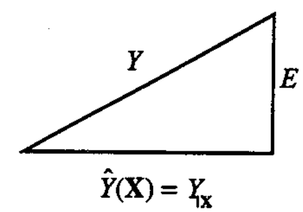

Figure 4.2 Orthogonality of MMSE linear estimate and error.

More generally, $\boldsymbol{a}$ may be uniquely determined by using an orthogonal basis for $V(\boldsymbol{X})$ with r_x elements, as discussed in Section 2.2.

To say that $Y_{|\boldsymbol{x}}$ is the closest variable to Y in $V(\boldsymbol{X})$ is to say that the variance of the difference variable $E = Y - Y_{|\boldsymbol{x}}$ is a minimum over all linear combinations of elements of $\boldsymbol{X}$. Therefore $Y_{|\boldsymbol{x}}$ is called the minimum-mean-squared-error (MMSE) linear estimate of Y given $\boldsymbol{X}$, and is alternatively denoted by

$$\hat{Y}(\boldsymbol{X}) = Y_{|\boldsymbol{x}} . \quad (4.37)$$

From the above development, any random variable Y may be written uniquely as

$$Y = \hat{Y}(\boldsymbol{X}) + E , \quad (4.38)$$

where $\hat{Y}(\boldsymbol{X})$ is in $V(\boldsymbol{X})$ and E is orthogonal to all variables in $V(\boldsymbol{X})$. This is illustrated by the right triangle of Figure 4.2. By the Pythagorean theorem for Hilbert spaces, the variance of Y is the sum of the variances of $\hat{Y}(\boldsymbol{X})$ and E. The estimation error variable E is zero if and only if $Y \in V(\boldsymbol{X})$. Since the mean of E is zero and E is orthogonal to $\boldsymbol{X}$, $\hat{Y}(\boldsymbol{X})$ is the conditional mean of Y given $\boldsymbol{X}$.

The above development generalizes straightforwardly to a set $\boldsymbol{Y} = \{Y_j\}$ of random variables Y_j. The MMSE linear estimate of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is the vector

$$\hat{\boldsymbol{Y}}(\boldsymbol{X}) = \boldsymbol{Y}_{|\boldsymbol{x}} \quad (4.39)$$

of MMSE linear estimates $\hat{Y}_j(\boldsymbol{X}) = \boldsymbol{a}_j^*\boldsymbol{X}$, so $\hat{\boldsymbol{Y}}(\boldsymbol{X}) = A^*\boldsymbol{X}$ for some matrix A. The components E_j of the estimation error vector

$$\boldsymbol{E} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}(\boldsymbol{X}) \quad (4.40)$$

are each orthogonal to all components of $\boldsymbol{X}$, and thus $\boldsymbol{E}$ satisfies the matrix equation

$$< \boldsymbol{X}, \boldsymbol{E} > = < \boldsymbol{X}, \boldsymbol{Y} > - < \boldsymbol{X}, \boldsymbol{X} > A = R_{\boldsymbol{xy}} - R_{\boldsymbol{xx}}A = 0 , \quad (4.41)$$

which yields the solution $A = R_{\boldsymbol{xx}}^{-1} R_{\boldsymbol{xy}}$ when $R_{\boldsymbol{xx}}$ is nonsingular. [2]

The orthogonality illustrated in Figure 4.2 continues to hold, since $\hat{\boldsymbol{Y}}(\boldsymbol{X})$ is a vector of elements of $V(\boldsymbol{X})$, while $\boldsymbol{E}$ is is vector of elements that are orthogonal to $V(\boldsymbol{X})$. However, the "Pythagorean theorem" now becomes

$$R_{\boldsymbol{yy}} = A^* R_{\boldsymbol{xx}} A + R_{\boldsymbol{ee}} \; ; \tag{4.42}$$

i.e., the covariance matrix of the diagonal is the sum of the covariance matrices of the two sides of the right triangle.

The covariance matrix $R_{\boldsymbol{ee}}$ of the minimum mean square linear estimation error $\boldsymbol{E}$ is minimum in every sense. Let $\hat{\boldsymbol{Y}}'(\boldsymbol{X}) = B^* \boldsymbol{X}$ be an arbitrary linear estimate of $\boldsymbol{Y}$ given $\boldsymbol{X}$, and let $\boldsymbol{E}' = \boldsymbol{Y} - \hat{\boldsymbol{Y}}'(\boldsymbol{X})$ be the corresponding error vector. Then since $\boldsymbol{Y} = A^* \boldsymbol{X} + \boldsymbol{E}$, it follows that $\boldsymbol{E}'$ has the orthogonal decomposition

$$\boldsymbol{E}' = (A^* - B^*)\boldsymbol{X} + \boldsymbol{E} = C^* \boldsymbol{X} + \boldsymbol{E} \; , \tag{4.43}$$

where $C^* \boldsymbol{X}$ is in $V(\boldsymbol{X})$ and $\boldsymbol{E}$ is orthogonal to $V(\boldsymbol{X})$. Consequently

$$R_{\boldsymbol{e}'\boldsymbol{e}'} = C^* R_{\boldsymbol{xx}} C + R_{\boldsymbol{ee}} \; , \tag{4.44}$$

where both $C^* R_{\boldsymbol{xx}} C$ and $R_{\boldsymbol{ee}}$ are nonnegative definite covariance matrices. It follows that $R_{\boldsymbol{ee}}$ is "less than" $R_{\boldsymbol{e}'\boldsymbol{e}'}$ in every sense; its determinant is less, its trace is less, its eigenvalues are less, its Cholesky factors are less, and so forth. For any vector $\boldsymbol{a}$, the variance of the linear combination $\boldsymbol{a}^* \boldsymbol{E}'$ is not less than that of $\boldsymbol{a}^* \boldsymbol{E}$, since

$$E[\boldsymbol{a}^* \boldsymbol{E}'(\boldsymbol{E}')^* \boldsymbol{a}] = \boldsymbol{a}^* R_{\boldsymbol{e}'\boldsymbol{e}'} \boldsymbol{a} \geq \boldsymbol{a}^* R_{\boldsymbol{ee}} \boldsymbol{a} \; , \tag{4.45}$$

by the nonnegative definiteness of $C^* R_{\boldsymbol{xx}} C$. Indeed, the nonnegative definiteness of a Hermitian-symmetric square matrix A is sometimes denoted by $A \geq 0$; in this notation, one may write

$$R_{\boldsymbol{e}'\boldsymbol{e}'} - R_{\boldsymbol{ee}} \geq 0 \; , \text{ or} \tag{4.46}$$

$$R_{\boldsymbol{e}'\boldsymbol{e}'} \geq R_{\boldsymbol{ee}} \; . \tag{4.47}$$

It follows that for any optimality criterion based on error variances, the vector MMSE linear estimate is optimum among all linear estimators.

2.4 Innovations representations via recursive MMSE prediction

The innovations representation of a random vector $\boldsymbol{X}$ may be developed by sequential MMSE linear prediction. Let $\boldsymbol{X}(j-1)$ denote the "past" relative to

[2]The inverse may be replaced by any one of many generalized inverses when $R_{\boldsymbol{xx}}$ is singular, see [21].

a component X_j of $\boldsymbol{X}$; i.e.,

$$\boldsymbol{X}(j-1) = \{X_k | k < j\} . \tag{4.48}$$

The MMSE linear prediction of X_j given $\boldsymbol{X}(j-1)$ is then the projection $X_j|_{\boldsymbol{X}(j-1)}$, and the j^{th} innovations variable W_j may then be defined as the prediction error

$$W_j = X_j - X_j|_{\boldsymbol{X}(j-1)} \; . \tag{4.49}$$

By the orthogonality principle, W_j is orthogonal to the past space $V(\boldsymbol{X}(j-1))$; however, $V(\boldsymbol{X}(j-1))$ and W_j together span $V(\boldsymbol{X}(j))$. It follows that $V(\boldsymbol{W}(j)) = V(\boldsymbol{X}(j))$, and thus that the elements of $\boldsymbol{W}$ are orthogonal (uncorrelated). An innovations variable is zero if and only if it is in the past space $V(\boldsymbol{X}(j-1))$. Since $X_j|_{\boldsymbol{X}(j-1)}$ may be expressed as a linear combination of the elements either of $\boldsymbol{X}(j-1)$ or of $\boldsymbol{W}(j-1)$, the prediction error equations may be expressed in matrix form as either

$$\boldsymbol{W} = L^{-1} \boldsymbol{X} \; , \tag{4.50}$$

or

$$\boldsymbol{X} = L \boldsymbol{W} \; , \tag{4.51}$$

where L and L^{-1} are both lower triangular and monic. Then

$$R_{\boldsymbol{xx}} = L R_{\boldsymbol{ww}} L^* \; , \tag{4.52}$$

is the Cholesky factorization of $R_{\boldsymbol{xx}}$ since such a factorization is unique.

2.5 Gaussian random vectors

Heretofore random vectors $\boldsymbol{X}$ have not been assumed to be Gaussian. However, Gaussian random vectors have particularly nice properties. In particular, information-theoretic quantities are simple functions of the second-order statistics (covariance matrices) of Gaussian random vectors. Therefore in a model in which only second-order statistics are given, it is often helpful to analyze the case in which all variables are Gaussian; this usually simplifies the analysis and yields structures and bounds that are useful for the general case.

The probability distribution of a zero-mean complex Gaussian random vector $\boldsymbol{X}$ is completely determined by its covariance matrix $R_{\boldsymbol{xx}}$. If $R_{\boldsymbol{xx}}$ is nonsingular, then

$$p_{\boldsymbol{x}}(\boldsymbol{X}) = \pi^{-r_x} |R_{\boldsymbol{xx}}|^{-1} e^{-\boldsymbol{X}^* R_{\boldsymbol{xx}}^{-1} \boldsymbol{X}} \; . \tag{4.53}$$

The separability property of this distribution implies that uncorrelated Gaussian random variables are independent.

More generally, as shown in Section 2.2, given $R_{\boldsymbol{xx}}$, a Gaussian vector $\boldsymbol{X}$ may be expressed as a linear combination $\boldsymbol{X} = F\boldsymbol{V}$ of r_x nonzero uncorrelated and thus independent Gaussian random variables V_j. If F is unitary, then this map from $\mathcal{S}_{\boldsymbol{V}}$ to $\mathcal{S}_{\boldsymbol{X}}$ is volume-preserving.

If $\boldsymbol{Y}$ and $\boldsymbol{X}$ are jointly Gaussian, then it is straightforward to show that the MMSE linear estimate $\hat{\boldsymbol{Y}}(\boldsymbol{X})$ is actually the unconstrained MMSE estimate of $\boldsymbol{Y}$ given $\boldsymbol{X}$, since $\boldsymbol{Y}$ may be written as

$$\boldsymbol{Y} = \hat{\boldsymbol{Y}}(\boldsymbol{X}) + \boldsymbol{E} , \tag{4.54}$$

where $\boldsymbol{E}$ is a Gaussian random vector that is independent of $\boldsymbol{X}$.

As shown in [2], the differential entropy of a complex Gaussian vector $\boldsymbol{X}$ of rank r_x with nonsingular covariance matrix $R_{\boldsymbol{xx}}$ is

$$h(\boldsymbol{X}) = r_x \log_2 \pi e |R_{\boldsymbol{xx}}|^{1/r_x} . \tag{4.55}$$

More generally, since the differential entropy is invariant under volume-preserving transformations, and a modal representation $\boldsymbol{X} = U\boldsymbol{M}$ is volume-preserving regardless of whether $\boldsymbol{X}$ has full rank, the differential entropy $h(\boldsymbol{X})$ is equal to $h(\boldsymbol{M})$, where $\boldsymbol{M}$ is a set of independent complex Gaussian variables M_j with variances λ_j^2 equal to the eigenvalues of $R_{\boldsymbol{xx}}$. Thus

$$h(\boldsymbol{X}) = h(\boldsymbol{M}) = \sum_{j \in J} \log_2 \pi e \lambda_j^2 , \tag{4.56}$$

where the sum is only over the set $J = \{j \mid \lambda_j^2 > 0\}$ of r_x indices corresponding to the r_x nonzero eigenvalues of $R_{\boldsymbol{xx}}$. In other words,

$$h(\boldsymbol{M}) = r_x \log_2 \pi e \|R_{\boldsymbol{mm}}\|^{1/r_x} , \tag{4.57}$$

where $\|R_{\boldsymbol{mm}}\|$ is the **effective determinant** of the diagonal covariance matrix $R_{\boldsymbol{mm}}$:

Definition 2 (Effective Determinant) *The effective determinant of a matrix is the product of its nonzero eigenvalues,*

$$\|R_{\boldsymbol{mm}}\| = \prod_{j \in J} \lambda_j^2 . \tag{4.58}$$

Note that $\|R_{\boldsymbol{mm}}\|^{1/r_x}$ is the geometric mean of the nonzero eigenvalues of $R_{\boldsymbol{xx}}$.

Since $\|R_{\boldsymbol{mm}}\|$ is the product of the nonzero eigenvalues of $R_{\boldsymbol{xx}}$, $\|R_{\boldsymbol{mm}}\|$ is invariant in any modal representation of $\boldsymbol{X}$. Therefore $\|R_{\boldsymbol{mm}}\| = \|R_{\boldsymbol{xx}}\|$, and the differential entropy of $\boldsymbol{X}$ is equal to

$$h(\boldsymbol{X}) = r_x \log_2 \pi e \|R_{\boldsymbol{xx}}\|^{1/r_x} . \tag{4.59}$$

Example 4.3 (Two-Dimensional Example continued) Again let $\boldsymbol{X}$ be a random vector $[X_1^*, X_2^*]^*$ with covariance matrix

$$R_{\boldsymbol{xx}} = \begin{bmatrix} a & b \\ b & a \end{bmatrix} , \tag{4.60}$$

with a, b real and $0 \le |b| \le a$. The eigenvalues of $R_{\boldsymbol{xx}}$ are $(a+b, a-b)$, and the rank r_x is 2 unless $|b| = a$. The effective determinant of $R_{\boldsymbol{xx}}$ is thus equal to

$$\|R_{\boldsymbol{xx}}\| = \begin{cases} |R_{\boldsymbol{xx}}| = a^2 - b^2, & \text{if } |b| < a \\ 2a, & \text{if } |b| = a . \end{cases} \tag{4.61}$$

Note that the effective determinant is equal to the product of the Cholesky factors of $R_{\boldsymbol{xx}}$ when $\boldsymbol{X}$ has full rank, but not when $r_x = 1$. Note also that there is a discontinuity in the differential entropy $h(\boldsymbol{X})$ as $|b| \to a$. This discontinuity often occurs when $R_{\boldsymbol{xx}}$ is optimized as in Section 6. These discontinuities leads to "symbol-rate" and "center-frequency" optimization for each used frequency band in the stationary case.

The differential entropy of any random vector $\boldsymbol{X}$ with covariance matrix $R_{\boldsymbol{xx}}$ is upperbounded by the differential entropy of a Gaussian vector with the same covariance matrix:

$$h(\boldsymbol{X}) \le r_x \log_2 \pi e \|R_{\boldsymbol{xx}}\|^{1/r_x} , \tag{4.62}$$

with equality if and only if $\boldsymbol{X}$ is Gaussian. The maximum entropy inference principle therefore suggests that if only the second-order statistics of $\boldsymbol{X}$ are known, then $\boldsymbol{X}$ should be presumed to be Gaussian. The effective determinant $\|R_{\boldsymbol{xx}}\|$ determines the differential entropy $h(\boldsymbol{X})$ of this presumed Gaussian density.

Since the mutual information between the input $\boldsymbol{X}$ and output $\boldsymbol{Y} = H\boldsymbol{X} + \boldsymbol{N}$ of a Gaussian ISI channel may be written as

$$I(\boldsymbol{X}; \boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}/\boldsymbol{X}) , \tag{4.63}$$

and the conditional differential entropy $H(\boldsymbol{Y}/\boldsymbol{X})$ is equal to $h(\boldsymbol{N})$, it follows that the mutual information is maximized for a given $R_{\boldsymbol{yy}}$ when $\boldsymbol{Y}$ is Gaussian, which in turn occurs when $\boldsymbol{X}$ is Gaussian.

These information-theoretic relations can be used to develop many determinantal inequalities, as shown by Cover and Thomas [2]. For example, Hadamard's inequality, which will be needed below, states that if R is a covariance matrix (a square Hermitian-symmetric nonnegative-definite matrix), then

$$|R| \le \prod_j R_{jj} , \tag{4.64}$$

with equality if and only if R is diagonal. For, suppose that $\boldsymbol{X}$ is a Gaussian random vector with covariance matrix R; then Hadamard's inequality follows from the information-theoretic inequality

$$h(\boldsymbol{X}) = \sum_j h(X_j|X_1, ..., X_{j-1}) \leq \sum_j h(X_j) , \qquad (4.65)$$

where equality holds if and only if the components X_j of $\boldsymbol{X}$ are independent.

3 EQUIVALENT CHANNEL MODELS, LINEAR ESTIMATION, AND MUTUAL INFORMATION

In this section, given a linear Gaussian channel model $\boldsymbol{Y} = H\boldsymbol{X}+\boldsymbol{N}$, equivalent forward and backward channel models that eliminate singularities and have many other nice properties are developed. Using these equivalent models, a number of relations are obtained between ML estimation, MMSE estimation, and mutual information (when $\boldsymbol{X}$ is Gaussian). In Section 4, the equivalent backward channel model will be used to develop the canonical GDFE receiver structure.

3.1 Forward and backward channel models

Given two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, either may be expressed uniquely as the sum of its MMSE linear estimate given the other and an orthogonal error vector:

$$\boldsymbol{Y} = \hat{\boldsymbol{Y}}(\boldsymbol{X}) + \boldsymbol{F} = A^*\boldsymbol{X} + \boldsymbol{F}; \qquad (4.66)$$

$$\boldsymbol{X} = \hat{\boldsymbol{X}}(\boldsymbol{Y}) + \boldsymbol{G} = B^*\boldsymbol{Y} + \boldsymbol{G} , \qquad (4.67)$$

where A and B are matrices to be determined, and $\boldsymbol{F}$ and $\boldsymbol{G}$ are orthogonal to $V(\boldsymbol{X})$ and $V(\boldsymbol{Y})$, respectively. The estimation error vector $\boldsymbol{F}$ is the innovations vector of $\boldsymbol{Y}$ given $\boldsymbol{X}$, while $\boldsymbol{G}$ is the innovations vector of $\boldsymbol{X}$ given $\boldsymbol{Y}$.

Suppose that the forward channel model

$$\boldsymbol{Y} = H\boldsymbol{X} + \boldsymbol{N} \qquad (4.68)$$

is given, where $\boldsymbol{N}$ is independent of $\boldsymbol{X}$, so $\boldsymbol{N}$ is orthogonal to $V(\boldsymbol{X})$. Then since the decomposition $\boldsymbol{Y} = A^*\boldsymbol{X} + \boldsymbol{F}$ is unique, $H\boldsymbol{X}$ must be the MMSE linear estimate $\hat{\boldsymbol{Y}}(\boldsymbol{X})$ of $\boldsymbol{Y}$ given $\boldsymbol{X}$, and $\boldsymbol{N}$ must be the estimation error or innovations of $\boldsymbol{Y}$ given $\boldsymbol{X}$.

The alternative representation above is then called the backward channel model, which may be written with the notation

$$\boldsymbol{X} = \hat{\boldsymbol{X}}(\boldsymbol{Y}) + \boldsymbol{E} = C\boldsymbol{Y} + \boldsymbol{E} , \qquad (4.69)$$

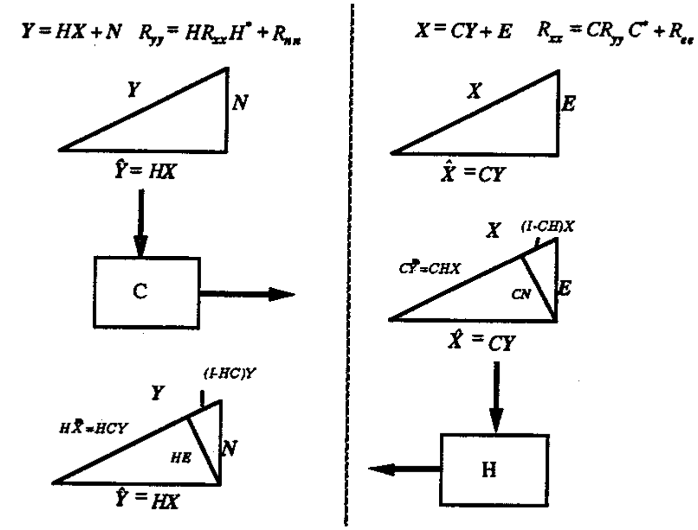

Figure 4.3 Pythagorean relations for forward and backward channel models.

where $C\boldsymbol{Y}$ denotes the MMSE linear estimate $\hat{\boldsymbol{X}}(\boldsymbol{Y})$ of $\boldsymbol{X}$ given $\boldsymbol{Y}$, and $\boldsymbol{E} = \boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{Y})$ is the estimation error or innovations vector of $\boldsymbol{X}$ given $\boldsymbol{Y}$. Thus, $B = C$ and $\boldsymbol{G} = \boldsymbol{E}$.

Figure 4.3 shows how the two Pythagorean representations of the forward and backward channel models may be combined, in two different ways. Thus in the forward channel

$$\hat{\boldsymbol{Y}}(\boldsymbol{X}) = H\boldsymbol{X} = HC\boldsymbol{Y} + H\boldsymbol{E} \qquad (4.70)$$

is the sum of the orthogonal vectors $HC\boldsymbol{Y} \in V(\boldsymbol{Y})$ and $H\boldsymbol{E} \in V(\boldsymbol{Y})^{\perp}$, and

$$\boldsymbol{N} = (I - HC)\boldsymbol{Y} - H\boldsymbol{E} . \qquad (4.71)$$

is also the sum of two orthogonal vectors. Similarly, in the backward channel there are orthogonal decompositions

$$\hat{\boldsymbol{X}}(\boldsymbol{Y}) = C\boldsymbol{Y} = CH\boldsymbol{X} + C\boldsymbol{N} \qquad (4.72)$$

$$\boldsymbol{E} = (I - CH)\boldsymbol{X} - C\boldsymbol{N} , \qquad (4.73)$$

where $CH\boldsymbol{X}$, $(I - CH)\boldsymbol{X} \in V(\boldsymbol{X})$ and $C\boldsymbol{N} \in V(\boldsymbol{X})^{\perp}$. All right triangles are geometrically similar; the "angle" between the two spaces $V(\boldsymbol{X})$ and $V(\boldsymbol{Y})$ is determined by the cross-correlation matrix $R_{\boldsymbol{xy}} = <\boldsymbol{X}, \boldsymbol{Y}>$.

3.2 Canonical forward and backward channel models

The principles of Section 2 and of optimum estimation theory are now used to reduce the general channel models of the previous section to canonical forms in which extraneous dimensions are eliminated, and which have other nice properties.

Definition 3 (Canonical Channel Model) *A channel model* $Y = HX + N$ *is canonical if* H *is square,* R_{xx} *and* R_{nn} *are nonsingular, and furthermore* $R_{nn} = H$*, which implies that* H *is a positive definite Hermitian-symmetric matrix.*

Our first observation is that any part of the input X that lies in the right null space (kernel) of H may be disregarded. In general, the matrix H defines a linear transformation $H\colon \mathcal{C}^m \to \mathcal{C}^n$ from the input space $\mathcal{C}^m$ of all possible complex m-vectors to the output space $\mathcal{C}^n$. The right null space of H is the kernel $K \subseteq \mathcal{C}^m$ of this transformation.

Any $x \in \mathcal{C}^m$ may be written uniquely as

$$x = x_{|K} + x_{|K^\perp} , \tag{4.74}$$

where $x_{|K}$ is the projection of x onto K and $x_{|K^\perp} = x - x_{|K}$ is the projection of x onto the orthogonal space $K^\perp$ to K. The signal component of the channel output then depends only on $x_{|K^\perp}$, since

$$Hx = Hx_{|K^\perp} , \tag{4.75}$$

independent of $x_{|K}$, since $Hx_{|K} = 0$. Thus, the input is effectively $x_{|K^\perp}$, and $x_{|K}$ does not affect the channel output. The projection $x_{|K}$ will be called the **undetectable part** of the input x, and $x_{|K^\perp}$ will be called the **effective input**.

If the input is a random vector X with covariance matrix R_{xx} and signal space $\mathcal{S}_X \subseteq \mathcal{C}^m$, then X may similarly be decomposed uniquely into

$$X = X_{|K} + X' , \tag{4.76}$$

where $X_{|K}$ is an undetectable input random vector defined on the space $K \cap \mathcal{S}_X$, while $X' = X_{|K^\perp}$ is an effective input random vector defined on the effective input space $\mathcal{S}_{X'} = K^\perp \cap \mathcal{S}_X$. The probability density of the effective input X' and its covariance matrix $R_{x'x'}$ are induced from those of X by this definition.

The output signal then depends only on X':

$$HX = HX' . \tag{4.77}$$

The linear transformation $H\colon \mathcal{S}_{X'} \to \mathcal{S}_{\hat{Y}}$ is one-to-one over these spaces (but is not necessarily one-to-one on the larger spaces $\mathcal{C}^m \to \mathcal{C}^n$), and the signal space $\mathcal{S}_{\hat{Y}}$ is the image of $\mathcal{S}_{X'}$ under the transformation H. It follows that $\mathcal{S}_{X'}$ and $\mathcal{S}_{\hat{Y}}$ both have the same dimension, which will be called the effective rank of the channel and denoted as $r_{\hat{y}}$. This rank often is less than the input or output dimensionality of the original channel matrix H, so that $r_{\hat{y}} = r_{x'} \le \min(n, r_x)$ and $r_{x'} \le r_x \le m$. Strict inequalities in fact often apply for optimized covariance R_{xx} as shown in later sections. Thus, both X' and $\hat{Y}(X) = HX = HX'$ have a dimensionality associated with the with $V(X')$, that is rank $r_{\hat{y}} = r_{x'}$. Clearly only the $r_{\hat{y}}$-dimensional effective input $X' = X_{|K^\perp}$ can convey information through the channel, and any power applied to the $(n - r_{\hat{y}})$-dimensional undetectable part $X_{|K}$ is wasted.

Since $R_{x'x'}$ has rank $r_{\hat{y}}$, the effective input X' may be represented as

$$X' = UM' , \tag{4.78}$$

where U is an $n \times n$ "unitary" matrix and is therefore an volume-preserving transformation, regardless of whether X' has full rank ($r_{\hat{y}} = r_{x'} = m$) or not ($r_{\hat{y}} < r_x \le m$), and M' is a set of random variables with covariance matrix $R_{m'm'}$. It may be desirable for the elements of M' to be uncorrelated, in which case (4.78) becomes the modal representation of Section 2.2. The rank and effective determinant of $R_{m'm'}$ are then the same as those of $R_{x'x'}$:

$$r_{m'} = r_{x'} = r_{\hat{y}} ; \tag{4.79}$$

$$\|R_{m'm'}\| = \|R_{x'x'}\| . \tag{4.80}$$

The identically zero components of M' and the associated columns of U may be eliminated to obtain an equivalent one-to-one volume-preserving transformation from $M \in \mathcal{C}^{r_{\hat{y}}}$ to $X' \in \mathcal{S}_{X'}$:

$$X' = U'M . \tag{4.81}$$

Then M has full rank $r_{\hat{y}}$, and the determinant of R_{mm} is equal to the effective determinant of $R_{m'm'}$:

$$m = r_{m'} = r_{x'} = r_{\hat{y}}; \tag{4.82}$$

$$|R_{mm}| = \|R_{m'm'}\| = \|R_{x'x'}\| . \tag{4.83}$$

Although the matrix U' is not square in general, the map U' remains a one-to-one volume-preserving transformation from $\mathcal{C}^{r_{\hat{y}}}$ to $\mathcal{S}_{X'}$. It is clear that estimation of M is equivalent to estimation of X'. Because M is full rank, then any characteristic representation of Section 2.2 in the form

$$M = FV \tag{4.84}$$

will have a volume preserving F of rank $r_{\hat{y}} = m = r_v$. A convenient form of the GDFE for H corresponding to stationary scalar channels will use the innovations decomposition in Section 4 while vector coding in Section 5 will use the modal decomposition.

The forward channel model may now be written as

$$Y = HU'M + N = GM + N \,, \tag{4.85}$$

where M is a complex random $r_{\hat{y}}$-vector with a covariance matrix R_{mm} that is positive definite and thus invertible, and N is a Gaussian noise vector independent of M with nonsingular covariance matrix R_{nn}.

Finally, a series of information-lossless linear transformations may be applied to the channel output Y to obtain the final form of a canonical model. First, let S be any square root of R_{nn}; i.e., let S be an invertible square matrix such that $R_{nn} = SS^*$. Then the invertible noise-whitening matrix S^{-1} applied to Y yields the equivalent model

$$Y' = S^{-1}Y = S^{-1}GM + S^{-1}N = G'M + N' \,, \tag{4.86}$$

where $N' = S^{-1}N$ is a Gaussian noise vector with an identity covariance matrix,

$$R_{n'n'} = S^{-1}R_{nn}S^{-*} = I \,, \tag{4.87}$$

and $G' = S^{-1}G = S^{-1}HU'$. The principle of the sufficiency of matched filtering may be applied: in the presence of white Gaussian noise, the outputs of a bank of matched filters matched to the responses of each input M_j (the columns of G') form a set of sufficient statistics for the detection of M. In matrix notation, this set of outputs is the $r_m = r_{\hat{y}}$-dimensional vector

$$Z = (G')^*Y' = (G')^*G'M + (G') * N' = R_f M + N'' \,, \tag{4.88}$$

where the $r_m \times r_m$ full-rank positive-definite matrix R_f is

$$R_f = (G')^*G' = (U')^*H^*S^{-*}S^{-1}HU' = (U')^*H^*R_{nn}^{-1}HU' \tag{4.89}$$

a Hermitian-symmetric matrix, M is a set of $r_{\hat{y}} = r_m$ uncorrelated random variables with nonsingular covariance matrix R_{mm}, and N'' is an independent Gaussian noise vector with covariance matrix

$$R_{n''n''} = (G')^*R_{n'n'}G' = (G')^*G' = R_f \,. \tag{4.90}$$

Thus the noise covariance matrix is equal to the equivalent channel-response matrix R_f. This yields our desired canonical forward channel model. This construction is summarized in Figure 4.4.

In summary:

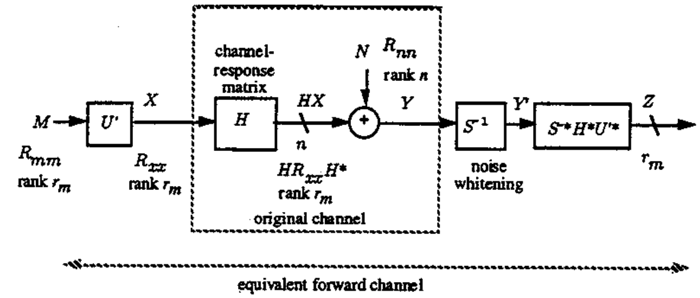

Figure 4.4 Construction of canonical forward channel model.

Theorem 4.1 (Equivalency of the Canonical Forward Channel Model) *Let $Y = HX + N$, where HX has rank $r_m = r_{\hat{y}}$ and N is a full-rank ($r_n = n$) Gaussian random vector independent of X. Without loss of optimality, an equivalent forward channel model is the canonical model*

$$Z = R_f M + N'' \,, \tag{4.91}$$

where the channel-response matrix R_f is a square $r_m \times r_m$ full-rank covariance matrix, the input M is a full-rank $r_{\hat{y}}$-vector with nonsingular covariance matrix R_{mm}, and the noise N'' is a full-rank Gaussian $r_{\hat{y}}$-vector independent of M whose covariance matrix $R_{n''n''}$ is equal to R_f. There is a one-to-one volume-preserving map from M to the effective part X' of the input X, and

$$r_{x'} = r_m = r_{\hat{y}} \le r_y; \|R_{x'x'}\| = |R_{mm}| \,. \tag{4.92}$$

The output Z is a sufficient statistic for detection of M or of X', and consequently the mutual information between input and output is the same in both models:

$$I(M;Z) = I(X';Y) = I(X;Y) \text{ bits/block.} \tag{4.93}$$

Since all $r_{\hat{y}} \times r_{\hat{y}}$ matrices are nonsingular, it is possible to solve explicitly for the corresponding backward channel model. The MMSE linear estimate of M given Z is $\hat{M}(Z) = R_b Z$, where the matrix R_b is determined by

$$R_b^* = R_{zz}^{-1}R_{zm} \,. \tag{4.94}$$

Since

$$R_{zm} = E[ZM^*] = R_f R_{mm}; \tag{4.95}$$

$$R_{zz} = E[ZZ^*] = R_f R_{mm} R_f + R_f = R_f R_{mm}(R_{mm}^{-1} + R_f) \,, \tag{4.96}$$

R_b is determined by the following fundamental formula:

$$R_b^{-1} = R_{\mathbf{mm}}^{-1} + R_f \, . \tag{4.97}$$

This formula shows that R_b is Hermitian-symmetric, $R_b^* = R_b$. Also, it shows that as the input covariance $R_{\mathbf{mm}}$ becomes large, R_b tends to the inverse of R_f, meaning the noise/errors $\boldsymbol{E}$ and $\boldsymbol{N}''$ can be ignored.

The covariance matrix $R_{\mathbf{ZZ}}$ is most easily determined from the following relationships between the four matrices R_f, R_b, $R_{\mathbf{mm}}$ and $R_{\mathbf{ZZ}}$:

$$R_{\mathbf{ZZ}} = R_f R_{\mathbf{mm}} R_b^{-1} = R_b^{-1} R_{\mathbf{mm}} R_f \, . \tag{4.98}$$

Since the covariance matrix $R_{\mathbf{ee}}$ of the estimation error vector $\boldsymbol{E} = \boldsymbol{M} - R_b \boldsymbol{Z}$ satisfies

$$R_{\mathbf{mm}} = R_b R_{\mathbf{ZZ}} R_b + R_{\mathbf{ee}} = R_{\mathbf{mm}} R_f R_b + R_{\mathbf{ee}} \, , \tag{4.99}$$

it follows that in the equivalent backward channel model the noise covariance matrix is again equal to the backward channel-response matrix:

$$R_{\mathbf{ee}} = R_{\mathbf{mm}} - R_{\mathbf{mm}} R_f R_b = R_{\mathbf{mm}}(R_b^{-1} - R_f) R_b = R_b \, . \tag{4.100}$$

In summary:

Theorem 4.2 (Equivalency of the Backward Canonical Model) *Under the same conditions as in Theorem 4.1, there is an equivalent canonical backward channel model*

$$\boldsymbol{M} = R_b \boldsymbol{Z} + \boldsymbol{E} \, , \tag{4.101}$$

where R_b is a square nonsingular Hermitian-symmetric channel-response matrix

$$R_b = (R_{\mathbf{mm}}^{-1} + R_f)^{-1} \, , \tag{4.102}$$

$\boldsymbol{Z}$ is a random "backward input" $r_{\tilde{y}}$-vector with nonsingular covariance matrix

$$R_{\mathbf{ZZ}} = R_f R_{\mathbf{mm}} R_b^{-1} = R_b^{-1} R_{\mathbf{mm}} R_f \, , \tag{4.103}$$

and $\boldsymbol{E}$ is a random error $r_m = r_{\tilde{y}}$-vector uncorrelated with $\boldsymbol{M}$ whose covariance matrix $R_{\mathbf{ee}}$ is equal to the channel-response matrix R_b.

Example 4.4 (Parallel Channels Example) Let the forward channel correspond to the previous "parallel channels" model of Section 1.1 and so any of the subchannels (with normalization of gain to one) is an ideal one-dimensional complex Gaussian channel $Y = X + N$ with signal and noise variances S_x and S_n, respectively. The corresponding equivalent canonical forward channel model for any such subchannel is

$$Z = S_n^{-1} Y = S_n^{-1} X + S_n^{-1} N = R_f M + N'' \, , \tag{4.104}$$

where $R_f = S_n^{-1}$, $M = X$ with $R_{\mathbf{mm}} = S_x$, and $N'' = S_n^{-1} N$ with $S_{n''n''} = S_n^{-1} = R_f$. The corresponding equivalent canonical backward channel model is

$$X = M = R_b Z + E \, , \tag{4.105}$$

where

$$R_b = (R_{\mathbf{mm}}^{-1} + R_f)^{-1} = (S_x^{-1} + S_n^{-1})^{-1} = S_x S_n / (S_x + S_n) \, ; \tag{4.106}$$

$$R_{\mathbf{ZZ}} = R_f R_{\mathbf{mm}} R_b^{-1} = (S_x + S_n)/S_n^2 \, ; \tag{4.107}$$

$$R_{\mathbf{ee}} = R_b = S_x S_n / (S_x + S_n) \, . \tag{4.108}$$

If any of the one-dimensional channels (or subchannel) had $S_x = 0$, then the reduction procedure from $\boldsymbol{X}$ to $\boldsymbol{M}$ would have resulted in this channel being eliminated from the set in the canonical forward and backward realizations, which would then consist of r_m subchannels corresponding to those with nonzero input energy. A subchannel with $h_i = 0$ would also suggest that the corresponding $S_{x,i}$ be set to zero and eliminated - that is that dimension is in the kernel $K \cap S_{\boldsymbol{X}}$ of H and so is eliminated. Thus the models in (4.105) and (4.106) correspond to only the used subchannels from the parallel set.

Thus the equivalent canonical backward channel model is similar to the forward model in that it is square, nonsingular, and has noise covariance equal to the channel matrix. It differs in that the elements of $\boldsymbol{Z}$ are not in general uncorrelated, and the "noise" $\boldsymbol{E}$ is not in general Gaussian; furthermore, $\boldsymbol{E}$ is merely uncorrelated with $\boldsymbol{Z}$ rather than independent of $\boldsymbol{Z}$. This is not surprising, since nothing in the derivation of the backward model depends on $R_{\mathbf{mm}}$ being diagonal, $\boldsymbol{N}''$ being Gaussian, or $\boldsymbol{N}''$ being independent of (rather than merely uncorrelated with) $\boldsymbol{M}$.

By rederiving the forward model from the backward model, or by direct substitution, one may obtain the symmetrical relations

$$R_f^{-1} = R_{\mathbf{ZZ}}^{-1} + R_b \, ; \tag{4.109}$$

$$R_{\mathbf{mm}} = R_b R_{\mathbf{ZZ}} R_f^{-1} = R_f^{-1} R_{\mathbf{ZZ}} R_b \, . \tag{4.110}$$

Many other matrix relations follow easily. In particular:

$$R_f R_b = I - R_f R_{\mathbf{ZZ}}^{-1} = I - R_{\mathbf{mm}}^{-1} R_b \, ; \tag{4.111}$$

$$R_b R_f = I - R_{\mathbf{ZZ}}^{-1} R_f = I - R_b R_{\mathbf{mm}}^{-1} \, ; \tag{4.112}$$

$$R_b^{-1} R_f^{-1} = I + R_b^{-1} R_{\mathbf{ZZ}}^{-1} = I + R_{\mathbf{mm}}^{-1} R_f^{-1} \, ; \tag{4.113}$$

$$R_f^{-1} R_b^{-1} = I + R_f^{-1} R_{\mathbf{mm}}^{-1} = I + R_{\mathbf{ZZ}}^{-1} R_b^{-1} \, . \tag{4.114}$$

Example 4.5 (Parallel Channels continued) For the ideal one-dimensional channel (or any of the used subchannels in the parallel set), these relations become

$$R_f R_b = S_x/(S_x + S_n) \; ; \tag{4.115}$$

$$R_f^{-1} R_b^{-1} = 1 + \frac{S_n}{S_x} \, . \tag{4.116}$$

From these equations, one may obtain the determinantal relations

$$\begin{aligned} |I - R_f R_b| &= |I - R_b R_f| = |R_f|/|R_{ZZ}| \\ &= |R_b|/|R_{mm}| \; ; & (4.117) \\ |R_f R_b|^{-1} &= |R_b R_f|^{-1} = |I + R_b^{-1} R_{ZZ}^{-1}| & (4.118) \\ &= |I + R_{mm}^{-1} R_f^{-1}| \\ &= |I + R_f^{-1} R_{mm}^{-1}| = |I + R_{ZZ}^{-1} R_b^{-1}| \, . & (4.119) \end{aligned}$$

Using these relations, one may verify that all the right triangles shown in Figure 4.4 are in fact similar, if the squared length of each side is identified with the determinant of its covariance matrix; the ratio of the squared lengths of the long side to the hypotenuse is always $|R_f R_b|$, and the ratio of the squared lengths of the short side to the hypotenuse is always $|I - R_f R_b|$. (Note again that here the "Pythagorean theorem" involves the sums of covariance matrices, not of their determinants.)

The forward and backward equivalent canonical channel models are two completely equivalent ways of specifying the joint probability distribution $p_{M/Z}(m, z)$. The forward model corresponds to specifying first M, then Z given M; i.e., to specifying $p_{M/Z}(m, z)$ as the product $p_M(m) p_{Z/M}(z|m)$. The backward model corresponds to specifying first Z, then M given Z; i.e., to specifying $p_{M/Z}(m, z)$ as the product $p_Z(z) p_{M/Z}(m|z)$.

In the forward channel, the conditional probability $p_{Z/M}(z|m)$ is specified by an independent Gaussian noise variable N'' via $p_{Z/M}(z|m) = p_{N''}(z - R_f m)$. If M and therefore Z and E are Gaussian, then a similar separation formula holds in the backward channel.

3.3 ML and MMSE estimation

ML Estimation

Given an output y, a maximum-likelihood (ML) estimator chooses the input $x \in C^m$ that maximizes the likelihood $p_{Y/X}(y|x) = p_N(y - Hx)$. Since $Hx = Hx'$, where x' is the effective part of the input, all that an ML estimator can actually do is estimate the effective input $x' \in S_{X'} \subset C^m$, or the corresponding $r_m = r_{\hat{y}}$-vector m such that $x = U'm$.

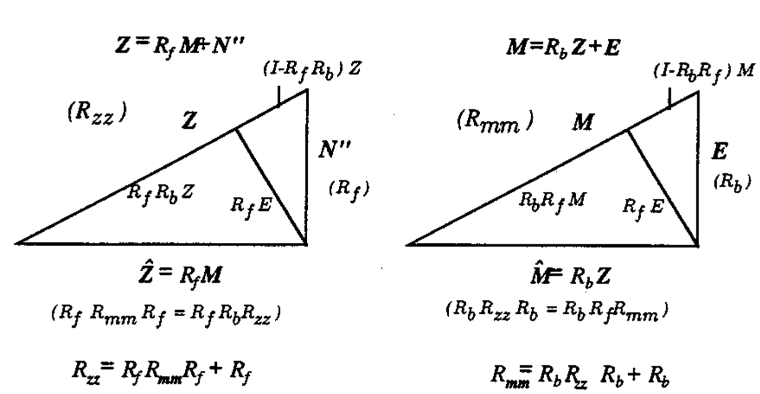

Figure 4.5 Similar right triangles.

Theorem 4.3 (The ML Estimator and Zero-Forcing Equalizer) *The ML estimates of* $\boldsymbol{X}'$ *or* $\boldsymbol{M}$ *from* $\boldsymbol{Y}$ *or* $\boldsymbol{Z}$ *are*

$$\hat{\boldsymbol{M}}(\boldsymbol{Z}) = R_f^{-1}\boldsymbol{Z} \; ; \tag{4.120}$$

$$\hat{\boldsymbol{X}}'(\boldsymbol{Z}) = U'\hat{\boldsymbol{M}}(\boldsymbol{Z}) = U'R_f^{-1}\boldsymbol{Z} \; . \tag{4.121}$$

Proof: Since $\boldsymbol{Z} = R_f\boldsymbol{M}+\boldsymbol{N}''$ is a sufficient statistic for estimation of $\boldsymbol{X}'$ or $\boldsymbol{M}$ from $\boldsymbol{Y}$, there is a one-to-one map between $S_{x'}$ or $S_m = C^{r_m}$ and $S_z = C^{r_m}$. Since $p_{\mathbf{N}''}(\boldsymbol{z} - R_f\boldsymbol{m})$ is maximized for the $\boldsymbol{m}$ such that $\boldsymbol{z} = R_f\boldsymbol{m}$ when $\boldsymbol{N}''$ is Gaussian, then the theorem follows. **QED.**
In other words, the block ML estimator simply computes the unique (effective) input that would give the observed matched-filter output vector $\boldsymbol{Z}$ in the absence of noise. For this reason the ML estimator is sometimes called a **zero-forcing equalizer.**

The ML estimation error is

$$\hat{\boldsymbol{E}} = \boldsymbol{M} - \hat{\boldsymbol{M}}(\boldsymbol{Z}) = \boldsymbol{M} - R_f^{-1}(R_f\boldsymbol{M} + \boldsymbol{N}'') = R_f^{-1}\boldsymbol{N}'' \; , \tag{4.122}$$

a Gaussian random vector with covariance matrix $R_f^{-1}R_{\boldsymbol{n}''\boldsymbol{n}''}R_f^{-1} = R_f^{-1}$.

ML Detection

It is important to distinguish an ML estimator from an ML detector, the latter of which is optimum for discrete uniform input distributions on $\boldsymbol{M}$. The ML estimator is only "optimum" when the input distribution for $\boldsymbol{M}$ is continuous uniform, which never occurs in practice. However, for some choices of receiver, an ML estimator followed by an ML detector designed only on knowledge of coding applied to $\boldsymbol{M}$ (and therefore not using any knowledge of the channel) to be optimum. The vector coding methods of Section 5 illustrate this property.

MMSE Estimation and MMSE Equalization

As observed in Section 2.3, for any optimality criterion based on error variances, the vector MMSE linear estimate is optimum among all linear estimators. Therefore, without more precisely specifying the optimality criterion, the linear MMSE estimator of $\boldsymbol{X}$ given $\boldsymbol{Y}$ (or the MMSE equalizer) may be defined as the vector MMSE linear estimate $\hat{\boldsymbol{X}}(\boldsymbol{Y}) = C\boldsymbol{Y}$. The linear estimation error vector $\boldsymbol{E} = \boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{Y})$ is then minimized in every sense among linear estimators.

Theorem 4.4 (MMSE Estimator and MMSE Equalizer) *The MMSE estimator is given by*

$$\hat{\boldsymbol{M}}(\boldsymbol{Z}) = R_b\boldsymbol{Z} \; ; \tag{4.123}$$

$$\hat{\boldsymbol{X}}'(\boldsymbol{Z}) = U'\hat{\boldsymbol{M}}(\boldsymbol{Z}) \tag{4.124}$$

$$= U'R_b\boldsymbol{Z} \; . \tag{4.125}$$

In other words, the block MMSE estimator simply computes the unique (effective) input that minimizes the error vector covariance and does not ignore noise. For this reason the MMSE estimator is sometimes called a **linear MMSE equalizer.**

Proof: Follows directly from the equivalent backward channel model. **QED.**

The linear MMSE estimation error $\boldsymbol{E} = \boldsymbol{M} - \hat{\boldsymbol{M}}(\boldsymbol{Z})$ has covariance matrix $R_{\boldsymbol{ee}} = R_b$, which is "less than" R_f^{-1} since

$$R_f^{-1} - R_b = R_{\boldsymbol{zz}}^{-1} \tag{4.126}$$

is a positive definite matrix (sometimes written $R_{\boldsymbol{zz}}^{-1} > 0$, or $R_f^{-1} > R_b$). However, as the signal-to-noise ratio becomes large, R_b approaches R_f^{-1}.
The estimation error for $\boldsymbol{X}'$ is simply

$$\boldsymbol{E}_{\boldsymbol{x}'} = \boldsymbol{X}' - \hat{\boldsymbol{X}}'(\boldsymbol{Z}) = U'\boldsymbol{M} - U'\hat{\boldsymbol{M}}(\boldsymbol{Z}) = U'\boldsymbol{E} \; . \tag{4.127}$$

Because U' is a volume-preserving transformation, the effective determinant of the covariance matrix of $\boldsymbol{E}_{\boldsymbol{x}'}$ is equal to $|R_{\boldsymbol{ee}}| = |R_b|$.

Example 4.6 (Parallel Channels continued) For the ideal one-dimensional channel (or one of several subchannels in a parallel set) $Y = X + N$ with $R_{\boldsymbol{xx}} = S_x$ and $R_{\boldsymbol{nn}} = S_n$, or the equivalent channel $Z = S_n^{-1}M + N''$ with $R_{\boldsymbol{mm}} = S_x$ and $R_{\boldsymbol{n}''\boldsymbol{n}''} = S_n^{-1}$, the ML estimate of $M = X$ is $S_nZ = Y$, which has error variance S_n. The MMSE estimate of $M = X$ is $(S_xS_n/(S_x+S_n))Z = (S_x/(S_x+S_n))Y$, which has error variance $R_{\boldsymbol{ee}} = S_xS_n/(S_x + S_n) < S_n$. For deleted subchannels, no estimate occurs.

MAP Detection and Estimation

If the input $\boldsymbol{X}$ is Gaussian, then $\boldsymbol{E}$ is Gaussian and the linear MMSE estimator is the unconstrained MMSE estimator. Furthermore, since it maximizes the a posteriori probability density $p_{\mathrm{M/Z}}(\boldsymbol{m}|\boldsymbol{z}) = p_{\mathrm{E}}(\boldsymbol{v} - R_b\boldsymbol{m})$, it may alternatively be called the maximum a posteriori (MAP) estimator. It is important to distinguish the MAP detector from the MAP estimator, just as it is important to distinguish the ML detector from the ML estimator. The detector is optimum when the input distribution is, as always the case in practice, discrete. The estimator is only defined for continuous distributions, and particularly for the this case, a continuous Gaussian distribution. Nonetheless, a MMSE estimator on a channel with a discrete input distribution for $\boldsymbol{M}$, followed by a detector that has a structure based only on $\boldsymbol{M}$ and not on the channel can be canonical, with specific structures illustrating this property in Sections 4 and 5.

Estimator and Detector Bias

The bias of an estimator $c\boldsymbol{X}$ of a random vector $\boldsymbol{X}$ is the difference between $\boldsymbol{X}$ and the expected value $E[c\boldsymbol{X}|\boldsymbol{X}]$. The ML estimator is unbiased, since

$$\bar{M}(Z) = M + R_f^{-1}N'' , \quad (4.128)$$

so $E[\bar{M}|M] = M$. Indeed, it is clear that the ML estimator is the unique unbiased linear estimator of $\boldsymbol{M}$ given $\boldsymbol{Z}$, since if $\bar{M} = CZ$, then $E[\bar{M}|M] = CR_fM$, which is equal to $\boldsymbol{M}$ everywhere if and only if $C = R_f^{-1}$.
The linear MMSE estimator is biased:

$$E[\hat{M}|M] = R_bR_fM = (I - R_bR_{mm}^{-1})M = M - R_bR_{mm}^{-1}M . \quad (4.129)$$

The bias is $R_bR_{mm}^{-1}M$, which tends to zero as the signal-to-noise ratio becomes large.

3.4 Mutual information

If the input $\boldsymbol{X}$ is Gaussian, then the mutual information $I(X;Y) = I(M;Z)$ between input and output may be expressed in either of two ways:

$$I(M;Z) = h(Z) - h(Z|M) = h(Z) - h(N'') = \log|R_{zz}|/|R_f| \; ; (4.130)$$

$$I(M;Z) = h(M) - h(M|Z) = h(M) - h(E) = \log|R_{mm}|/|R_b| (4.131)$$

These relations recall the determinantal relations derived earlier,

$$|R_f|/|R_{zz}| = |R_b|/|R_{mm}| = |I - R_fR_b| = |I - R_bR_f| , \quad (4.132)$$

from which it follows that

$$I(M;Z) = -\log|I - R_fR_b| = -\log|I - R_bR_f| . \quad (4.133)$$

The determinant $|R_{mm}|$ is equal to the effective determinant $\|R_{x'x'}\|$ of the effective input $\boldsymbol{X}'$ in its r_y-dimensional space $\mathcal{S}_{\boldsymbol{X}'}$. The determinant $|R_b|$ is equal to the effective determinant $\|R_{e'e'}\|$ of the error of the MMSE estimator $U'M$ of $\boldsymbol{X}'$. Therefore there is an interesting connection between MMSE estimation and mutual information, as follows:

Theorem 4.5 (Sufficiency of Canonical Transmission with the Forward and Backward Channel Models) *Given a channel model $Y = HX + N$ where N is full-rank and Gaussian, let $\|R_{x'x'}\|$ be the effective determinant of the effective input X', and let $\|R_{e'e'}\|$ be the effective determinant of the linear MMSE estimate of X' given Y. Then the mutual information $I(X;Y)$ when X is Gaussian is given by*

$$I(X;Y) = \log\|R_{x'x'}\|/\|R_{e'e'}\| . \quad (4.134)$$

The above theorem implies the potential existence of canonical transmission systems that use only the forward or backward canonical models, thus ignoring all eliminated dimensions and inputs. As noted earlier, the mutual information $I(X;Y)$ when $\boldsymbol{X}$ is Gaussian is an upper bound to the mutual information when $\boldsymbol{X}$ is an arbitrary random vector with the same second-order statistics.

Example 4.7 (Parallel Channels (cont.)) On the ideal one-dimensional Gaussian channel, the input variance is $R_{xx} = S_x$ and the MMSE error variance is $R_{ee} = S_xS_n/(S_x + S_n)$, so $I(X;Y) = \log R_{xx}/R_{ee} = log(S_x + S_n)/S_n$.

Matrix SNRs

Mutual information results suggest some matrix SNR definitions that allow generalization of many of the results in [1].

Definition 4 (MMSE-SNR Matrix) *Define the square matrix SNR_{GDFE} by*

$$SNR_{\text{GDFE}} = R_{mm}R_b^{-1} \; ; \quad (4.135)$$

then

$$I(M;Z) = I(X;Y) = \log|SNR_{\text{GDFE}}| . \quad (4.136)$$

(Alternatively, the same result is obtained using the conjugate transpose $\mathbf{SNR}^* = \mathbf{SNR} = R_b^{-1}R_{mm}$.*)*

SNR_{GDFE} is the matrix generalization of $\text{SNR}_{\text{MMSE-DFE}}$ for the infinite-length MMSE-DFE. The SNR is well-understood through the ratio of the transmitted message energy (covariance) R_{mm} to the minimized square-error power (covariance) $R_{ee} = R_b$. Similarly, recalling that the ML error variance is R_f^{-1}, define

$$SNR_{\text{ML}} = R_{mm}R_f , \quad (4.137)$$

(or alternatively $SNR^*_{\text{ML}} = SNR_{\text{ML}} = R_fR_{mm}$). Then, since $R_b^{-1} = R_{mm}^{-1} + R_f$, it follows that

$$SNR_{\text{GDFE}} = I + SNR_{\text{ML}} . \quad (4.138)$$

Equation (4.138) is the matrix equivalent of the expression $\text{SNR}_{\text{MMSE-DFE}} = \text{SNR}_{\text{MMSE-DFE,U}} + 1$ in [1].
In summary:

Theorem 4.6 *Given canonical forward and backward channel models* $\boldsymbol{Z} = R_f\boldsymbol{M} + \boldsymbol{N}''$ *and* $\boldsymbol{M} = R_b\boldsymbol{Z} + \boldsymbol{E}$, *define* $\boldsymbol{SNR}_{\text{GDFE}} = R_{\boldsymbol{mm}}R_b^{-1} = R_f^{-1}R_{\boldsymbol{zz}}$, *and define* $\boldsymbol{SNR}_{\text{ML}} = R_{\boldsymbol{mm}}R_f = R_bR_{\boldsymbol{zz}}$. *Then the mutual information* $I(\boldsymbol{M};\boldsymbol{Z})$ *when* $\boldsymbol{M}$ *is Gaussian is given by*

$$I(\boldsymbol{M};\boldsymbol{Z}) = \log|\boldsymbol{SNR}_{\text{GDFE}}| = \log|I + \boldsymbol{SNR}_{\text{ML}}| . \quad (4.139)$$

Bias Results

Since the ML estimator is the unique unbiased linear estimator, this result may also be interpreted as a relation between mutual information, linear MMSE estimation, and unbiased linear estimation.

Since the MMSE estimator is $R_b\boldsymbol{Z}$ and the unbiased ML estimator is $R_f^{-1}\boldsymbol{Z}$, an MMSE estimate may be converted to an unbiased ML estimate by multiplication by $\boldsymbol{SNR}_{\text{ML}}^{-1}\boldsymbol{SNR}_{\text{GDFE}}$, or by $\boldsymbol{SNR}_{\text{GDFE}}\boldsymbol{SNR}_{\text{ML}}^{-1}$. The bias of the MMSE estimate is equal to

$$R_bR_{\boldsymbol{mm}}^{-1}\boldsymbol{M} = \boldsymbol{SNR}_{\text{GDFE}}^{-1}\boldsymbol{M} . \quad (4.140)$$

Example 4.8 (Example 2 (cont.)) On the ideal one dimensional Gaussian channel, $\boldsymbol{SNR}_{\text{GDFE}} = R_{\boldsymbol{xx}}/R_{\boldsymbol{ee}} = (S_x + S_n)/S_n$ and $\boldsymbol{SNR}_{\text{ML}} = R_{\boldsymbol{xx}}/R_{\boldsymbol{nn}} = S_x/S_n = \boldsymbol{SNR}_{\text{GDFE}} - 1$. The biased MMSE estimate $(S_x/(S_x + S_n))Y$ may be converted to the unique unbiased linear estimate Y by multiplication by $\text{SNR}_{\text{MMSE-DFE}}/\text{SNR}_{\text{ML}} = (S_x + S_n)/S_x$.

Notice that $\boldsymbol{SNR}_{\text{GDFE}}$ and $\boldsymbol{SNR}_{\text{ML}}$ are diagonalized by the same unitary transformations U and therefore commute; for if

$$\boldsymbol{SNR}_{\text{ML}} = U\Lambda_{ML}^2U*, \quad (4.141)$$

then

$$\boldsymbol{SNR}_{\text{GDFE}} = I + \boldsymbol{SNR}_{\text{ML}} = U(I + \Lambda_{ML}^2)U* = U\Lambda_{MMSE}^2U* . \quad (4.142)$$

This implies that the eigenvalues of $\boldsymbol{SNR}_{\text{GDFE}}$ are equal componentwise to 1 plus the eigenvalues of $\boldsymbol{SNR}_{\text{ML}}$:

$$\lambda_{MMSE,j}^2 = 1 + \lambda_{ML,j}^2 , \quad (4.143)$$

regardless of whether the matrices R_f, $R_{\boldsymbol{mm}}$, R_b and $R_{\boldsymbol{zz}}$ commute. Thus each of the individual modes in the block channel has a relationship between MMSE SNR and unbiased SNR that parallels the relationship established in [1], namely $\text{SNR}_{\text{MMSE-DFE}} = \text{SNR}_{\text{MMSE-DFE,U}} + 1$.

4 THE GENERALIZED DFE RECEIVER STRUCTURE

This section introduces and develops the canonical GDFE receiver structure for a general block Gaussian ISI channel $\boldsymbol{Y} = H\boldsymbol{X} + \boldsymbol{N}$ with arbitrary H, $R_{\boldsymbol{xx}}$ and (full-rank) $R_{\boldsymbol{nn}}$. This structure is an apparently novel structure that is a finite-length generalization of the usual infinite-length MMSE decision-feedback equalization (MMSE-DFE).

The starting point is an equivalent canonical backward channel model

$$\boldsymbol{M} = R_b\boldsymbol{Z} + \boldsymbol{E} . \quad (4.144)$$

Since R_b is a nonsingular covariance matrix, it has a unique Cholesky factorization

$$R_b = L_bD_b^2L_b^* , \quad (4.145)$$

where L_b is a monic lower triangular matrix and D_b^2 is a nonsingular positive-definite diagonal matrix.

Premultiplication by the lower triangular matrix L_b^{-1} yields the equivalent channel model

$$\boldsymbol{M}' = L_b^{-1}\boldsymbol{M} = D_b^2L_b^*\boldsymbol{Z} + L_b^{-1}|\boldsymbol{E} = \boldsymbol{Z}' + \boldsymbol{E}' , \quad (4.146)$$

where $\boldsymbol{Z}' = D_b^2L_b^*\boldsymbol{Z}$ may be viewed as the result of passing $\boldsymbol{Z}$ through an upper-triangular "feedforward filter" $D_b^2L_b^*$, and the noise $\boldsymbol{E}' = L_b^{-1}\boldsymbol{E}$ has a diagonal covariance matrix

$$R_{\boldsymbol{e'e'}} = L_b^{-1}R_{\boldsymbol{ee}}L_b^{-*} = D_b^2 ; \quad (4.147)$$

i.e., its components are uncorrelated.

The usual assumptions of decision-feedback equalization are now invoked:

- symbol-by-symbol decisions may be made on the components M_j of $\boldsymbol{M}$;
- in the detection of M_j, it may be assumed that all previous decisions are correct (the "ideal DFE assumption").

Now since L_b^{-1} is lower triangular, M_j is a linear combination of Z_j', E_j' and previous components $\boldsymbol{M}(j-1) = [M_1, ..., M_{j-1}]$. The MMSE symbol estimate of M_j given Z_j' and $\boldsymbol{M}(j-1)$ is therefore equal to Z_j' minus the linear combination of the past components $[M_1, ..., M_{j-1}]$ that is specified by the j^{th} row of L_b^{-1} (the "feedback filter" at time j). The error in this estimate is E_j'. The signal-to-noise ratio for the j^{th} symbol is thus

$$\text{SNR}_j = E[|M_j|^2]/E[|E_j'|^2] = \lambda_{m,j}^2/d_{b,j}^2 , \quad (4.148)$$

where $\lambda_{m,j}^2$ is the j^{th} diagonal element of the diagonal matrix $R_{\boldsymbol{mm}}$, and $d_{b,j}^2$ is the j^{th} Cholesky factor of R_b.

Theorem 4.7 (GDFE is Canonical) *If R_{mm} is diagonal, or equivalently, the input vector $\boldsymbol{M}$ has uncorrelated elements, then the GDFE is canonical.*

Proof: The product of the symbol SNRs is $|\boldsymbol{SNR}_{\text{GDFE}}|$, since

$$\prod_j \text{SNR}_j = |R_{mm}|/|D_b^2| = |R_{mm}|/|R_b| = |\boldsymbol{SNR}_{\text{GDFE}}| \ . \quad (4.149)$$

This expression is the key to showing that this receiver structure is canonical. To complete the proof, assume that $\boldsymbol{X}$ and thus all random vectors are Gaussian. Then

$$I(\boldsymbol{M};\boldsymbol{Z}) = \log|\boldsymbol{SNR}_{\text{GDFE}}| \ . \quad (4.150)$$

Furthermore, from the chain rule of information theory,

$$I(\boldsymbol{M};\boldsymbol{Z}) = \prod_j I(M_j; Z_j|\boldsymbol{M}(j-1)) \ . \quad (4.151)$$

The mutual information in the jth symbol transmission may be expressed as

$$\begin{aligned} I(M_j; Z_j|\boldsymbol{M}(j-1)) &= h(M_j|\boldsymbol{M}(j-1)) - h(M_j|Z_j', \boldsymbol{M}(j-1)) \\ &= h(M_j) - h(E_j') && (4.152) \\ &= \log \frac{\lambda_{m,j}^2}{d_{b,j}^2} && (4.153) \\ &= \log \text{SNR}_j \ , && (4.154) \end{aligned}$$

since M_j is independent of $\boldsymbol{M}(j-1)$ when $\boldsymbol{M}$ is Gaussian and E_j' is the estimation error (innovations) for M_j given $[Z_j', \boldsymbol{M}(j-1)]$. Thus

$$I(\boldsymbol{M};\boldsymbol{Z}) = \prod_j \log \text{SNR}_j \ . \quad (4.155)$$

Now use a long code of rate arbitrarily close to $\log \text{SNR}_j$ on each subchannel that has an arbitrarily low error probability. Decode the subchannels in order so that the "past" decisions $M_1, ..., M_{j-1}$ are available when decoding M_j (which justifies the ideal DFE assumption). Then one can send at an aggregate rate arbitrarily close to $I(\boldsymbol{M};\boldsymbol{Z}) = \log|\boldsymbol{SNR}_{\text{GDFE}}|$ per block with arbitrarily low probability of error. Hence this block MMSE-DFE receiver structure is canonical.**QED**.

In practice, as in vector coding systems, one can code "across subchannels" to avoid excessive decoding delay and buffering. The ideal DFE assumption then fails, but this problem may be elegantly handled by a kind of "transmitter precoding" similar to the precoding techniques that have been developed for single-channel transmission systems.

4.1 The Packet GDFE - stationary special case

The GDFE is general but does not converge to the usual MMSE-DFE for infinite-length packets on a stationary channel without the additional transmitter alterations in this subsection. In general, these alterations add additional complexity for no improvement in performance, other than they allow a recursive implementation of the transmit filter via Cholesky factorization.

The input vector $\boldsymbol{M}$ can be decomposed according to its innovation representation as in Section 2.2 as

$$\boldsymbol{M} = L\boldsymbol{W} \ , \quad (4.156)$$

where $R_{mm} = LR_{ww}L^*$ and L is lower-triangular and nonsingular, and where R_{ww} is diagonal. The elements of $\boldsymbol{W}$ are the innovations of $\boldsymbol{M}$. For the Packet GDFE, the elements of $\boldsymbol{W}$ are considered the coded input sequence and should be the values estimated by the GDFE receiver. The alteration necessary to the receiver is simply to replace the feedback section by the rows of $\tilde{L} = L^{-1}L_b^{-1}$ instead of the rows of L_b^{-1}. This new matrix feedback section is still lower triangular and previous decisions on elements of $\boldsymbol{W}$ can be used to aid future decisions just as the elements of $\boldsymbol{M}$ were used in the diagonal-R_{mm} case.

In the transmitter, the input becomes

$$\boldsymbol{X} = U'\boldsymbol{M} = U'L\boldsymbol{V} \quad (4.157)$$

and $\|R_{xx}\| = |R_{mm}| = |R_{vv}|$ so that $I(\boldsymbol{X};\boldsymbol{Y}) = I(\boldsymbol{M};\boldsymbol{Z}) = I(\tilde{\boldsymbol{X}};\boldsymbol{Y})$.

Lemma 4.1 (Packet GDFE is Canonical) *The packet GDFE, which estimates $\boldsymbol{W}$ directly in the feedback section and uses the additional transmit filter of L is canonical.*

Proof: The proof is identical to the proof for the GDFE with diagonal $\boldsymbol{M}$ with $\boldsymbol{W}$ replacing $\boldsymbol{M}$. **QED**.

The transmit signal decomposition has an interesting interpretation:

- The lower-triangular filter L relates the innovations or underlying transmitted signal to the filtered channel output for whatever transmit covariance R_{xx} is used. When R_{xx} is stationary, the rows of L will converge for long packet length to the filters that relate the innovations to the channel input. Different parts of L may converge to different filters, corresponding to the different frequency bands used.
- The filter U' is not triangular and is necessary when dimensions have been reduced from the original channel H. This filter combines different sets of L into a single transmit signal - the transmit signal thus contains potentially nonoverlapping frequency bands in the the stationary case and

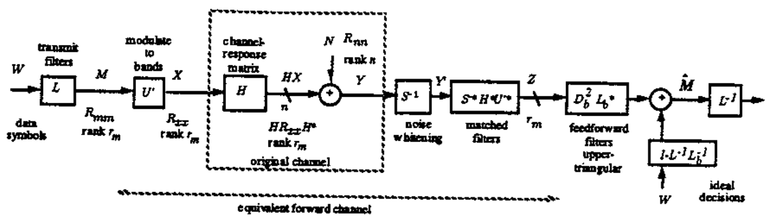

Figure 4.6 Packet GDFE.

U' is an orthogonal matrix that is volume preserving and corresponds to essentially modulation of the various baseband signals generated by the triangular L into the different frequency bands.

Figure 4.6 illustrates the various parts of the GDFE, including the special packet case. When U' corresponds to a modal decomposition, $L = I$ and $M = V$, and the input cannot be realized by triangular filtering.

5 TRANSMITTER OPTIMIZATION AND VECTOR CODING

To this point the input covariance matrix R_{xx} has been assumed to be given. In this section R_{xx} will be chosen subject to a power constraint to maximize the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ assuming Gaussian input statistics, or equivalently to maximize the determinant $|\boldsymbol{SNR}_{\text{GDFE}}|$.

An optimized R_{xx} has a natural diagonal representation that suggests the well-known method of vector coding, which yields an alternative canonical receiver structure for the optimum R_{xx}. When R_{xx} is optimized, or equivalently R_{mm} is optimized, then VC and the GDFE have the same canonical performance.

5.1 Input covariance optimization

The channel model is the general linear Gaussian channel $\boldsymbol{Y} = H\boldsymbol{X} + \boldsymbol{N}$ as before, with H and R_{nn} given, and R_{xx} to be optimized subject to a constraint on the average input energy $E[\boldsymbol{X}\boldsymbol{X}^*]$, which is the trace of R_{xx}- i.e., the sum $\sum_j R_{xx}(jj)$ of the variances $R_{xx}(jj)$ of the components X_j of $\boldsymbol{X}$.

As discussed earlier, a general input vector $\boldsymbol{X}$ may be decomposed into an undetectable part $\boldsymbol{X}_{|K}$ in the right null space K of H and an effective input $\boldsymbol{X}_{|K^\perp}$ in the orthogonal subspace $K^\perp$. It is clear that no energy should be wasted on $X_{|K}$ and therefore an optimized $\boldsymbol{X}'$ should be constrained to lie in $K^\perp$. The dimension $r_{\tilde{y}} = r_m$ of $K^\perp$ is equal to the dimension of the range space of H, since H is a one-to-one map from $K^\perp$ to its range space.

It is convenient to choose an orthonormal basis for $K^\perp$ consisting of $r_{x'}$ orthonormal vectors $\boldsymbol{U} = \{\boldsymbol{u}_j, 1 \le j \le r_{x'}\}$. Then any $\boldsymbol{x} \in K^\perp$ may be written as a linear combination of the basis vectors, $\boldsymbol{x} = U\boldsymbol{m}$, and furthermore because of orthonormality, $\|\boldsymbol{x}\|^2 = \|\boldsymbol{m}\|^2$. Then an average energy constraint on $\boldsymbol{X}$ translates directly into an equivalent average energy constraint on the random r_x-vector $\boldsymbol{M}$. The VC design then chooses R_{mm} in the channel model $\boldsymbol{Y} = HU\boldsymbol{M} + \boldsymbol{N}$ to maximize $I(\boldsymbol{X};\boldsymbol{Y}) = I(\boldsymbol{M};\boldsymbol{Y})$ subject to an average energy constraint on $\boldsymbol{M}$.

Again noise whitening and matched filtering may be used without loss of optimality to reduce to obtain an equivalent canonical channel model

$$\boldsymbol{Z} = R_f\boldsymbol{M} + \boldsymbol{N}'' . \tag{4.158}$$

where $R_f = U^*H^*R_{nn}^{-1}HU$ is a square full-rank covariance matrix that is both the channel-response matrix and the covariance matrix of $\boldsymbol{N}''$. Again $I(\boldsymbol{M};\boldsymbol{Z}) = I(\boldsymbol{X};\boldsymbol{Y})$. This equivalent canonical forward channel model is similar to that derived in Section 3, except that R_{mm} is yet to be determined and is not necessarily diagonal.

Now it is desired to maximize $I(\boldsymbol{M};\boldsymbol{Z}) = \log|I + \boldsymbol{SNR}_{\text{ML}}| = \log|I + R_{mm}R_f|$:

Theorem 4.8 (Optimum transmit vectors) *The optimum R_{mm} must have the same eigenvectors as R_f: i.e., R_{mm} must commute with R_f.*

Proof: Let V be a unitary transformation that diagonalizes R_f; i.e., $V^*R_fV = \Lambda_f^2$, where $\Lambda_f^2 = \text{diag}\{\lambda_{f,j}^2\}$ is a diagonal matrix whose diagonal components are the eigenvalues $\lambda_{f,j}^2$ of R_f. Let $R_{m,j}^2$ be the diagonal elements of $V^*R_{mm}V$. Then the diagonal elements of $I + V^*R_{mm}VV^*R_fV = I + (V^*R_{mm}V)\Lambda_f^2$ are $\{1 + R_{m,j}^2\lambda_{f,j}^2\}$, so by Hadamard's inequality,

$$|I + R_{mm}R_f| = |I + V^*R_{mm}VV^*R_fV| \le \prod_j (1 + R_{m,j}^2\lambda_{f,j}^2) , \tag{4.159}$$

with equality if and only if $V^*R_{mm}V$ is diagonal. Since V is a unitary transformation the trace (sum of the diagonal components) of $V^*R_{mm}V$ is the same as the trace of R_{mm}; therefore setting the off-diagonal components of $V^*R_{mm}V$ to zero will not change the average energy of $\boldsymbol{M}$, but will necessarily decrease $I(\boldsymbol{M};\boldsymbol{Y}) = I(\boldsymbol{M};\boldsymbol{Z})$ unless $V^*R_{mm}V$ is already diagonal. Thus the optimum $V^*R_{mm}V$ is diagonal. **QED.**

Given that $V^*R_{mm}V$ is a diagonal matrix $\Lambda_m^2 = diag\{\lambda_{m,j}^2\}$, the optimum variances $\lambda_{m,j}^2$ may then be determined by discrete water-pouring in the usual manner, with the result that

$$\lambda_{m,j}^2 = K - \frac{1}{\lambda_{f,j}^2}, \text{if } K \geq \frac{1}{\lambda_{f,j}^2}; \tag{4.160}$$

$$\lambda_{m,j}^2 = 0 \text{ , otherwise,} \tag{4.161}$$

where K is a constant chosen so that the average energy constraint on $\prod_j \lambda_{m,j}^2$ is met. Thus this water-pouring optimization may cause some of the subchannels to be unused and thus reduce the effective rank of the channel below $r_{x'}$ to a new value of r_m that would then force $r_{x'}$ to be smaller and equal to r_m through the original definitions of these ranks, which depend on the choice of R_{xx}, or equivalently, R_{mm}.

The optimum R_{xx} is then determined from the optimum Λ_m^2 via

$$R_{xx} = UR_{mm}U^* = UV\Lambda_M^2V^*U^* . \tag{4.162}$$

A canonical GDFE receiver may then be constructed from this optimum R_{xx} and may be used to approach the maximized $I(\boldsymbol{M};\boldsymbol{Y})$, namely the channel capacity of the given linear Gaussian packet channel.

Finally, since

$$R_{xx} = UR_{mm}U^* ; \tag{4.163}$$

$$H^*R_{nn}^{-1}H = UR_fU^* , \tag{4.164}$$

it follows readily from the orthonormality of U that if R_{mm} and R_f commute, then R_{xx} and $H^*R_{nn}^{-1}H$ commute.

5.2 Commuting channels

The above argument shows that an optimum R_{mm} commutes with R_f. A canonical channel model $\boldsymbol{Z} = R_f\boldsymbol{M} + \boldsymbol{N}$ will be called commuting if R_{mm} and R_f commute:

$$R_{mm}R_f = R_fR_{mm} . \tag{4.165}$$

Equivalently, since $(R_{mm}R_f)^* = R_fR_{mm}$, a canonical channel is commuting if $R_{mm}R_f$ is Hermitian-symmetric. Since

$$\boldsymbol{SNR}_{\text{ML}} = R_{mm}R_f ; \tag{4.166}$$

$$\boldsymbol{SNR}_{\text{GDFE}} = I + \boldsymbol{SNR}_{\text{ML}}, \tag{4.167}$$

a channel is commuting if either SNR_{ML} or $\text{SNR}_{\text{MMSE-DFE}}$ is Hermitian-symmetric. A one-dimensional channel is necessarily commuting.

The inverses R_{mm}^{-1} and R_f^{-1} of commuting covariance matrices commute with with R_{mm} and R_f and with each other. Moreover, from the defining equations for R_b and R_{zz},

$$R_b = (R_{mm}^{-1} + R_f)^{-1} ; \tag{4.168}$$

$$R_{zz} = R_fR_{mm}R_b^{-1} = R_b{-1}R_{mm}R_f , \tag{4.169}$$

it follows that R_b and R_{zz} and their inverses also commute with each other and with R_{mm} and R_f. Thus the corresponding backward canonical channel model is commuting as well.

5.3 Vector coding

If $\boldsymbol{Z} = R_f\boldsymbol{M} + \boldsymbol{N}$ is a commuting equivalent forward channel model of rank r_y, then $R_{mm} = VL\Lambda_m^2V^*$ and $R_f = V\Lambda_f^2V^*$ for some unitary matrix V, so

$$\boldsymbol{Z} = V\Lambda_f^2V^*\boldsymbol{M} + \boldsymbol{N} ; \tag{4.170}$$

$$\boldsymbol{Z}' = V^*\boldsymbol{Z} = \Lambda_f^2V^*\boldsymbol{M} + V^*\boldsymbol{N} = \Lambda_f^2\boldsymbol{M}' + \boldsymbol{N}' , \tag{4.171}$$

where $\boldsymbol{Z}' = V^*\boldsymbol{Z}$, $\boldsymbol{M}' = V^*\boldsymbol{M}$, and $\boldsymbol{N}' = V^*\boldsymbol{N}$; i.e., the random vectors are represented in the basis determined by the unitary transformation V. Now

$$R_{m'm'} = V^*R_{mm}V = \Lambda_m^2 ; \tag{4.172}$$

$$R_{n'n'} = V^*R_fV = \Lambda_f^2 . \tag{4.173}$$

The channel therefore naturally decomposes into r_y decoupled one-dimensional subchannels of the form

$$Z_j' = \lambda_{f,j}^2M_j' + N_j', 1 \leq j \leq r_y , \tag{4.174}$$

where the variance of M_j' is $\lambda_{m,j}^2$ and N_j' is an independent Gaussian variable of variance $\lambda_{f,j}^2$. This is just a standard one-dimensional Gaussian channel model of the type of Example 2, with $S_{x,j} = \lambda_{f,j}^4\lambda_{m,j}^2, S_{n,j} = \lambda_{f,j}^2$, and therefore

$$\frac{S_{x,j}}{S_{n,j}} = \lambda_{f,j}^2\lambda_{m,j}^2 . \tag{4.175}$$

The mutual information over such a channel is

$$I(M_j';Z_j') = \log\left(1 + \frac{S_{x,j}}{S_{n,j}}\right) = \log(1 + \lambda_{f,j}^2\lambda_{m,j}^2) . \tag{4.176}$$

The aggregate mutual information of all parallel subchannels is

$$I(M';Z') = \prod_j \log(1+\lambda_{f,j}^2\lambda_{m,j}^2) = \log|I+\lambda_f^2\Lambda_m^2|$$
$$= \log|I + R_f R_{mm}| = \log|SNR_{\mathrm{GDFE}}| \,. \qquad (4.177)$$

It follows that this structure, called vector coding, is canonical for any commuting channel. In particular, it is canonical for any channel for which R_{mm} or R_{xx} has been optimized.

Lemma 4.2 (Optimality and Canonical properties of VC) *Vector Coding is both optimal and canonical for a commuting channel.*

Proof: It follows directly from (4.177) that VC is canonical. VC is also a ML estimator for which each subchannel can use an ML detector for the applied code. If the input X is uniform discrete over the $r_{\tilde{y}}$-dimensional subspace, then this ML detector minimizes error probability. **QED.**

If the channel is not commuting, however, then it cannot be decomposed into completely decoupled one-dimensional subchannels in this way; i.e., vector coding is not well defined for noncommuting channels. Thus in certain cases where R_{xx} is predetermined and cannot be optimized, the GDFE structure may be the only canonical receiver structure available.

DMT - Discrete Multitone

DMT or Discrete Multitone is a special case of VC when the channel correlation matrix matrix $HR_{nn}^{-1}H^*$ is circulant. This circulant property is forced by the use of a cyclic prefix in each transmitted packet, which is simply a repeat of the last few samples at the beginning and end of a packet. The eigenvectors needed for a commutative channel and for the optimized input are essentially the vectors associated with a Discrete Fourier Transform, thus allowing very efficient optimal and canonical implementations through the use of Fast Fourier Transform methods.

6 LIMITING RESULTS WITH INCREASING PACKET LENGTH

The results in this paper all converge to generalizations of the known results in [1] for infinite-length (continuous non-packet) transmission on a stationary dispersive channel with additive Gaussian noise. This convergence requires that the individual elements of the vectors X and N are successive samples from stationary random processes and that H for any values of $m \geq n$ has each successive row moved one position to the right with respect to the previous

row, but the row elements are otherwise the same. That is, H is "Toeplitz" as $n \to \infty$.

Perhaps not well established in [1] is the situation in which these well-known results exist, namely that the input process X must have nonsingular covariance as $n \to \infty$, which requires a resampling or "optimization of symbol and center frequencies" as a function of the channel, which tacitly may involve multiple disjoint frequency bands and multiple MMSE-DFE's. The GDFE more accurately describes these multiple MMSE-DFE's in the limit, each of which exhibits the properties discussed in [1].

6.1 Channel Models

The D-transform of a discrete time sequence or random process X_k (the samples of X as $m \to \infty$) is $X(D) \triangleq \sum_k X_k D^k$. Convolution of sequences in discrete time corresponds to multiplication of their D-transforms. The matrix channel with Toeplitz H corresponds to convolution of $X(D)$ with $h(D)$ (the D-transform of the first row of H). Thus, the dual channel model becomes:

$$Y = HX + N \Longrightarrow Y(D) = h(D)X(D) + N(D) \qquad (4.178)$$
$$X = CY + E \Longrightarrow X(D) = c(D)Y(D) + E(D) \,. \qquad (4.179)$$

Multiplication of a vector by H^* corresponds to convolution with $h^*(D^{-*})$. Thus, a matched filter output is

$$Z(D) = h^*(D^{-*})Y(D) \,. \qquad (4.180)$$

For stationary sequences, the autocorrelation function $r_{xx,k} = E[X_l X_{l-k}^*]$ has a D-transform

$$R_{xx}(D) = \sum_k r_{xx,k} D^k \,. \qquad (4.181)$$

Pythagorean relationships are

$$R_{yy}(D) = h(D)R_{xx}(D)h^*(D^{-*}) + R_{nn}(D) \qquad (4.182)$$
$$R_{xx}(D) = c(D)R_{yy}(D)c^*(D^{-*}) + R_{ee}(D) \,. \qquad (4.183)$$

Also,

$$R_{ee}(D) = R_{ex}(D) = R_{xx}(D) - c(D)R_{xy}(D) \,. \qquad (4.184)$$

6.2 Limiting Entropy, Mutual Information, and SNR

The innovations are stationary when $X(D)$ is stationary and critical to the generalization of entropy. A particularly crucial problem in establishing limiting results, and required by a stationary process, is the singularity of the process $X(D)$. Subsection 6.2 reviews results when $X(D)$ is nonsingular and Subsection 6.2 extends and generalizes these results in a heuristic way to the nonsingular case.

Nonsingular input sequences

A stationary random sequence $x(D)$ satisfies the Paley-Wiener Criterion:

$$S_w = |\frac{1}{2\pi}\int_{-\pi}^{\pi} \log R_{xx}(e^{-\jmath\theta})d\theta| < \infty \ , \tag{4.185}$$

which means it is also nonsingular. In practice, satisfaction of the PW criterion means the power spectral density $R_{xx}(e^{-\jmath\theta})$ cannot be infinite nor zero at more than a few discrete frequencies, a requirement often not met if the input sequence $X(D)$ tries to zero energy in certain regions of the band that water-filling arguments might dictate should be zeroed. For a nonsingular sequence the vector $\boldsymbol{X}$ will have a nonsingular $R_{\boldsymbol{xx}}$ for all packet lengths as $m \to \infty$. The types of singular processes of interest in Subsection 6.2 are actually very close to stationary in that within certain frequency bands (or at the right sampling rates and center/carrier frequencies), PW is individually satisfied for each of several disjoint bands.

When $X(D)$ is stationary and therefore nonsingular, the relation

$$\boldsymbol{W} = L^{-1}\boldsymbol{X} \tag{4.186}$$

directly corresponds to the chain rule for entropy ([2]) when $\boldsymbol{X}$ is Gaussian:

$$H(\boldsymbol{X}) = H(X_1) + H(X_2/X_1) + ... + H(X_m/\{X_1, X_2, ...X_{m-1}\}) \ . \tag{4.187}$$

That is W_k is the MMSE sample corresponding to the estimate of X_k, given all previous values of $X(D)$,

$$H(\boldsymbol{X}) = H(W_1) + H(W_2) + ...H(W_m) \ . \tag{4.188}$$

L^{-1} is a linear prediction filter operating on $\boldsymbol{X}$ to produce $\boldsymbol{W}$. Clearly, since X_k is stationary, this filter is constant, meaning that L^{-1} tends towards a Toeplitz matrix when m gets large, and $R_{\boldsymbol{ww}}$ tends towards a constant diagonal matrix with linear minimum mean square error S_w along the diagonal.

In this case, Cholesky factorization corresponds to

$$R_{\boldsymbol{xx}} = LR_{\boldsymbol{ww}}L^* \Longrightarrow R_{xx}(D) = l(D)S_w l^*(D^{-*}) \ , \tag{4.189}$$

where $l(D)$ is monic ($l_0 = 1$), causal ($l_k = 0 \ \forall \ k < 0$) and minimum-phase (all roots and poles outside the unit circle), and

$$S_w = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log R_{xx}(e^{-\jmath\theta})d\theta \ , \tag{4.190}$$

implying that $R_{xx}(D)$ satisfies the discrete-time Paley-Wiener criterion [22]. The linear prediction filter is $1/l(D)$, and the innovations sequence is $w(D) = x(D)/l(D)$.

The entropy for a stationary process is defined as

$$H(X) = \lim_{m\to\infty} \frac{mS_w}{m} = \lim_{m\to\infty} \frac{H(\boldsymbol{X})}{m} \ . \tag{4.191}$$

For the Gaussian random process $X(D)$, this value is clearly

$$H(X) = log(\pi e S_w) \ , \tag{4.192}$$

and because $R_{\boldsymbol{ww}}$ is a constant diagonal, the prediction error sequence or innovations $W(D)$ is white. Similarly, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly stationary and Gaussian, the limit is found using Toeplitz distribution results [23],

$$H(X/Y) = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \log_2 R_{ee}(e^{-\jmath\theta})d\theta \ . \tag{4.193}$$

Essentially, the conditional entropy is equal to the entropy of the error sequence associated with estimation of the random variable based on the given random variable. This stationary Gaussian error sequence itself also has a innovations representation and the conditional entropy is thus also equal to the entropy of this innovations sequence. Thus,

$$\begin{aligned} H(X/Y) &= H(X_k/(Y, [E_{k-1}, E_{k-2}...])) = H(X_k/(Y, [X_{k-1}, X_{k-2}...])) \\ &= \log_2(\pi e \frac{N_0}{D_b}) \ , \end{aligned} \tag{4.194}$$

where the rightmost relation is obtained by recognizing that the MMSE estimation associated with $H(X_k/(Y, [X_{k-1}, X_{k-2}...]))$ is that of the MMSE-DFE. Further, D_b must also converge to a constant since the matrix R_b^{-1} is Toeplitz when $R_{\boldsymbol{ww}}$ is constant, which it must be when the system is stationary and infinite length. The error sequence for the MMSE-DFE is white because this sequence is the innovations sequence for the linear prediction of the error sequence corresponding to the linear MMSE of $X(D)$ given $Y(D)$.

The value of D_b is determined from the spectral factorization $R_b^{-1}(D) = l_b(D)D_b l_b^*(D^{-*})$ (where $l(D)$ is causal, monic, and minimum-phase, $D_b > 0$ is real, and $l^*(D)$ is anticausal, monic, and maximum-phase - see [1])

$$D_b = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \log R_b^{-1}(e^{-\jmath\theta})d\theta \ , \tag{4.195}$$

where,

$$R_b^{-1}(D) = R_{mm}^{-1}(D) + R_f(D) = R_m^{-1}(D) + h(D)R_{nn}^{-1}(D)h^*(D^{-*}); . \tag{4.196}$$

The factorization of (4.196) is called the "key equation" in [1].

The single-band MMSE-DFE: The mutual information for jointly stationary and Gaussian $X(D)$ and $Y(D)$ also has a limiting definition

$$I(X;Y) = \lim_{m\to\infty} \bar{I}(\boldsymbol{X};\boldsymbol{Y}) . \tag{4.197}$$

The formula $I(X;Y) = H(X) - H(X/Y) = H(Y) - H(Y/X)$ leads to

$$I(X;Y) = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \log_2(\text{SNR}(\theta))d\theta , \tag{4.198}$$

where

$$\text{SNR}(\theta) = \frac{S_w}{R_{ee}(e^{-\jmath\theta})} = \frac{S_w|l(e^{-\jmath\theta})h(e^{-\jmath\theta})|^2 + R_{nn}(e^{-\jmath\theta})}{R_{nn}(e^{-\jmath\theta})} . \tag{4.199}$$

The MMSE-DFE is biased, but simple scaling can remove the bias and the relation

$$I(X;Y) = \log_2(1 + \text{SNR}_{\text{MMSE-DFE,U}}) = \log_2(S_w D_b) , \tag{4.200}$$

shows the MMSE-DFE to be canonical for a given fixed choice of input spectrum given by $R_{xx}(D) = l(D)S_w l^*(D^{-*})$. Thus, a maximum likelihood detector is not necessary because best performance can be attained by applying a good code with as small a gap from mutual information (on an AWGN) to an intersymbol interference channel that uses a MMSE-DFE and still maintain that same small gap from mutual information.

The Case of Singular Input

Technically, a singular input sequence is not stationary because it does not satisfy the Paley-Wiener Criterion. However, it is often possible in practice to resample a sequence at a lower rate, and possibly with a carrier offset in bandpass processes, so that an equivalent complex baseband random process is stationary. Such stationary processes can be added together, again with carrier offset, to create a nonstationary process (in this case, cyclostationary with period equal to the greatest common multiple of carrier periods or in the finite-length case to the packet period).

In effect, each of the frequency bands used now has a stationary process within it and all the results of Section 6.2 apply individually to each band. The data rate is of course the sum of the data rates. The SNR is the geometric average where each band's SNR is weighted by its ratio of bandwidth to the total used by all bands. The union of all these disjoint bands is denoted by Ω and a modification of the PW criterion holds such that

$$\frac{1}{2\pi}\int_{\Omega} \log R_{xx}(e^{-\jmath\theta})d\theta| < \infty . \tag{4.201}$$

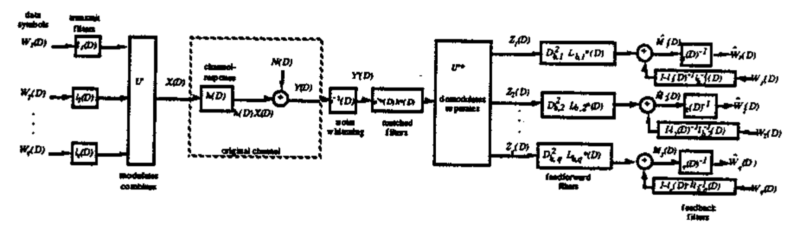

Figure 4.7 Singular GDFE in the limit.

With the GDFE, this situation is illustrated much more clearly than in [1]. The transmit filter U' of the canonical channel models combines the various bands via interpolation and translation. Translation in frequency is a unitary matrix operation. Recall U' was $m \times r_m$ "unitary" matrix, thus allowing for interpolation of the input to effectively a higher sampling rate for the combined signals The matrix L does not converge to a single filter, but rather essentially becomes triangular with disjoint blocks, each of which internally exhibits the convergence of the rows to the innovations filter for the corresponding band. U' then combines these signals into an aggregate (cyclostationary) packet transmit signal.

This situation is depicted in Figure 4.7.

6.3 Vector Coding to Multitone

The VC case, as in Section 5, corresponds to the forward canonical model

$$\boldsymbol{Z} = R_f \boldsymbol{M} + \boldsymbol{N}'' . \tag{4.202}$$

for which the GDFE is both canonical and ML if the input $\boldsymbol{M}$ is already, or is decomposed by, a modal decomposition

$$\boldsymbol{M} = V\boldsymbol{M}' \tag{4.203}$$

where both $\boldsymbol{M}$ and $\boldsymbol{M}'$ have full rank $r_m \le m$. Singularity is trivially handled by U' in the VC case as it corresponds to ignoring subchannels for which $\lambda_{f,i} = 0$ or for which $\lambda_{m,i} = 0$.

In the limit as packet length goes to zero, Toeplitz distribution arguments ([24]) lead to the limit

$$\begin{aligned}\lim_{m\to\infty} \log_2 |\boldsymbol{SNR}_{\mathrm{ML}}|^{1/m} &= I(X;Y) \\ &= \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \log_2\left(1+\lambda_{m;i}|\lambda_{f,i}|^2\right) \\ &= \frac{1}{2\pi}\int_{\Omega} \log_2\left(1+\frac{S_x(\theta)|H(e^{-\jmath\theta})|^2}{R_{nn}(e^{-\jmath\theta})}\right) d\theta \end{aligned} \quad (4.204)$$

The vector-coding system becomes equivalent to a "multitone" transmission system as packet length goes to infinity. Thus, the GDFE and VC converge to the highest performance levels possible, namely a data rate possibility of $I(X;Y)$ if good known codes for the AWGN are applied. Both must use the same frequency bands and the channel is always commutative at infinite length.

6.4 Infinite-length Transmit Optimization

The well-known water-filling energy distribution [24] [1] satisfies

$$S_x(\theta) + \frac{R_{nn}(e^{-\jmath\theta})}{|H(e^{-\jmath\theta})|^2} = \kappa \ . \quad (4.205)$$

The solution must exhibit $S_x(\theta) \geq 0$. There is a band Ω^* such that for all $\theta \in \Omega^*$, $S_x(\theta) > 0$. When $|\Omega^*| = 2\pi$, an innovations representation of the thus stationary input can be found through the canonical factorization

$$R_{xx}(D) = l(D)S_w l^*(D^{-*}) \ . \quad (4.206)$$

Then, $l(D)$ is the stationary MMSE-DFE transmit filter that acts on the input data innovations $w(D)$ to produce the proper water-fill spectrum of the channel input sequence $x(D)$. When $S_x(\theta) = 0$ over a measurable band, then separate MMSE-DFE's should be applied to each of the measurable frequency bands for which $S_x > 0$ for all but a countable number of discrete points. The bit rate for each connected subregion of Ω^*, and the GDFE will converge to a constant on all dimensions used by water-filling that correspond to a connected subregion. Each band may have a different symbol rate (equal to the measure of the corresponding connected region of used frequencies) and possibly a carrier-frequency (corresponding to the center frequency of each such band). In effect, one independently designs a MMSE-DFE and takes limits for each of the connected sub-bands of Ω^*. The limiting case of the GDFE is the infinite-length canonical transmission structure called the MMSE-DFE in [1] in each of the optimum bands of Ω^*, which is used by either the VC GDFE for which the feedback section is trivially zero or for the packet GDFE for which feedback sections are nontrivial.

7 SUMMARY AND CONCLUSION

The concept of canonical transmission has been refined to characterize systems that may not be optimum detectors, but for which nevertheless the highest possible data rates may be transmitted with the careful application of the same good codes that near capacity on the ideal additive white Gaussian noise channel. The GDFE structure is a generalization of decision feedback that allows for any characteristic representation of an input and derives from canonical forward and backward channel models that remove unnecessary dimensions and force nonsingular transmission over only those dimensions that can carry information. Various forms of the GDFE, corresponding for instance to an innovations representation of the input, i.e., the "packet GDFE," or to a modal representation of the input, otherwise known as Vector Coding. The VC case is indeed very special, because it is both canonical and optimal and the feedback section of the GDFE trivially disappears, avoiding the need for precoding methods. The VC case, however, must use only special inputs that commute with the forward channel characterization matrix R_f while the GDFE exists in general when this condition is not met. Other characteristic representations could also be used to form other types of GDFE's.

The GDFE is always canonical. The GDFE, however, is not equivalent to the fixed DFE's in common use in data transmission, the latter of which are decidely suboptimum and not canonical unless special conditions hold that are often not met. For this reason, the GDFE is the preferred method for high-performance design of transmission on channels with ISI and additive Gaussian noise. Various methods can be used to simplify a GDFE, most notably the elimination of the feedback section with the Vector-Coding GDFE, which can be further simplified through the use of fast Fourier Transform methods in the implementation known as DMT.

Other areas of simplification of the GDFE remain open to study in addition to the study of specific performance differences on various channels, which can run from very small to very large. The existence of a packet channel model $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{N}$ has been postulated and indeed is a research topic in itself as to appropriate ways to synthesize a channel design such that this relationship holds exactly or approximately.

Acknowledgement: The authors would like to thank Dr. R. Gallager of MIT and Dr. C. Rohrs of Tellabs Research for their contributions to some of the basic concepts in this paper. The first author would like to acknowledge a series of conversations with Dr. Paul Algoet in 1990 that precipitated a few of the concepts in this paper.

Both authors would like to dedicate this paper to Dr. Thomas Kailath of Stanford University on his 60th Birthday. Through Prof. Kailath's efforts, many of the concepts used in this paper have been popularized and developed over his long and distinguished career. The first author also adds his appreciation

for the profound impact that TK has had on his career and for the patient and continued support through many stages of progress in the understanding of all aspects of information systems.

REFERENCES

[1] J.M. Cioffi, G.P. Dudevoir, M. V. Eyuboglu, and G.D. Forney. "MMSE Decision-Feedback Equalizers and Coding – Parts I and II". *IEEE transactions on Communications*, 43(10):2582–2604, October 1995.

[2] T.M. Cover and J.A. Thomas. *"Elements of Information Theory"*. Wiley & Sons, New York, 1991.

[3] M.L. Doelz, E.T. Heald, and D.L. Martin. "Binary Data Transmission Techniques for LInear Systems". *Proceedings of the IRE*, 45:656–661, May 1957.

[4] J.A.C. Bingham. "Multicarrier Modulation for Data Transmission: An Idea Whose Time has Come". *IEEE Communications Magazine*, 28(4):5–14, April 1990.

[5] A. Ruiz, J.M. Cioffi, and S. Kasturia. "Discrete Multiple Tone Modulation with Coset Coding for the Spectrally Shaped Channel". *IEEE Transactions on Communications*, 40(5), May 1992.

[6] J.C. Rault, D. Castelain, and B. Le Floch. "The Coded Orthogonal Frequency Division Multiplexing (COFDM) Technique, and Its Application to Digital Radio Broadcasting toward Mobile Receivers". In *Proceedings of Globecom 1989*, Dallas, TX, November 1989.

[7] American National Standards Institute (ANSI). "Metallic Interfaces for Asymmetric Digital Subscriber Lines (ADSL)". In *ANSI Standard T1.413*, Washington, D.C., December 1995.

[8] G. Plenge. "DAB - A new sound broadcasting system – Status of the development – Routes to its introduction". *European Broadcasting Union Review*, (246), April 1991.

[9] Per Applequist. "HD-DEVINCE, a Scandinavian terrestrial HDTV project". In *Proceedings 1993 National Association of Broadcasters*, Las Vegas, NV, April 1993.

[10] S. Kasturia, J. Aslanis, and J.M. Cioffi. "Vector Coding for Partial-Response Channels". *IEEE Transactions on Information Theory*, 36(4):741–762, July 1990.

[11] L.C. Barbosa. "Maximum Likelihood Sequence Estimators: A Geometric View". *IEEE Transactions on Information Theory*, 35(2):419–427, March 1989.

[12] N. Al-Dhahir and J.M. Cioffi. "MMSE Decision-Feedback Equalizers: Finite-Length Results". to appear, IEEE Transactions on Information Theory, 1995.

[13] N. Al-Dhahir and J.M. Cioffi. "Optimal Finite-Complexity Transmit Filters for Packet-Based Data Transmission on Dispersive Channels with Application to the FIR MMSE-DFE". to appear, IEEE transactions on Information Theory, May 1993.

[14] G.K. Kaleh. "Channel Equalization for Block Transmission Systems". *IEEE Journal on Selected Areas in Communication*, 13(1):110–121, January 1995.

[15] N. Zervos, S. Pasupathy, and A. Venetsanopoulos. "The Unified Decision Theory of Non-Linear Equalization". In *Proceedings, IEEE 1984 Globecom*, pages 683–687, Atlanta, December 1984.

[16] M. Austin. "Decision Feedback Equalization for Digital Communication over Dispersive Channels". *M.I.T. Research Lab Electronics Technical Report. 461*, August 1967.

[17] C.A. Belfiore and J.H. Park Jr. "Decision Feedback Equalization". *Proceedings IEEE*, 67(8):1143–1156, August 1979.

[18] R. Price. "Nonlinearly Feedback-Equalized PAM vs Capacity for Noisy Filter Channels". In *International Conference on Communications*, June 1972. 22-12 - 22-17.

[19] J. Salz. "Optimum Mean-Square Decision Feedback Equalization". *Bell System Technical Journal*, 52(8):1341, October 1973.

[20] E.A. Lee and D. G. Messerschmitt. *Digital Communications.* Kluwer, Boston, 1988.

[21] C.L. Lawson and R. J. Hanson. *Solving Least Squares Problems.* Prentice-Hall, Englewood Cliffs, NJ, 1974.

[22] T. Kailath. *Lectures on Least-Squares Estimation.* Springer-Verlag, New York, 1976.

[23] R.E. Blahut. *Principles and Practice of Information Theory.* Addison-Wesley, Menlo Park, CA, 1987.

[24] R.G. Gallager. *Information Theory and Reliable Communication.* Wiley, New York, NY, 1968.

[25] R. Wesel and J.M. Cioffi. "Fundamentals of Coding for Broadcast OFDM". In *Proceedings of the 29th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, October 1995.