TECHNICAL
RESEARCH
REPORT

SYSTEMS
RESEARCH
CENTER

*Supported by the
National Science Foundation
Engineering Research Center
Program (NSFD CD 8803012),
Industry and the University*

Joint Design of Block Source Codes and Modulation Signal Sets

by V. Vaishampayan and N. Farvardin

TR 90-4

Joint Design of Block Source Codes and Modulation Signal Sets *

V. Vaishampayan
Department of Electrical Engineering
Texas A&M University
College Station, Texas 77843

N. Farvardin
Electrical Engineering Department
Institute for Advanced Computer
Studies
and
Systems Research Center
University of Maryland
College Park, MD 20740

Abstract

We consider the problem of designing a bandwidth-efficient, power-limited digital communication system for transmitting information from a source with known statistics over a noisy waveform channel. Each output vector of the source is encoded by a block encoder to one of a finite number of signals in a modulation signal set. The received waveform is processed in the receiver by an estimation-based decoder. The goal is to design an encoder, decoder and modulation signal set so as to minimize the mean squared-error (MSE) between the source vector and its estimate in the receiver. For highly noisy Gaussian channels we justify restricting the estimator to the class of linear estimators. With this restriction, we derive necessary conditions for optimality of the encoder, decoder and the signal set and develop a convergent algorithm for solving these necessary conditions. We prove that the MSE of the digital system designed here is bounded from below by the MSE of an analog modulation system. Performance results for the digital system and signal constellation designs are presented for first-order Gauss-Markov sources and a white Gaussian channel. Comparisons are made against a standard vector quantizer (VQ)-based system, the bounding analog modulation system and the optimum performance theoretically attainable. The results indicate that for a correlated source, a sufficiently noisy channel and specific source block sizes and bandwidths, the digital system performance coincides with the optimum performance theoretically attainable. Further, significant performance improvements over the standard VQ-based system are demonstrated when the channel is noisy. Situations in which the linearity assumption results in poor performance are also identified.

*This work was supported in part by National Science Foundation grants NSFD MIP-86-57311 and NSFD CDR-85-00108, and in part by a grant from Martin Marietta Laboratories.

I Introduction

We consider the problem of communicating information from a continuous alphabet source over a noisy waveform channel using a digital communication system. Our objective is to study the best performance that can be achieved by a block-structured digital communication system. In this paper we consider a simple communication system which consists of an encoder that maps the output of a vector source to one of a given number of modulation signals. The decoder, based on the output of the waveform channel forms an estimate of the transmitted source vector. We formulate the problem as an optimization problem in which the objective is to minimize the overall average distortion by suitably selecting a signal set, an encoder and a decoder, while maintaining constraints on the transmitted signal energy and bandwidth.

The problem that we consider here is a joint source-channel coding problem. Recent interest in this area appears to have begun with the work of Modestino and Daut [1], in which it was demonstrated that for noisy channels the joint design of a source and channel encoder and decoder can lead to dramatic performance improvements. These improvements were obtained by lowering the rate of the source encoder (thus increasing the distortion introduced by the source encoder), and using the available bandwidth to judiciously provide channel error protection.

One of the earliest formulations of the joint source-channel coding problem as an optimization problem appears to have been by Fine [2], who developed necessary conditions for an optimum digital encoder/decoder pair, for the transmission of continuous amplitude data over a digital channel. Kurtenbach and Wintz [3] using a less general approach, considered a simple zero-memory quantizer and a binary symmetric channel (BSC) for which optimum quantizer thresholds and reconstruction levels were determined for a *fixed* codeword assignment to the quantization levels. The mean squared-error (MSE) criterion was used in both these papers.

It was later shown in [4] that the algorithm proposed in [3] need not converge, and that it is necessary to impose a realizability constraint on the quantizer thresholds in order to guarantee convergence. Further, in [4] the codeword assignment problem was also solved. A key observation here was that not all the available codewords are used for transmission in an optimal system. On the one hand this can be interpreted as a form of channel coding, but on the other it can be interpreted as a weakness of the discrete alphabet channel that results from the signal constellation being used. The work in [4] has been extended to vector quantizers (VQ's) in [5].

Ayanoglu and Gray [6] applied the generalized Lloyd algorithm to design joint source and channel optimized trellis and predictive trellis waveform coders for a variety of distortion criteria. They demonstrated significant improvements in performance over a codebook designed for a noiseless channel, and improvements in highly noisy situations over separately optimized source and channel coding schemes connected in tandem. They noted that the jointly optimized trellis coding system is simpler to implement than a trellis source coder connected in tandem with a convolutional channel coder.

All the work that we have made reference to so far, considers the joint design of source and channel encoders, but assumes a fixed modulation system. The main contribution of this paper consists of including the modulation signal set in the joint design process. The problem of signal design is an old one. A majority of the work has considered the problem of optimizing the signal constellation so as to minimize a probability of error criterion, e.g., [7], [8], [9], to name a few. It is usually assumed that the significant error event is a demodulation error to the closest signal in the signal constellation. Such an assumption becomes necessary because of difficulties in integrating Gaussian densities over irregular regions of multi-dimensional space and is good only for relatively high values of the channel signal-to-noise ratio (CSNR). It is also usually assumed that signals in the signal set are used with equal probability. This assumption is debatable since most real source coding schemes do not produce equally likely symbols. Finally, two signal constellations having the same average error probability could have a different MSE performance, since the MSE depends on the individual entries of the channel transition probability matrix.

Several authors have considered the problem of designing signal sets in order to improve the MSE performance of zero-memory quantizers over a noisy waveform channel. Bedrosian [10] assumes a linear PCM system and a uniform source distribution in which each bit of the PCM codeword is transmitted by an antipodal signal set whose energy varies with the significance of that bit. He considers single bit errors as the significant error events. For 7-bit linear PCM, he reports an extension of 1.9 dB in the usable CSNR. Sundberg [11] extended this work to include general source distributions and nonlinear PCM and more general modulation formats such as QAM. His approach is to define for each transmitted bit position, an A-factor, which is a measure of the sensitivity of that bit position to a specific channel error pattern averaged over all possible source outputs. By varying the energy allocated for the transmission of each bit, it is possible to reduce the error probability for the highly sensitive bits at the expense of the less sensitive bits in such a way that the overall distortion is minimized. He reports extensions in usable CSNR of 1.85 dB for 8-bit speech PCM. He also reports larger gains for higher dimensional signal constellations such as 16-level QAM (16-QAM) than for antipodal signaling. Note that the results presented in [11] assume a high CSNR so that only single bit errors are significant. The A-factor analysis mentioned earlier is accurate only when binary signaling is employed. Specifically, the problem arises in non-binary signaling formats because the probability of a specific bit error pattern depends on the signal transmitted and hence on the source symbol.

In order to overcome this problem, the so called D-factors were introduced by Steele, Sundberg and Wong in [12] and employed to derive weighted QAM constellations for 8-bit PCM in [13]. In [13] extensions of 3 to 5 dB in usable CSNR are reported for 8-bit logarithmic PCM transmitted using 16-level and 256-level weighted QAM, respectively.

It turns out that the problem we treat here has close links to a series of amplitude modulation and block filtering problems that have appeared in the literature. The similarity arises primarily because the MSE distortion measure suggests the use of estimation-based demodulation rather than detection-based demodulation in the receiver. Lee and Peterson [14] consider (what we refer to as) a block pulse amplitude modulation system (BPAM) in which information from a vector source is to be communicated over a noisy vector channel. The encoder and decoder are assumed to be linear transformations from the source space to the channel space and vice-versa, respectively. An energy constraint is imposed on the transmitted signal, and the encoder and decoder maps are jointly optimized for a weighted MSE criterion.

Tufts considers a classical PAM system [15] in which it is assumed that the information from a discrete-time source is to be communicated over a noisy waveform channel. Two cases are treated, the first being the design of transmitter and receiver filters subject to a constraint on the transmitted power and an additional constraint of zero intersymbol interference (ISI). The second case treats the same problem as the above, except that the restriction of zero ISI is dropped. The joint optimization problem is solved for the restrictive case of time-limited transmitter signals. Berger and Tufts [16] consider the same problem as in [15] without any constraints on the ISI, and develop a general solution for the joint design problem. They also make comparisons against the optimum performance theoretically attainable (OPTA), which is obtained by using information theoretic arguments. A key observation is that for the special case of an independent and identically distributed (iid) Gaussian source, transmitted over an ideally bandlimited, additive white Gaussian noise channel, at a signaling rate equal to the Nyquist rate of the channel, the OPTA and PAM performance curves coincide. To the best of our knowledge, a similar comparison to the OPTA has not been made for BPAM systems and correlated sources. We present such results here.

Our main objective in this paper is to include the signal constellation in the joint optimization problem. We consider linear estimator-based decoders and demonstrate that such digital systems can be treated as optimum, finite encoding rate approximations to the BPAM system, and that in the limit of infinite source encoding rates, while maintaining a fixed energy and transmission bandwidth, their MSE performance converges to that of the BPAM system from above.

The paper is organized into six sections as follows. In Section II, a general communication system model is described, notation is developed and the design problem is formulated as an optimization problem with bandwidth and energy constraints. The nature of the bandwidth constraint is made precise and a formula is developed for the optimum nonlinear estimator-based decoder. Finally justification is provided for restricting the decoder to the class of linear estimators.

In Section III, necessary conditions for optimality are developed for the system that uses a linear estimator-based decoder. An algorithm for iterative system design is presented. The structure of the optimum encoder and its effect on encoding complexity is also discussed.

In Section IV, a lower bound on the MSE performance of the class of digital systems that we propose is derived and it is proved this bound is asymptotically tight in the limit of infinite encoding rates. An iterative algorithm for system design is presented. The structure of the optimum encoder and its effect on the encoding complexity is also discussed.

In Section V, numerical results are presented for the performance of the optimum system for a range of correlated Gaussian sources and channel signal-to-noise ratios. These results are compared to the performance of a standard communication system, as well as against_the BPAM performance and the OPTA. Finally in Section VI a summary of the paper is provided, conclusions are drawn regarding the systems designed in this paper and several open issues are discussed.

II Problem Formulation

II.1 Preliminaries and Notation

It is desired to transmit information from a discrete-time, continuous-amplitude source over a noisy, bandlimited, waveform channel using a power-limited transmitter. It is assumed that the source emits a sample every τ_s seconds and is represented by a zero-mean, finite-variance, stationary, ergodic random process, $\{X_n,\ n \in \mathbb{Z}\}$. Let $\{\mathbf{X}_n\}$ be an L-dimensional vector random process constructed from $\{X_n\}$ according to $\mathbf{X}_n = (X_{nL}, X_{nL+1}, \ldots,\ X_{(n+1)L-1})^T$. Assume that $\mathbf{X}_n$ has a known L-fold probability distribution $P_{\mathbf{X}}$ and density $p_{\mathbf{X}}$. We regard the vector rather than the scalar process as our primary source and assume that a vector is produced at the fixed rate of one every $T = L\tau_s$ seconds. We will represent the source encoder, channel encoder and modulator by the single encoder map $\gamma(\cdot)$, and the source decoder, channel decoder and demodulator by the decoder map $g(\cdot)$, as illustrated in Fig. II.1.

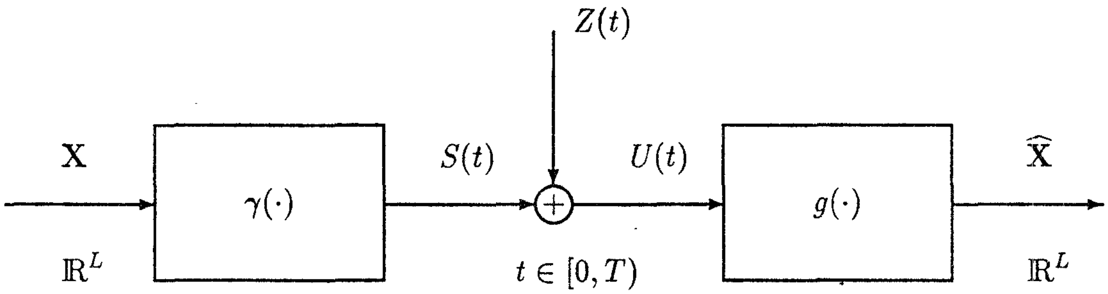

Figure II.1: General System Block Diagram

The encoder maps the L-dimensional source vector $\mathbf{X}$, to a modulation signal in the signal set $\mathcal{S} = \{s_i(t),\ i = 1, 2, \ldots,\ N,\ t \in [0,T)\}$. The output of the encoder is represented by the random process $\{S(t),\ t \in [0,T)\}$. The modulator signal $S(t)$ is transmitted over an additive white Gaussian noise (AWGN) channel having a onesided power spectral density N_0. The received waveform, $U(t) = S(t) + Z(t),\ t \in [0, T)$, is mapped by the decoder $g(\cdot)$ to an L-dimensional vector $\widehat{\mathbf{X}}$.

We denote by A_i, the inverse image of $s_i(t),\ t \in [0, T)$, under the encoder map $\gamma(\cdot)$. The encoder partitions $\mathbb{R}^L$ into N encoding regions; we denote the partition by $\mathcal{A} = \{A_i,\ i = 1, 2, \ldots,\ N\}$ and the probability of the set A_i, under the source distribution by $P_i,\ i = 1, 2, \ldots,\ N$.

The problem we wish to solve is that of minimizing the per sample MSE, $E(\|\mathbf{X} - \widehat{\mathbf{X}}\|^2)/L$, by suitably selecting the encoder and decoder maps, subject to constraints on the average transmitted power $E(\int\limits_0^T S^2(t)dt)/T$ and the transmitted signal bandwidth[1].

We mention certain notational details before leaving this section. In the following we will use $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ to denote the Euclidean norm and the inner product, respectively. We will assume that all vectors are column vectors and will denote the trace of a square matrix by $tr(\cdot)$. The null space and the range space of a linear map $\mathbf{\Gamma}$ will be represented by $\mathcal{N}(\mathbf{\Gamma})$ and $\mathcal{R}(\mathbf{\Gamma})$, respectively. Upper case characters will be used for random quantities and boldface characters for vectors and matrices. We will use 0 for both the vector as well as the scalar zero; the meaning will be clear from the context. We will write $\mathbf{X} \geq 0$ if each component of $\mathbf{X}$ is non-negative and $\mathbf{X} < \infty$ if each component of $\mathbf{X}$ is finite. $\mathbf{R_{XY}}$ will be used to denote the covariance matrix of two random vectors $\mathbf{X}$ and $\mathbf{Y}$. Finally, we will use the superscript $\perp$ to denote the orthogonal complement of a linear subspace and the operator $\oplus$ for the direct sum of two linear subspaces.

II.2 Bandwidth Constraint

Our goal is to design a signal set, each signal of which has a duration of T seconds and is bandlimited in some sense. Since these signals are time-limited they cannot

[1] We will be more precise about the nature of the bandwidth constraint in Section II.2.

be strictly bandlimited. Several methods of imposing bandwidth constraints do exist in the literature. In [17] it is suggested that a signal set be designed subject to a constraint on the fractional out of band energy, i.e., if $S(f)$ is the Fourier transform of a transmitted signal $s(t)$, $0 \leq t < T$ and $H(f)$ $(h(t))$ is the frequency (impulse) response of a given bandpass filter, then the bandwidth constraint should have the form

$$\frac{\int_{-\infty}^{\infty} |H(f)S(f)|^2 df}{\int_{-\infty}^{\infty} |S(f)|^2 df} \geq \eta, \tag{II.1}$$

where $0 < \eta < 1$ is a suitably chosen constant.

The response of a general bandlimited filter to a time-limited signal is not time-limited. However, our decoder uses only that portion of the channel output which lies in the same interval as the transmitted signal. Hence, assuming that the transmitted signal lies in the interval $[0, T)$, the received SNR will be maximized if the energy of the channel filter response that lies in $[0, T)$, is maximized. Therefore, rather than maximize the energy in the frequency band as in (II.1), we seek to maximize the energy of the filter output that lies in the interval $[0, T)$.

We will now proceed to show that for this viewpoint of the signal bandwidth it suffices to consider a finite dimensional representation for the signal set, and that the bandwidth increases as the dimension of the signal set increases. Our development is based on Theorem 8.4.1 in [18].

Define $h(t, \tau)$ by

$$h(t,\tau) = \begin{cases} h(t-\tau)\ , & t, \tau \in [0, T) \\ 0\ , & \text{elsewhere,} \end{cases} \tag{II.2}$$

and define $R(\tau_1, \tau_2)$ by

$$R(\tau_1, \tau_2) = \int_0^T h(t, \tau_1) h(t, \tau_2) dt. \tag{II.3}$$

Assume that $\int_0^T \int_0^T h^2(t, \tau) d\tau \, dt < \infty$, i.e., $h(t, \tau)$ is square integrable. There exists a sequence of orthonormal functions $\phi_i(t)$, $i \in \mathbb{Z}^+$, $t \in [0, T)$ and a sequence of real numbers $1 \geq \mu_1 \geq \mu_2 \ldots > 0$, in one-to-one correspondence with these functions, such that

$$\int_0^T R(\tau_1, \tau_2)\phi_i(\tau_2)d\tau_2 = \mu_i \phi_i(\tau_1), \quad i \in \mathbb{Z}^+ . \tag{II.4}$$

Further, there exists another sequence of orthonormal functions $\theta_i(t)$, $t \in [0, T)$, such that

$$\int_0^T h(t, \tau)\phi_i(\tau)d\tau = \sqrt{\mu_i}\theta_i(t), \quad i \in \mathbb{Z}^+ . \tag{II.5}$$

It follows that if any signal $s(t)$, $t \in [0, T)$ can be expressed as

$$s(t) = \sum_{i=1}^{K} s_i \phi_i(t), \tag{II.6}$$

then the output of the filter $y(t)$ in the interval $[0, T)$, corresponding to an input $s(t)$, can be expressed as

$$y(t) = \sum_{i=1}^{K} s_i \sqrt{\mu_i}\theta_i(t). \tag{II.7}$$

For a given input signal $s(t)$, $t \in [0, T)$, expressed by (II.6) and having energy $\mathcal{E}_{in}$, let $\mathcal{E}_{out}$ be the energy at the output of the filter in the interval $[0, T)$. Then,

$$\frac{\mathcal{E}_{out}}{\mathcal{E}_{in}} = \frac{\sum\limits_{i=1}^{K} s_i^2 \mu_i}{\sum\limits_{i=1}^{K} s_i^2} \geq \mu_K . \tag{II.8}$$

Hence if we restrict the signal set to lie in the span of $\{\phi_i(t),\ i = 1, 2, \ldots,\ K\}$, it is guaranteed that the fraction μ_K of the signal energy will be recovered in the interval $[0, T)$. Hence for a given value of T the received signal energy in the interval $[0, T)$ will increase (or, equivalently, the bandwidth will decrease) as the value K decreases. Since T is proportional to L, we will use the ratio K/L as a measure of the signal bandwidth. We emphasize at this point that the ratio K/L is not a linear measure of the signal bandwidth expressed in Hz.

We can now represent the signal $S(t)$, noise $Z(t)$ and received waveform $U(t)$ by $\mathbf{S} = (S^1, S^2, \ldots,\ S^K)^T$, $\mathbf{Z} = (Z^1, Z^2, \ldots,\ Z^K)^T$ and $\mathbf{U} = (U^1, U^2, \ldots, U^K)^T$, respectively. Here S^i, Z^i and U^i denote the projections of $S(t), Z(t)$ and $U(t)$ onto the ith basis function $\phi_i(t)$, respectively. The signal set $\mathcal{S}$ will now be considered to consist of vectors $\mathbf{s}_i = (s_i^1, s_i^2, \ldots, s_i^K)^T$ that represent projections of the signals $s_i(t)$, $i = 1, 2, \ldots, N$, on the basis functions. It is a well-known fact that the above finite dimensional representations are sufficient for the problem that we wish to solve.

We note here that, in the sequel, we will ignore the effects of ISI by assuming that a 'genie' provides the receiver with the actual interfering portion of the filter output, thus allowing the receiver to subtract off the effects of ISI. In the real world the 'genie' can take the form of any of the ISI cancellation algorithms available [19]. The bandwidth constraint that we have imposed does, in a sense, restrict the effect of ISI since it minimizes the energy in the filter output that lies outside the signaling interval. Further, if $1/T$ is small as compared to the filter bandwidth, we would expect the effects of ISI to be small.

II.3 Problem Statement

Given an L-dimensional, stationary, zero-mean vector source $\{\mathbf{X}\}$, an additive white Gaussian vector channel having a variance of $N_0/2$ per channel dimension, and N, the cardinality of the signal set $\mathcal{S}$, we wish to minimize the MSE, $E\{\|\mathbf{X} - \widehat{\mathbf{X}}\|^2\}/L$, subject to the average energy constraint

$$\mathcal{E}(\mathcal{S}) = \frac{1}{L} E(\mathbf{S}^T\mathbf{S}) \leq \mathcal{E}_a = \frac{1}{L} P_a T \, , \tag{II.9}$$

and the bandwidth constraint

$$dim(\mathcal{S}) \leq K, \tag{II.10}$$

by suitably selecting the encoder $\gamma(\cdot)$, the decoder $g(\cdot)$ and the signal set $\mathcal{S}$. It is assumed that the value of K in (II.10) has been selected as described in Section II.2.

The parameters N, K and L can be easily related to parameters commonly used in the communications literature. For a given bandpass filter and constant source sample generation rate, the ratio K/L is a measure of the transmitted signal bandwidth. The source encoder rate is given by $R_s = \log N/L$. This rate places a fundamental lower limit on the average distortion that a source encoder can achieve. One is accustomed to interpreting R_s as a measure of the bandwidth required to transmit the source data. This is not the case here since for a fixed ratio K/L, the bandwidth is *independent* of R_s. The cardinality of the signal set, as we shall see later, is an indication of the encoding complexity.

II.4 Optimum Decoder and Approximations

From estimation theory it is well-known that for the squared-error distortion criterion, the optimum decoder must compute the conditional expectation of the source vector based on the channel output, i.e.,

$$g(\mathbf{u}) = E(\mathbf{X}|\mathbf{U} = \mathbf{u}). \tag{II.11}$$

It is also known that the optimum estimator is unbiased, i.e., $E(g(\mathbf{U})) = E(\mathbf{X}) = 0$.

By conditioning on the event that $\mathbf{s}_i$ is transmitted, it is easy to see that $g(\mathbf{u})$ can be expressed as a convex combination of the centroids of the encoding regions, i.e.,

$$g(\mathbf{u}) = \sum_{i=1}^{N} E(\mathbf{X}|i)P(\mathbf{s}_i|\mathbf{u}), \tag{II.12}$$

where $E(\mathbf{X}|i)$ is the centroid of the region A_i and $P(\mathbf{s}_i|\mathbf{u})$ is the probability that $\mathbf{s}_i$ is transmitted, given the channel output $\mathbf{u}$. Equation (II.12) can further be expressed as

$$g(\mathbf{u}) = \sum_{i=1}^{N} E(\mathbf{X}|i)\frac{p(\mathbf{u}|\mathbf{s}_i)P_i}{p(\mathbf{u})}, \tag{II.13}$$

where the lower case p's are used to represent probability density functions.

In order to compute the performance of this estimator and in order to be able to optimize the signal set, it is necessary to compute the variance of the estimate. The variance of $\widehat{\mathbf{X}}$, is given by

$$\mathrm{var}(\widehat{\mathbf{X}}) \quad = \quad E(g^T(\mathbf{U})g(\mathbf{U}))$$

$$= \int\limits_{\mathbb{R}^K} \sum_{i=1}^{N}\sum_{j=1}^{N} E(\mathbf{X}|i)^T E(\mathbf{X}|j) \frac{p(\mathbf{u}|\mathbf{s}_i)p(\mathbf{u}|\mathbf{s}_j)}{p(\mathbf{u})} P_i P_j d\mathbf{u}$$
$$= \sum_{i=1}^{N}\sum_{j=1}^{N} E(\mathbf{X}|i)^T E(\mathbf{X}|j) P_i P_j \int\limits_{\mathbb{R}^K} \frac{p(\mathbf{u}|\mathbf{s}_i)p(\mathbf{u}|\mathbf{s}_j)}{p(\mathbf{u})} d\mathbf{u}. \qquad \text{(II.14)}$$

The evaluation of the integral in (II.14) is the main bottleneck in evaluating the estimator variance.

Fortunately, for very noisy Gaussian channels, the optimum nonlinear estimator can be well approximated by a *linear* estimator. We now prove a result which is motivated by an earlier result due to Gardner [20], in order to make the above statement precise.

Theorem 2.1: Assume that $E(\mathbf{X}) = 0$, $E(\mathbf{S}) = 0$ and $E(\mathbf{S}^T\mathbf{S}) = \mathcal{E}$. Let $\mathbf{Z}$ be a zero mean, Gaussian, K-dimensional noise vector with autocovariance matrix $(N_0/2)\,\mathbf{I}$. The optimum estimate of the transmitted vector $\mathbf{X}$, based on the received vector $\mathbf{u}$, can be expressed as,

$$g(\mathbf{u}) = \mathbf{G}^T\mathbf{u} + \mathbf{o}(\sqrt{\mathcal{E}}, \mathbf{u}), \qquad \text{(II.15)}$$

where $\|\mathbf{o}(\sqrt{\mathcal{E}}, \mathbf{u})\|/\sqrt{\mathcal{E}} \longrightarrow 0$ as $\mathcal{E} \longrightarrow 0$, and $\mathbf{G}^T$ is a linear map, $\mathbf{G}^T : \mathbb{R}^K \longrightarrow \mathbb{R}^L$.

Proof: Rewrite (II.13) as

$$g(\mathbf{u}) = \frac{\sum\limits_{i=1}^{N} E(\mathbf{X}|i) p_{\mathbf{u}|\mathbf{s}_i}(\mathbf{u}|\mathbf{s}_i) P_i}{\sum\limits_{j=1}^{N} p_{\mathbf{u}|\mathbf{s}_j}(\mathbf{u}|\mathbf{s}_j) P_j}, \qquad \text{(II.16)}$$

where we use the subscript on p to identify the density being used. Now use the fact that $p_{\mathbf{u}|\mathbf{s}_i}(\mathbf{u}|\mathbf{s}_i) = p_{\mathbf{Z}}(\mathbf{u} - \mathbf{s}_i)$ and expand $p_{\mathbf{Z}}(\mathbf{u} - \mathbf{s}_i)$ in a Taylor series around $\mathbf{u}$ to get

$$p_{\mathbf{u}|\mathbf{s}_i}(\mathbf{u}|\mathbf{s}_i) = p_{\mathbf{Z}}(\mathbf{u} - \mathbf{s}_i) = p_{\mathbf{Z}}(\mathbf{u}) - \frac{dp_{\mathbf{Z}}(\mathbf{u})}{d\mathbf{u}}\mathbf{s}_i + o(\mathbf{s}_i, \mathbf{u}), \qquad \text{(II.17)}$$

where $|o(\mathbf{s}_i, \mathbf{u})|/\|\mathbf{s}_i\| \rightarrow 0$ as $\|\mathbf{s}_i\| \rightarrow 0$. Since $\mathbf{Z}$ is Gaussian with covariance matrix $(N_0/2)\,\mathbf{I}$, it follows that

$$\frac{dp_{\mathbf{Z}}(\mathbf{u})}{d\mathbf{u}} = \left(-\frac{2}{N_0}\right) p_{\mathbf{Z}}(\mathbf{u})\mathbf{u}^T. \qquad \text{(II.18)}$$

Now substitute (II.18) and (II.17) into (II.16) and use the fact that $E(\mathbf{X}) = 0$ and $E(\mathbf{S}) = 0$ to get,

$$\begin{aligned} g(\mathbf{u}) &= \frac{2 \sum_{i=1}^{N} P_i E(\mathbf{X}|i)\mathbf{s}_i^T \mathbf{u}/N_0 + \sum_{i=1}^{N} P_i E(\mathbf{X}|i) o(\mathbf{s}_i, \mathbf{u})/p_{\mathbf{z}}(\mathbf{u})}{1 + \sum_{i=1}^{N} P_i o(\mathbf{s}_i, \mathbf{u})/p_{\mathbf{z}}(\mathbf{u})} \\ &= 2\sum_{i=1}^{N} P_i E(\mathbf{X}|i)\mathbf{s}_i^T \mathbf{u}/N_0 + \mathbf{o}\left(\sqrt{\mathcal{E}}, \mathbf{u}\right). \end{aligned} \tag{II.19}$$

By identifying $\mathbf{G}^T = 2 \sum_{i=1}^{N} P_i E(\mathbf{X}|i)\mathbf{s}_i^T/N_0$, the theorem is proved. The details that lead to the last step have been proved in Appendix A.

We also present evidence to support our claim in II.2. We consider a uniformly distributed scalar source ($L = 1$) encoded by an N-level uniform scalar quantizer and transmitted by an N-level PAM signal set ($K = 1$). The map from the quantizer output to the signal set preserves the order, i.e., the leftmost quantizer output is mapped to the leftmost signal and so on. In both Figs. II.2a and II.2b the optimum estimator output, as given by II.12 is plotted as a function of the channel output over a region having a probability of 0.99. The estimator output was obtained by computing its value at 25 equally spaced points in the above mentioned region. In Fig. II.2a, $N = 4$ and the curves are presented for three values of the CSNR ($= 10\log(\mathcal{E}/LN_0)$). It is clearly seen that the optimum estimator tends to become linear as the CSNR decreases. In Fig. II.2b, CSNR= 7.0dB and curves are plotted for $N = 2, 4$ and 8. Here we notice that the optimum estimator tends to become linear as N increases. The results of this section provide the motivation for restricting attention to the class of linear estimator-decoders, and we do so in the sequel.

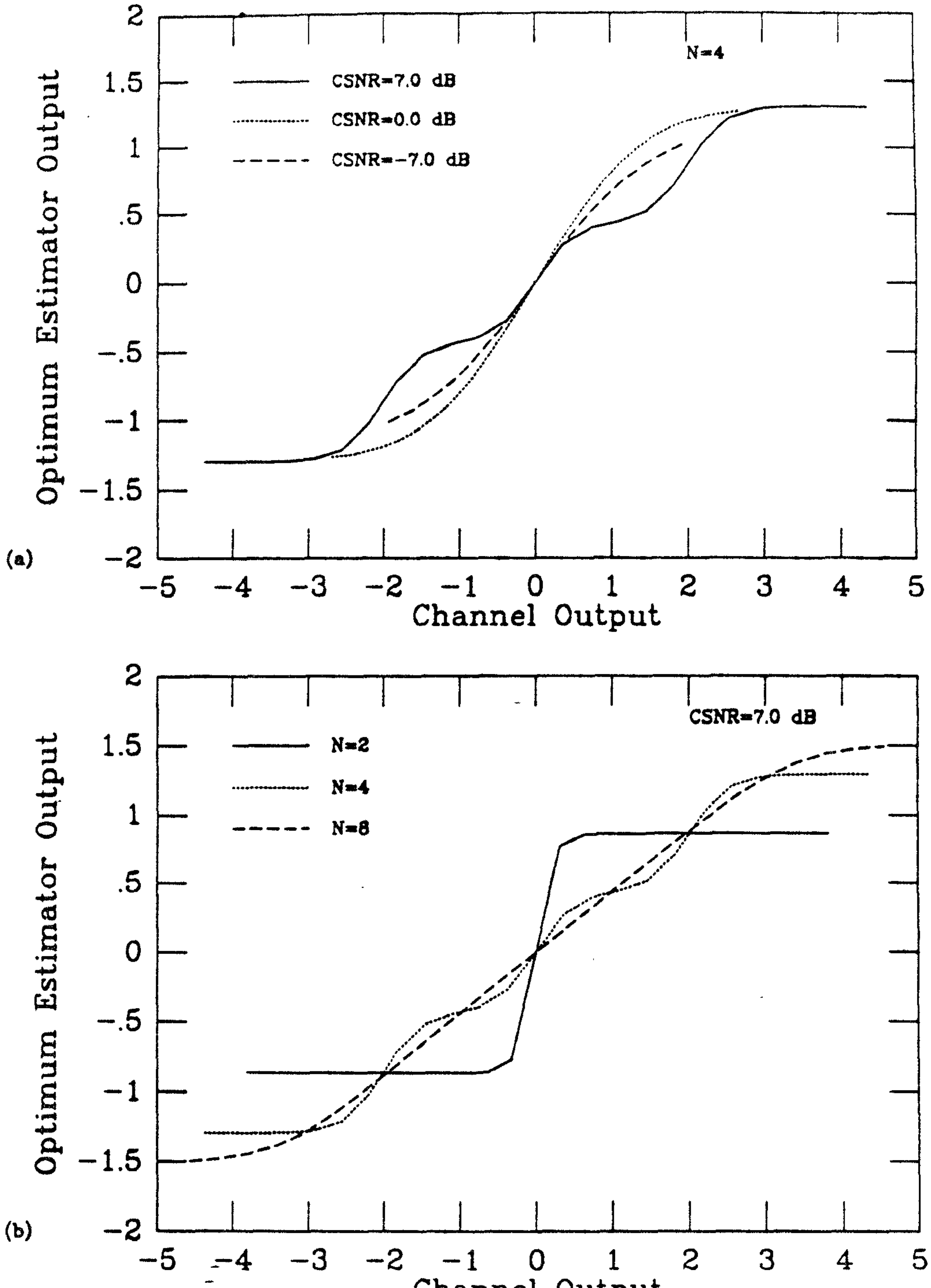

Figure II.2: Optimum Estimator Examples

III Optimum Signal Design for Linear Estimator-Based Decoders

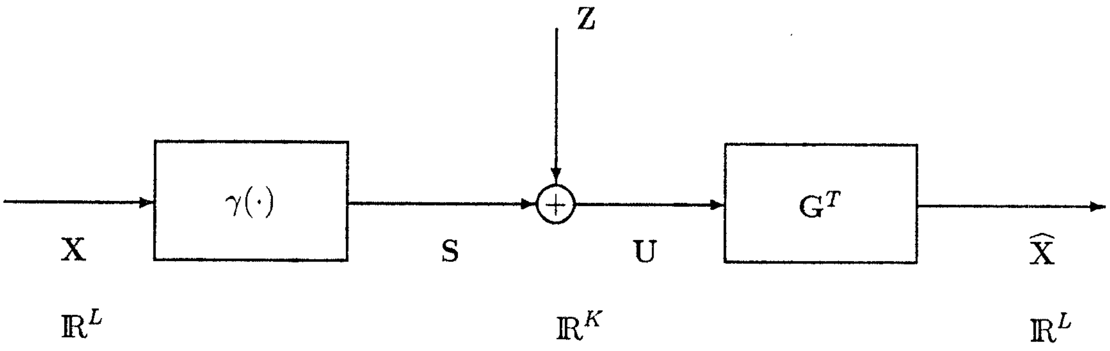

Figure III.1: System Block Diagram

We will now proceed to develop necessary conditions for optimality for the problem stated in Section II.3 with the restriction that the decoder be a linear map $\mathbf{G}^T : \mathbb{R}^K \to \mathbb{R}^L$. In order to do so we first state, without proof, some useful formulae that are simple to derive, and write a general expression for the MSE in terms of the parameters to be optimized.

With reference to Fig. III.1, assume that $\mathbf{X}$ and $\mathbf{Z}$ are independent and have zero means. The following relationships are easily verified:

$$\mathbf{R_{XU}} = \mathbf{R_{XS}} = \sum_{i=1}^{N} P_i E(\mathbf{X}|i)\mathbf{s}_i^T, \tag{III.1}$$

$$\mathbf{R}_{\mathbf{X}\widehat{\mathbf{X}}} = \mathbf{R_{XS}}\mathbf{G}, \tag{III.2}$$

$$\mathbf{R_{UU}} = \mathbf{R_{SS}} + \mathbf{R_{ZZ}}, \tag{III.3}$$

$$\mathbf{R}_{\widehat{\mathbf{X}}\widehat{\mathbf{X}}} = \mathbf{G}^T\mathbf{R_{UU}}\mathbf{G} = \mathbf{G}^T(\mathbf{R_{SS}} + \mathbf{R_{ZZ}})\mathbf{G}, \tag{III.4}$$

and

$$\mathbf{R_{SS}} = \sum_{i=1}^{N} P_i \mathbf{s}_i \mathbf{s}_i^T. \tag{III.5}$$

The MSE is given by

$$\frac{1}{L}E(tr(\mathbf{X}-\widehat{\mathbf{X}})(\mathbf{X}-\widehat{\mathbf{X}})^T) = \frac{1}{L}tr\left(\mathbf{R}_{\mathbf{XX}} - 2\mathbf{R}_{\mathbf{X}\widehat{\mathbf{X}}} + \mathbf{R}_{\widehat{\mathbf{X}}\widehat{\mathbf{X}}}\right)$$

$$= \frac{1}{L}tr\left(\mathbf{R}_{\mathbf{XX}} - 2\sum_{i=1}^{N} P_i E(\mathbf{X}|i)\mathbf{s}_i^T\mathbf{G} + \mathbf{G}^T\left(\sum_{i=1}^{N} P_i\mathbf{s}_i\mathbf{s}_i^T + \mathbf{R}_{\mathbf{ZZ}}\right)\mathbf{G}\right). \quad \text{(III.6)}$$

For completeness we state the energy constraint

$$\frac{1}{L}E(\mathbf{S}^T\mathbf{S}) \le \mathcal{E}_a, \qquad \text{(III.7)}$$

where it is assumed that $\mathcal{E}_a = P_aT/L$, for a given constant transmitter power P_a. Necessary conditions for optimality are to be developed with respect to the encoder map $\gamma(\cdot)$ and the linear decoder map $\mathbf{G}^T$. The encoder map $\gamma(\cdot)$ can be expressed as

$$\gamma(\mathbf{x}) = \sum_{i=1}^{N} I_{A_i}(\mathbf{x})\mathbf{s}_i, \qquad \text{(III.8)}$$

where I_{A_i} is the indicator function of the set A_i. From (III.8) it follows that the encoder is fully specified by the partition $\mathcal{A} = \{A_i,\ i = 1,2,\ldots,\ N\}$ and the signal set $\mathcal{S} = \{\mathbf{s}_i,\ i = 1,2,\ldots,\ N\}$. Therefore it suffices to develop necessary conditions for optimality with respect to the partition $\mathcal{A}$, the signal set $\mathcal{S}$ and the decoder $\mathbf{G}^T$. In the sequel, the MSE per source sample and the average energy per source sample for the system in Fig. III.1 will be denoted by $D(\mathcal{A},\mathcal{S},\mathbf{G}^T)$ and $\mathcal{E}(\mathcal{A},\mathcal{S})$, respectively[2].

III.1 Optimum Encoder Partition

Assume that the signal set $\mathcal{S}$ and the decoder $\mathbf{G}^T$ are fixed. Our goal is to minimize (III.6) subject to the energy constraint (III.7), by suitably choosing an encoder partition $\mathcal{A}$. Note that the encoder partition affects the energy constraint through

[2] For notational brevity we will frequently suppress the fixed arguments of D and $\mathcal{E}$. For example, if the signal set and decoder are fixed, we will use the notation $D(\mathcal{A})$ and $\mathcal{E}(\mathcal{A})$ to represent the MSE and the average energy, respectively.

the probabilities of the sets A_i, $i = 1, 2, \ldots, N$; hence the energy constraint *must* be imposed.

In order to make the discussion in this section clear, we make explicit the dependence of the transmitted signal vector on the source vector by writing $\mathbf{s}(\mathbf{x})$. The first step is to construct the Lagrangian

$$L(\mathcal{A}, \beta) = D(\mathcal{A}) + \beta(\mathcal{E}(\mathcal{A}) - \mathcal{E}_a), \tag{III.9}$$

which can be written in integral form as

$$L(\mathcal{A}, \beta) = \frac{1}{L} \int\limits_{\mathbb{R}^L} F(\mathbf{x}, \beta) p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \tag{III.10}$$

where

$$F(\mathbf{x}, \beta) = E(tr(\mathbf{x} - \widehat{\mathbf{X}})(\mathbf{x} - \widehat{\mathbf{X}})^T | \mathbf{X} = \mathbf{x}) + \beta(tr(\mathbf{s}(\mathbf{x})\mathbf{s}(\mathbf{x})^T) - L\mathcal{E}_a). \tag{III.11}$$

A well-known sufficient condition for optimality [21], states that if there exists a partition $\mathcal{A}^*$ and a multiplier $\beta^* \geq 0$, which together satisfy

$$L(\mathcal{A}^*, \beta) \leq L(\mathcal{A}^*, \beta^*) \leq L(\mathcal{A}, \beta^*), \quad \forall \beta \geq 0, \ \forall \mathcal{A} \ , \tag{III.12}$$

then $\mathcal{A}^*$ is a globally optimum partition subject to the imposed energy constraint. For a given value of $\beta \geq 0$, we will determine in Theorem 3.1, a partition $\mathcal{A}(\beta)$, for which

$$L(\mathcal{A}(\beta), \beta) \leq L(\mathcal{A}, \beta), \quad \forall \mathcal{A} \ . \tag{III.13}$$

We will then prove in Theorem 3.2, that $\exists \ \beta^* \geq 0$ such that

$$L(\mathcal{A}(\beta^*), \beta) \leq L(\mathcal{A}(\beta^*), \beta^*) \ , \quad \forall \beta \geq 0 \ . \tag{III.14}$$

Therefore, upon defining $\mathcal{A}^* = \mathcal{A}(\beta^*)$, where β^* is chosen to satisfy (III.14), it follows that (III.12) is satisfied by the pair $(\mathcal{A}^*, \beta^*)$ and hence $\mathcal{A}^*$ is a globally optimum partition.

Theorem 3.1: Given $\beta \geq 0$, define $A_i(\beta)$, $i = 1, 2, \ldots, N$, by

$$A_i(\beta) = \left\{ \mathbf{x} : 2\langle \mathbf{x}, \mathbf{G}^T(\mathbf{s}_i - \mathbf{s}_j) \rangle \geq \|\mathbf{G}^T\mathbf{s}_i\|^2 - \|\mathbf{G}^T\mathbf{s}_j\|^2 + \beta(\|\mathbf{s}_i\|^2 - \|\mathbf{s}_j\|^2),\ \forall\, j \neq i \right\}. \tag{III.15}$$

Let $\mathcal{A}(\beta) = \{A_i(\beta),\ i = 1, 2, \ldots,\ N\}$. Then $\mathcal{A}(\beta)$ and β together satisfy (III.13).

Proof: For a given $\beta \geq 0$, we wish to determine a partition $\mathcal{A}$ that minimizes $L(\mathcal{A}, \beta)$. Since $p_{\mathbf{X}}(\mathbf{x}) \geq 0$ for all $\mathbf{x}$, it suffices to minimize $F(\mathbf{x}, \beta)$ for all $\mathbf{x} \in \mathbb{R}^L$. For each $\mathbf{x} \in \mathbb{R}^L$, $F(\mathbf{x}, \beta)$ can take on one of N different values depending on the signal to which $\mathbf{x}$ is mapped by the encoder[3]. Denote by $F_i(\mathbf{x}, \beta)$ the value of $F(\mathbf{x}, \beta)$ under the assumption that $\mathbf{x}$ is mapped to $\mathbf{s}_i$. Since the optimum encoder must map $\mathbf{x}$ to $\mathbf{s}_i$, where i is the index that minimizes $F_i(\mathbf{x}, \beta)$, it follows that

$$A_i(\beta) = \{\mathbf{x} : F_i(\mathbf{x}, \beta) \leq F_j(\mathbf{x}, \beta),\ \forall\, j \neq i\}. \tag{III.16}$$

It is easy to see that $F_i(\mathbf{x}, \beta)$ is given by,

$$F_i(\mathbf{x}, \beta) = \|\mathbf{x}\|^2 - 2\langle \mathbf{x}, \mathbf{G}^T\mathbf{s}_i \rangle + \|\mathbf{G}^T\mathbf{s}_i\|^2 + tr(\mathbf{G}^T R_{\mathbf{ZZ}}\mathbf{G}) + \beta(\|\mathbf{s}_i\|^2 - L\mathcal{E}_a). \tag{III.17}$$

Simple algebra then yields (III.15), thus proving the theorem.

It can be proved using (III.17) that the sets $A_i(\beta)$ are convex sets that are separated from each other by hyperplanes. To see this, define $A_{ij}(\beta)$ by

$$A_{ij}(\beta) = \{\mathbf{x} :\ 2\langle \mathbf{x}, \mathbf{G}^T(\mathbf{s}_i - \mathbf{s}_j) \rangle \geq \|\mathbf{G}^T\mathbf{s}_i\|^2 - \|\mathbf{G}^T\mathbf{s}_j\|^2 + \beta(\|\mathbf{s}_i\|^2 - \|\mathbf{s}_j\|^2)\}. \tag{III.18}$$

Then

$$A_i(\beta) = \bigcap_{\substack{j=1 \\ j \neq i}}^{N} A_{ij}(\beta), \tag{III.19}$$

and $A_{ij}(\beta)$ as defined in (III.18) is an open half space of $\mathbb{R}^L$, the separating hyperplane, $H_{ij}(\beta)$, of which is defined by,

$$H_{ij}(\beta) = \left\{ \mathbf{x}\ :\ 2\langle \mathbf{x}, \mathbf{G}^T(\mathbf{s}_i - \mathbf{s}_j) \rangle = \|\mathbf{G}^T\mathbf{s}_i\|^2 - \|\mathbf{G}^T\mathbf{s}_j\|^2 + \beta(\|\mathbf{s}_i\|^2 - \|\mathbf{s}_j\|^2) \right\}. \tag{III.20}$$

[3] We assume here that the map from the source sample to the modulation signal is non-random; it is easy to show that the optimum map is deterministic.

Since the $A_{ij}(\beta)$ are convex, it follows from (III.19) that $A_i(\beta)$ is also convex. Equation (III.20) is the equation of a hyperplane that is perpendicular to the line joining the points $\mathbf{G}^T\mathbf{s}_i$ and $\mathbf{G}^T\mathbf{s}_j$ and whose shortest distance to the origin is given by $\left(\|\mathbf{G}^T\mathbf{s}_i\|^2 - \|\mathbf{G}^T\mathbf{s}_j\|^2 + \beta(\|\mathbf{s}_i\|^2 - \|\mathbf{s}_j\|^2)\right)/2\|\mathbf{G}^T(\mathbf{s}_i - \mathbf{s}_j)\|$. Hence, for a fixed decoder and signal set, the only effect that the multiplier β has on the separating hyperplane $H_{ij}(\beta)$ is that of moving the hyperplane without changing its orientation, i.e., $H_{ij}(\beta)$ and $H_{ij}(\beta')$ are parallel hyperplanes separated by a distance that is proportional to $(\beta - \beta')$.

We now wish to prove the existence of a $\beta^* \geq 0$ for which (III.14) holds. This result is proved in Theorem 3.2, but in order to prove this theorem we will need the following Lemma. For a given $\mathcal{S}$, $\mathbf{G}^T$ and $\beta \geq 0$, let $\mathcal{E}(\beta)$ and $D(\beta)$ be the average transmitted energy and the MSE, respectively, corresponding to the partition $\mathcal{A}(\beta)$.

Lemma 3.1: $\mathcal{E}(\beta)$ is a non-increasing continuous function of the Lagrange multiplier β, under the assumption that the source distribution $P_{\mathbf{X}}$ is absolutely continuous.

Proof: (i) (<u>$\mathcal{E}(\beta)$ is non-increasing in β</u>)

For a fixed signal set $\mathcal{S}$, let $\beta > \beta'$. Since $(\mathcal{A}(\beta), \beta)$ and $(\mathcal{A}(\beta'), \beta')$ satisfy (III.13), it follows by using (III.9) that

$$D(\beta) + \beta(\mathcal{E}(\beta) - \mathcal{E}_a) \leq D(\beta') + \beta(\mathcal{E}(\beta') - \mathcal{E}_a), \tag{III.21}$$

and that

$$D(\beta') + \beta'(\mathcal{E}(\beta') - \mathcal{E}_a) \leq D(\beta) + \beta'(\mathcal{E}(\beta) - \mathcal{E}_a). \tag{III.22}$$

Equations (III.21) and (III.22), in turn, imply that

$$\beta(\mathcal{E}(\beta') - \mathcal{E}(\beta)) \geq \beta'(\mathcal{E}(\beta') - \mathcal{E}(\beta)), \tag{III.23}$$

and since $\beta > \beta'$ it follows that $\mathcal{E}(\beta') \geq \mathcal{E}(\beta)$, which proves that $\mathcal{E}(\beta)$ is a non-increasing function of β.

(Remark: This result is useful in the implementation of an algorithm to determine the optimum partition.)

(ii) (continuity of the $\mathcal{E}(\beta)$)

It suffices to prove that the probabilities of the partition sets $A_i(\beta)$ are continuous functions of β.

The proof proceeds as follows:

$$|P(A_i(\beta)) - P(A_i(\beta'))| \leq P(A_i(\beta)\Delta A_i(\beta')), \tag{III.24}$$

where Δ denotes the symmetric difference between two sets. Then, it follows from (III.19) that

$$A_i(\beta)\Delta A_i(\beta') \subset \bigcup_{j\neq i}(A_{ij}(\beta)\Delta A_{ij}(\beta')). \tag{III.25}$$

$A_{ij}(\beta)\Delta A_{ij}(\beta')$ is a region in $\mathbb{R}^L$ sandwiched between two parallel hyperplanes whose separation, as we have already mentioned, is proportional to $(\beta - \beta')$. It is easy to see, assuming that $P_{\mathbf{X}}$ is absolutely continuous, that given $\epsilon > 0$, a $\delta > 0$ can be chosen such that $|\beta - \beta'| < \delta$ implies $P(A_{ij}(\beta)\Delta A_{ij}(\beta')) < \epsilon/N$, $\forall j \neq i$. Hence from (III.24) and (III.25) it follows that $|P(A_i(\beta)) - P(A_i(\beta'))| < \epsilon$, thus proving continuity of $P(A_i(\beta))$ w.r.t. β.

Theorem 3.2: Assume $\mathcal{S}$ is such that $\min_{\mathbf{s}\in\mathcal{S}} \|\mathbf{s}\|^2 < L\mathcal{E}_a$ and that the conditions of Lemma 3.1 hold. Then $\exists\ \beta^* \geq 0$ that satisfies (III.14).

Proof: For $\beta = 0$ either $\mathcal{E}(\beta) \leq \mathcal{E}_a$ or $\mathcal{E}(\beta) > \mathcal{E}_a$. If $\mathcal{E}(0) \leq \mathcal{E}_a$ then (III.14) holds for $\beta^* = 0$. If $\mathcal{E}(0) > \mathcal{E}_a$, then $\exists\ \beta_0 > 0$ such that $\mathcal{E}(\beta_0) < \mathcal{E}_a$. In order to prove this assume that k is the index of the minimum energy signal in $\mathcal{S}$. To simplify the details of the proof we will assume that no other signal in $\mathcal{S}$ has the same energy; this assumption is *not* necessary for the theorem to hold. By Theorem 3.1,

$$A_k(\beta) = \{\mathbf{x} :\ 2\langle\mathbf{x}, \mathbf{G}^T(\mathbf{s}_j - \mathbf{s}_k)\rangle \leq \|\mathbf{G}^T\mathbf{s}_j\|^2 - \|\mathbf{G}^T\mathbf{s}_k\|^2 + \beta(\|\mathbf{s}_j\|^2 - \|\mathbf{s}_k\|^2),\ \forall\, j \neq k\}. \tag{III.26}$$

Define[4]

$$\tilde{A}_k(\beta) = \left\{ \mathbf{x} : \quad \|\mathbf{x}\| \quad \le \quad \frac{\min\limits_{j\neq k} \left(\|\mathbf{G}^T\mathbf{s}_j\|^2 - \|\mathbf{G}^T\mathbf{s}_k\|^2 + \beta(\|\mathbf{s}_j\|^2 - \|\mathbf{s}_k\|^2)\right)}{2\max\limits_j \|\mathbf{G}^T(\mathbf{s}_j - \mathbf{s}_k)\|} \right\}. \tag{III.27}$$

It is clear that $\tilde{A}_k(\beta) \subset A_k(\beta)$. Further, since $(\|\mathbf{s}_j\|^2 - \|\mathbf{s}_k\|^2) > 0$, $\forall j \neq k$, it follows that for any $\delta > 0$, $\exists\, \beta(\delta)$ such that $\beta > \beta(\delta)$ implies $P(\tilde{A}_k(\beta)) > 1 - \delta$ and hence $\mathcal{E}(\beta) \le \|\mathbf{s}_k\|^2 + \delta(\max_j \|\mathbf{s}_j\|^2)$. On the other hand, since $\|\mathbf{s}_k\|^2 < L\mathcal{E}_a$, it is possible to find a $\delta > 0$ such that $(\|\mathbf{s}_k\|^2 + \delta(\max_j \|\mathbf{s}_j\|^2)) < L\mathcal{E}_a$, which implies that $\mathcal{E}(\beta) < \mathcal{E}_a$, for $\beta > \beta(\delta)$. This establishes the existence of $\beta_0 > 0$.

Since $\mathcal{E}(\beta)$ is continuous and $\mathcal{E}(0) > \mathcal{E}_a > \mathcal{E}(\beta_0)$ for some β_0, it follows that $\exists\, \beta^* \in (0, \beta_0)$ such that $\mathcal{E}(\beta^*) = \mathcal{E}_a$. For this value of β^*, $L(\mathcal{A}(\beta^*), \beta) = D(\mathcal{A}(\beta^*)) = L(\mathcal{A}(\beta^*), \beta^*)$ and hence (III.14) holds.

It is instructive to study the constraint that the linear decoder places on the optimum encoder partition. As we prove in the next theorem, it turns out that in some cases the optimum partition is equivalent to the partition of a lower dimensional Euclidean space.

Theorem 3.3: For a given decoder $\mathbf{G}^T$ and signal set $\mathcal{S}$, let $\mathcal{A}$ be the optimum partition and let β be the corresponding Lagrange multiplier. Then the null space of $\mathbf{G}$, $\mathcal{N}(\mathbf{G}) \subset A_i$, where i is the index that satisfies

$$\|\mathbf{G}^T\mathbf{s}_i\|^2 + \beta\|\mathbf{s}_i\|^2 \le \|\mathbf{G}^T\mathbf{s}_j\|^2 + \beta\|\mathbf{s}_j\|^2, \ \forall j \neq i. \tag{III.28}$$

Further, if $\mathbf{x}$ is mapped by the encoder to $\mathbf{s}_j$, then so is $\mathbf{x} + \mathbf{x}'$, $\forall \mathbf{x}' \in \mathcal{N}(\mathbf{G})$, $\forall j$.

Proof: Assume $\mathbf{x} \in \mathcal{N}(\mathbf{G})$. Then since $\langle \mathbf{x}, \mathbf{G}^T(\mathbf{s}_i - \mathbf{s}_j)\rangle = \langle \mathbf{G}\mathbf{x}, (\mathbf{s}_i - \mathbf{s}_j)\rangle = 0$, it follows from (III.15) that $\mathbf{x} \in A_i(\beta)$, for the index i which satisfies (III.28). Further, assume that $\mathbf{x} \in A_k$ and let $\mathbf{x}' \in \mathcal{N}(\mathbf{G})$. Since $\langle \mathbf{x} + \mathbf{x}', \mathbf{G}^T(\mathbf{s}_k - \mathbf{s}_j)\rangle = \langle \mathbf{G}(\mathbf{x}+\mathbf{x}'), (\mathbf{s}_k - \mathbf{s}_j)\rangle = \langle \mathbf{G}\mathbf{x}, (\mathbf{s}_k - \mathbf{s}_j)\rangle = \langle \mathbf{x}, \mathbf{G}^T(\mathbf{s}_k - \mathbf{s}_j)\rangle$ it follows by using (III.15) that $\mathbf{x} + \mathbf{x}' \in A_k(\beta)$.

[4] As we will see later (Lemma 3.2), it suffices to choose $\mathbf{s}_i$, $i = 1, 2, \ldots, N$, so that they lie in $\mathcal{N}(\mathbf{G}^T)^{\perp}$, hence $\max\limits_j \|\mathbf{G}^T(\mathbf{s}_j - \mathbf{s}_k)\| \neq 0$.

This theorem indicates that the encoder partition is insensitive to translations of the source vector along the basis vectors of $\mathcal{N}(\mathbf{G})$, i.e., it is impossible to cross the boundary of any of the encoding regions by traversing a path parallel to vectors in $\mathcal{N}(\mathbf{G})$. Due to this it is possible to significantly reduce the computation required during the encoding procedure in cases where $\dim(\mathcal{N}(\mathbf{G}))$ is large, as the following example illustrates.

Suppose we wish to design an encoder for a source vector of dimension 2 and a signal set of dimension 1. Then $\mathcal{N}(G)$ has a dimension of 1. Assume that it is oriented as shown in Fig. III.2. The above theorem implies that the encoder

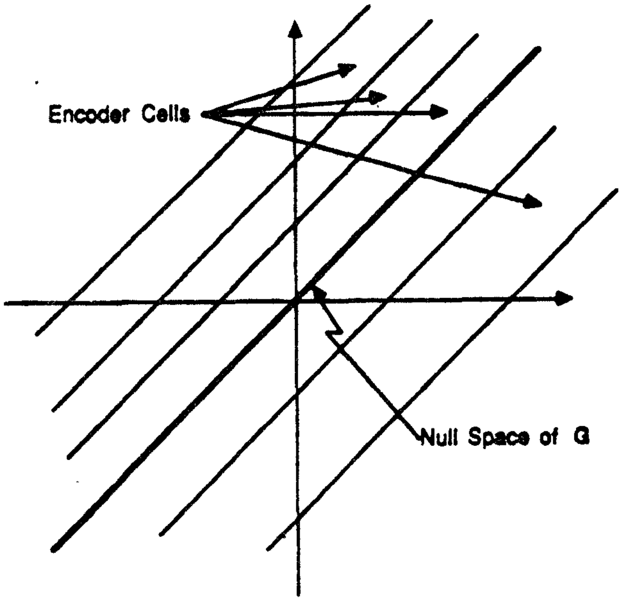

Figure III.2: Optimum Encoder Partition

partition boundaries are lines parallel to $\mathcal{N}(G)$, i.e., in this case the source encoder partition is equivalent to that of a scalar quantizer operating on a source vector that has been transformed using a linear transformation, hence a less complex search is required than would be for a general 2-D VQ.

III.2 Optimum Signal Set

Assume that the partition $\mathcal{A}$ and the decoder $\mathbf{G}^T$ are fixed. We wish to minimize the average distortion given by (III.6) subject to the energy constraint (III.7), by suitably choosing a signal set $\mathcal{S}$.

The Lagrangian for the optimization problem is given by

$$\begin{aligned} L(\mathcal{S},\psi) &= D(\mathcal{S}) + \psi(\mathcal{E}(\mathcal{S}) - \mathcal{E}_a) \\ &= \frac{1}{L} tr \left(\mathbf{R_{XX}} - 2\sum_{i=1}^{N} P_i E(\mathbf{X}|i)\mathbf{s}_i^T \mathbf{G} + \mathbf{G}^T(\mathbf{R_{SS}} + \mathbf{R_{ZZ}})\mathbf{G} \right) \\ &\quad + \psi \left(tr \left(\frac{1}{L} \sum_{i=1}^{N} P_i \mathbf{s}_i \mathbf{s}_i^T \right) - \mathcal{E}_a \right). \end{aligned} \tag{III.29}$$

Differentiate $L(\mathcal{S},\psi)$ with respect to $\mathbf{s}_i$, $i = 1, 2, \ldots, N$, and set the derivatives to zero in order to arrive at the following necessary conditions for optimality:

$$(\mathbf{G}\mathbf{G}^T + \psi\mathbf{I})\mathbf{s}_i = \mathbf{G}E(\mathbf{X}|i), \; i = 1, 2, \ldots, N, \tag{III.30}$$

$$\psi \geq 0, \tag{III.31}$$

and

$$\psi \left(tr \left(\frac{1}{L} \sum_{i=1}^{N} P_i \mathbf{s}_i \mathbf{s}_i^T \right) - \mathcal{E}_a \right) = 0. \tag{III.32}$$

It is important to note that $E(\mathbf{X}|i)$ is the centroid of A_i and hence (III.30) implies that the optimum signals used to transmit the source vectors in A_i are *linearly* related to the centroids of the region A_i. Hence there exists a linear map[5] $\mathbf{\Gamma}: \mathbb{R}^L \rightarrow \mathbb{R}^K$, such that,

$$\mathbf{s}_i = \mathbf{\Gamma} E(\mathbf{X}|i), \quad i = 1, 2, \ldots, N. \tag{III.33}$$

Thus the problem of optimum signal design is equivalent to that of selecting an optimal linear map $\mathbf{\Gamma}$ as in (III.33). From (III.30), it follows that the optimum

[5] This is a non-trivial statement. If $\psi > 0$ then $(\mathbf{G}\mathbf{G}^T + \psi I)$ is nonsingular, hence invertible and the existence of $\mathbf{\Gamma}$ follows. If $\psi = 0$, then $(\mathbf{G}\mathbf{G}^T + \psi I)$ may not be invertible. However, it is possible to show that there exists an optimal signal set that lies in $N(\mathbf{G}^T)^{\perp}$ from which follows the existence of $\mathbf{\Gamma}$.

linear map $\mathbf{\Gamma}$, must satisfy $(\mathbf{G}\mathbf{G}^T + \psi I)\mathbf{\Gamma} = \mathbf{G}$. We also note here that the optimum encoder can be factored into a vector quantizer (VQ) followed by a linear map from the VQ centroids to the modulation signal set, thus allowing us to depict the optimum system as in Fig. III.3. Note that as a consequence of (III.33), the signal set lies in the range space of the linear map $\mathbf{\Gamma}$. Since the dimension of the range space of $\mathbf{\Gamma}$ satisfies $\dim(\mathcal{R}(\mathbf{\Gamma})) \leq L$, it follows that if we had chosen $K > L$ then at least $K - L$ dimensions of the signal space would not be utilized, which in turn implies that the optimum signal set is incapable of resulting in a bandwidth expansion. Needless to say, this restriction on the bandwidth arises due to the linear decoder assumption. In the sequel, we will therefore assume that $K \leq L$.

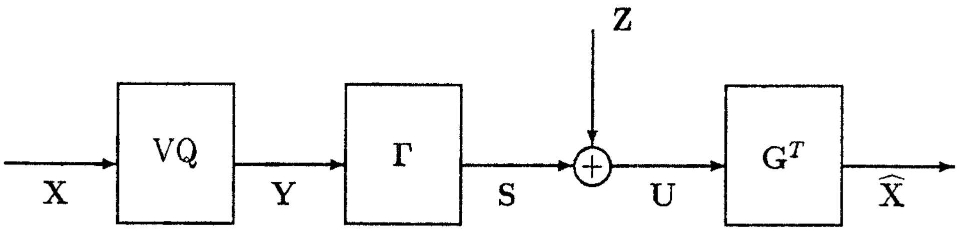

Figure III.3: Optimum System

III.3 Optimum Decoder

Assume that the encoder partition $\mathcal{A}$ and the signal set $\mathcal{S}$ are fixed and that we wish to determine the optimum decoder $\mathbf{G}^T$. Note that the choice of the decoder does not affect the energy constraint, hence we have an unconstrained optimization problem to solve. The solution to this problem is well known in linear estimation theory and is obtained by using the orthogonality principle which states that

$$E((\mathbf{X} - \mathbf{G}^T\mathbf{U})(\mathcal{L}\mathbf{U})^T) = 0, \qquad \text{(III.34)}$$

for any $\mathcal{L}$ in the space of linear maps from $\mathbb{R}^K$ to $\mathbb{R}^L$. The optimum decoder matrix is obtained by solving (III.34), and is given by

$$\mathbf{G}^T = \mathbf{R}_{\mathbf{XU}}\mathbf{R}_{\mathbf{UU}}^{-1}, \qquad \text{(III.35)}$$

from which it follows by using (III.1) and (III.3) that

$$\mathbf{G}^T = \mathbf{R}_{\mathbf{XS}}(\mathbf{R}_{\mathbf{SS}} + \mathbf{R}_{\mathbf{ZZ}})^{-1}. \qquad \text{(III.36)}$$

Our next goal is to simultaneously solve the necessary conditions developed in Sections III.1–III.3, in order to determine a solution to the necessary conditions for optimality. It turns out that the necessary conditions for the optimum signal set and decoder can be solved analytically; we shall do so in the next section, following which we will describe an iterative algorithm for determining a system that satisfies all the necessary conditions for optimality.

III.4 Simultaneous Solution of the Optimum Signal Set and the Optimum Decoder

In this section we will simultaneously solve (III.30)–(III.32) and (III.36). We will do so by first stating and then solving a problem which has an identical solution. With reference to Fig. III.4, let $\mathbf{Y}$ be an L-dimensional random vector which takes values in the discrete set $\{E(\mathbf{X}|i),\ i = 1, 2, \ldots, N\}$ with corresponding probabilities P_i, $i = 1, 2, \ldots, N$. We wish to determine an optimum linear encoder $\mathbf{\Gamma}$ and a linear decoder $\mathbf{G}^T$ so as to minimize $E(tr(\mathbf{Y} - \widehat{\mathbf{Y}})(\mathbf{Y} - \widehat{\mathbf{Y}})^T)$, subject to an energy constraint $tr(\mathbf{\Gamma}\mathbf{R}_{\mathbf{YY}}\mathbf{\Gamma}^T) \leq L\mathcal{E}_a$.

Theorem 3.4: Let $\mathbf{\Gamma}$ and $\mathbf{G}^T$, as described above, minimize $D' = E(tr(\mathbf{Y} - \widehat{\mathbf{Y}})(\mathbf{Y} - \widehat{\mathbf{Y}})^T)$ subject to $tr(\mathbf{\Gamma}\mathbf{R}_{\mathbf{YY}}\mathbf{\Gamma}^T) \leq L\mathcal{E}_a$. Then $\mathbf{\Gamma}$ and $\mathbf{G}^T$ simultaneously satisfy (III.30)–(III.32) and (III.36).

Proof: We can write

$$D' = \frac{1}{L}tr\left(\mathbf{R}_{\mathbf{YY}} - 2\mathbf{R}_{\mathbf{Y}\widehat{\mathbf{Y}}} + \mathbf{R}_{\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}}\right), \qquad \text{(III.37)}$$

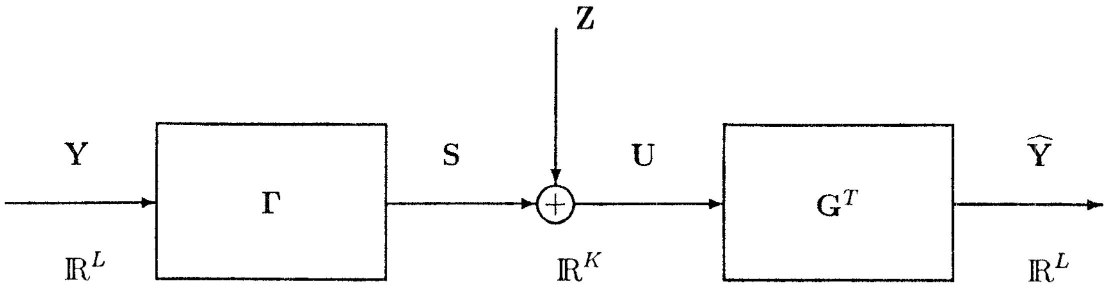

Figure III.4: An Equivalent System

which simplifies to

$$D' = \frac{1}{L}tr\left(\mathbf{R_{YY}} - 2\sum_{i=1}^{N} P_i E(\mathbf{X}|i)\mathbf{s}_i^T\mathbf{G} + \mathbf{G}^T(\sum_{i=1}^{N} P_i\mathbf{s}_i\mathbf{s}_i^T + \mathbf{R_{ZZ}})\mathbf{G}\right), \qquad \text{(III.38)}$$

where $\mathbf{s}_i = \mathbf{\Gamma}E(\mathbf{X}|i)$, $i = 1, 2, \ldots,\ N$. Except for the first term, $tr(R_{\mathbf{YY}})$, which does not depend on $\mathbf{\Gamma}$ and $\mathbf{G}$, (III.38) is identical to (III.6). Further, the average transmitted energy is identical for the two problems. Hence $\mathbf{\Gamma}$ and $\mathbf{G}$ simultaneously satisfy (III.30)-(III.32) and (III.36), thus proving the theorem.

We will now proceed to determine the optimum linear encoder map $\mathbf{\Gamma}$ and decoder $\mathbf{G}^T$ that minimize $E\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2$ for the system in Fig. III.4 subject to an average energy constraint. For future reference, the MSE expression for the system in Fig. III.4, is given by

$$D' = \frac{1}{L}tr\left(\mathbf{R_{YY}} - 2\mathbf{R_{YY}}\mathbf{\Gamma}^T\mathbf{G} + \mathbf{G}^T(\mathbf{\Gamma}\mathbf{R_{YY}}\mathbf{\Gamma}^T + \mathbf{R_{ZZ}})\mathbf{G}\right), \qquad \text{(III.39)}$$

where D and D' are related by

$$D = D' + \frac{1}{L}tr\left(\mathbf{R_{XX}} - \mathbf{R_{YY}}\right). \qquad \text{(III.40)}$$

We define the source encoder distortion by $\frac{1}{L}tr(\mathbf{R_{XX}} - \mathbf{R_{YY}})$ and the channel distortion by D'. The energy constraint is given by

$$\frac{1}{L}tr\left(\mathbf{\Gamma}\mathbf{R_{YY}}\mathbf{\Gamma}^T\right) \leq \mathcal{E}_a. \qquad \text{(III.41)}$$

The optimum encoder map, for a given decoder $\mathbf{G}^T$ is obtained by solving

$$\left(\mathbf{G}\mathbf{G}^T + \psi\mathbf{I}\right)\mathbf{\Gamma} = \mathbf{G}, \tag{III.42}$$

$$\psi \geq 0, \tag{III.43}$$

and

$$\psi\left(tr\left(\mathbf{\Gamma}\mathbf{R_{YY}}\mathbf{\Gamma}^T\right) - L\mathcal{E}_a\right) = 0. \tag{III.44}$$

The optimum decoder map for a given encoder is given by

$$\mathbf{G}^T = \mathbf{R_{YS}}(\mathbf{\Gamma}\mathbf{R_{YY}}\mathbf{\Gamma}^T + \mathbf{R_{ZZ}})^{-1}. \tag{III.45}$$

Our goal, now, is to solve (III.42)-(III.44) and (III.45) simultaneously. In order to solve these equations we will proceed as follows. We will first show that $\mathcal{R}(\mathbf{G}) = \mathcal{R}(\mathbf{\Gamma})$ and $\mathcal{N}(\mathbf{G}) = \mathcal{N}(\mathbf{\Gamma})$, where $\mathbf{\Gamma}$ is the optimum encoder for a given decoder $\mathbf{G}^T$. This fact and (III.42) will enable us to show that there exist bases for $\mathbb{R}^K$ and $\mathbb{R}^L$, called the *singular bases* of $\mathbf{G}^T$, in which $\mathbf{G}^T$, $\mathbf{G}\mathbf{G}^T$ *and* $\mathbf{\Gamma}$, the optimum encoder for the given decoder $\mathbf{G}^T$, have zero off-diagonal terms. Finally, we will prove that if $\mathbf{G}^T$ is the optimum decoder map, the singular basis for $\mathbb{R}^L$ is the set of eigenvectors of $\mathbf{R_{YY}}$. This will enable us to express the overall distortion as a summation, each term of which is a convex $\bigcup$ function of the average energy transmitted on a given channel dimension. By determining the optimum distribution of energy among the channel dimensions, we will be able to determine the optimum encoder map, following which the optimum decoder map will be determined..The solution to the optimization problem is developed in the sequence of theorems that follow.

Lemma 3.2: For a given linear decoder $\mathbf{G}^T$, $\exists$ an optimum encoder map $\mathbf{\Gamma}$ that satisfies:

$$\text{(i)}\ \ \mathcal{N}(\mathbf{\Gamma}) \subset \mathcal{N}(\mathbf{G}) = \mathcal{R}(\mathbf{G}^T)^{\perp}, \tag{III.46}$$

and

$$\text{(ii)}\ \ \mathcal{R}(\mathbf{\Gamma}) \subset \mathcal{R}(\mathbf{G}) = \mathcal{N}(\mathbf{G}^T)^{\perp}. \tag{III.47}$$

Proof: (i) Based on (III.42) we can write

$$(\mathbf{G}\mathbf{G}^T + \psi \mathbf{I})\mathbf{\Gamma}\mathbf{x} = \mathbf{G}\mathbf{x} \; . \tag{III.48}$$

Let $\mathbf{x} \in \mathcal{N}(\mathbf{\Gamma})$. The left side of (III.48) is zero, from which it follows that $\mathbf{G}\mathbf{x} = 0$ and hence $\mathcal{N}(\mathbf{\Gamma}) \subset \mathcal{N}(\mathbf{G})$. The equality in (III.46) is a well known fact in linear algebra and hence is not proved here.

(ii) It is always possible to decompose the encoder map as $\mathbf{\Gamma} = \mathbf{\Gamma}^o + \mathbf{\Gamma}^n$, where $\mathcal{R}(\mathbf{\Gamma}^o) \subset \mathcal{N}(\mathbf{G}^T)^\perp$ and $\mathcal{R}(\mathbf{\Gamma}^n) \subset \mathcal{N}(\mathbf{G}^T)$, since $\mathbb{R}^K = \mathcal{N}(\mathbf{G}^T)^\perp \oplus \mathcal{N}(\mathbf{G}^T)$. We now prove that if $\mathbf{\Gamma}$ is replaced by $\mathbf{\Gamma}^o$, the MSE remains unchanged and the energy constraint continues to be satisfied. Substitute $\mathbf{\Gamma} = (\mathbf{\Gamma}^o + \mathbf{\Gamma}^n)$ in the expression for the MSE given by (III.39). Thus

$$\begin{aligned} D' \;=\; & \frac{1}{L} tr \Big(\mathbf{R_{YY}} - 2\mathbf{R_{YY}}(\mathbf{\Gamma}^o + \mathbf{\Gamma}^n)^T \mathbf{G} + \\ & \mathbf{G}^T (\mathbf{\Gamma}^o + \mathbf{\Gamma}^n) \mathbf{R_{YY}} (\mathbf{\Gamma}^o + \mathbf{\Gamma}^n)^T \mathbf{G} + \mathbf{G}^T \mathbf{R_{ZZ}} \mathbf{G} \Big) \, . \end{aligned} \tag{III.49}$$

Since $\mathcal{R}(\mathbf{\Gamma}^n) \subset \mathcal{N}(\mathbf{G}^T)$, it follows that $\mathbf{G}^T\mathbf{\Gamma}^n = 0$ and hence $\mathbf{\Gamma}^{n^T}\mathbf{G} = 0$ which implies that D' in (III.39) is unchanged when $\mathbf{\Gamma}$ is replaced by $\mathbf{\Gamma}^o$.

The energy constraint states that $tr(\mathbf{\Gamma}\mathbf{R_{YY}}\mathbf{\Gamma}^T) \leq L\mathcal{E}_a$. But $tr(\mathbf{\Gamma}\mathbf{R_{YY}}\mathbf{\Gamma}^T) = tr(\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{R_{YY}}) = tr(\mathbf{\Gamma}^{o^T}\mathbf{\Gamma}^o\mathbf{R_{YY}}) + tr(\mathbf{\Gamma}^{n^T}\mathbf{\Gamma}^n\mathbf{R_{YY}})$, the terms containing $\mathbf{\Gamma}^{o^T}\mathbf{\Gamma}^n$ and $\mathbf{\Gamma}^{n^T}\mathbf{\Gamma}^o$ being zero because $\mathcal{R}(\mathbf{\Gamma}^o) \subset \mathcal{R}(\mathbf{\Gamma}^n)^\perp = \mathcal{N}(\mathbf{\Gamma}^{n^T})$ and similarly $\mathcal{R}(\mathbf{\Gamma}^n) \subset \mathcal{N}(\mathbf{\Gamma}^{o^T})$. Finally, $tr(\mathbf{\Gamma}^{n^T}\mathbf{\Gamma}^n\mathbf{R_{YY}}) = tr(\mathbf{\Gamma}^n\mathbf{R_{YY}}\mathbf{\Gamma}^{n^T}) \geq 0$ since $\mathbf{R_{YY}}$ is positive semi-definite, hence $L\mathcal{E}_a \geq tr(\mathbf{\Gamma}\mathbf{R_{YY}}\mathbf{\Gamma}^T) \geq tr(\mathbf{\Gamma}^o\mathbf{R_{YY}}\mathbf{\Gamma}^{o^T})$. We therefore do not lose optimality if we replace $\mathbf{\Gamma}$ by $\mathbf{\Gamma}^o$, which, by construction, satisfies (ii).

An immediate consequence of Lemma 3.2 is that, $\mathcal{R}(\mathbf{\Gamma}) = \mathcal{R}(\mathbf{G})$ and $\mathcal{N}(\mathbf{\Gamma}) = \mathcal{N}(\mathbf{G})$. This is so because,

$$\dim \mathcal{R}(\mathbf{\Gamma}) + \dim \mathcal{N}(\mathbf{\Gamma}) = L = \dim \mathcal{R}(\mathbf{G}) + \dim \mathcal{N}(\mathbf{G}), \tag{III.50}$$

which implies that

$$\dim \mathcal{R}(\mathbf{\Gamma}) - \dim \mathcal{R}(\mathbf{G}) = \dim \mathcal{N}(\mathbf{G}) - \dim \mathcal{N}(\mathbf{\Gamma}), \tag{III.51}$$

which can be true only if $\mathcal{R}(\Gamma) = \mathcal{R}(\mathbf{G})$ and $\mathcal{N}(\Gamma) = \mathcal{N}(\mathbf{G})$.

As we mentioned earlier, a natural set of basis vectors for $\mathrm{I\!R}^K$ and $\mathrm{I\!R}^L$ are the so-called singular bases associated with the decoder map $\mathbf{G}^T$, since both $\mathbf{G}$ and $\mathbf{G}\mathbf{G}^T$ have a convenient representation with respect to these bases. We now construct the singular bases with respect to $\mathbf{G}^T$. Details and proofs can be found in basic linear algebra texts such as [22].

$\mathbf{G}\mathbf{G}^T$ is a positive semi-definite operator, hence it has an orthonormal set of eigenvectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$ that span $\mathrm{I\!R}^K$. Let $\{g_i^2,\ i = 1, 2, \ldots, K\}$ be the associated set of eigenvalues and assume, without loss of generality, that $g_1^2 \geq g_2^2 \geq \ldots \geq g_t^2 > 0$ and that $g_{t+1}^2 = \ldots = g_K^2 = 0$, for some t, $0 \leq t \leq K$. Then $\{\mathbf{e}_1, \ldots, \mathbf{e}_t\}$ lies in and spans $\mathcal{N}(\mathbf{G}\mathbf{G}^T)^{\perp}$ and hence also $\mathcal{N}(\mathbf{G}^T)^{\perp}$, whereas, $\{\mathbf{e}_{t+1}, \ldots, \mathbf{e}_K\}$ spans $\mathcal{N}(\mathbf{G}^T)$. Let $\mathbf{v}_i = \mathbf{G}^T\mathbf{e}_i/g_i$ $, 1 \leq i \leq t$. Then $\{\mathbf{v}_i, 1 \leq i \leq t\}$ forms an orthonormal basis for $\mathcal{R}(\mathbf{G}^T) \subset \mathrm{I\!R}^L$. To complete the construction of an orthonormal basis for $\mathrm{I\!R}^L$ choose $\{\mathbf{v}_{t+1}, \ldots, \mathbf{v}_L\}$ as an arbitrary basis that spans $\mathcal{R}(\mathbf{G}^T)^{\perp}$. Here, $\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_L\}$ are called the singular bases of $\mathbf{G}^T$. With respect to the singular bases, the map $\mathbf{G}^T$ has the representation

$$\mathbf{G}^T = \left[\begin{array}{ccc|c} g_1 & & & \\ & \ddots & & 0 \\ & & g_t & \\ \hline & 0 & & 0 \end{array}\right], \qquad \text{(III.52)}$$

and $\mathbf{G}\mathbf{G}^T$ has the representation

$$\mathbf{G}\mathbf{G}^T = \begin{bmatrix} g_1^2 & & & & & \\ & \ddots & & & & \\ & & g_t^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}. \qquad \text{(III.53)}$$

Theorem 3.5: For a fixed decoder map $\mathbf{G}^T$, an optimum encoder map Γ which satisfies Lemma 3.2 has a zero off-diagonal (ZOD) representation with respect to the singular bases of $\mathbf{G}^T$ i.e., $\mathbf{\Gamma}$ is of the form,

$$\mathbf{\Gamma} = \left[\begin{array}{ccc|c} \gamma_1 & & & \\ & \ddots & & 0 \\ & & \gamma_t & \\ \hline & 0 & & 0 \end{array}\right]. \tag{III.54}$$

Proof: From (III.42) it follows that

$$\mathbf{\Gamma}^T[\mathbf{G}\mathbf{G}^T + \psi\mathbf{I}]\mathbf{e}_i = \mathbf{G}^T\mathbf{e}_i. \tag{III.55}$$

Now use the fact that for $1 \le i \le t$, $\mathbf{G}\mathbf{G}^T\mathbf{e}_i = g_i^2\mathbf{e}_i$ and $\mathbf{G}^T\mathbf{e}_i = g_i\mathbf{v}_i$ to get $\mathbf{\Gamma}^T\mathbf{e}_i = g_i/(g_i^2 + \psi)\mathbf{v}_i$, $1 \le i \le t$. For $t < i \le K$, $\mathbf{\Gamma}^T\mathbf{e}_i = 0$ since $\mathbf{e}_i \in \mathcal{N}(\mathbf{G}^T)$ which equals $\mathcal{N}(\mathbf{\Gamma}^T)$ as a direct consequence of Lemma 3.2. Hence if we define $\gamma_i = g_i/(g_i^2 + \psi)$, $1 \le i \le t$, it follows that $\mathbf{\Gamma}$ has a representation of the form (III.54).

So far we have succeeded in deriving a representation for the optimum $\mathbf{\Gamma}$ for a given $\mathbf{G}^T$ in terms of the singular bases of $\mathbf{G}^T$. However, the optimum decoder map $\mathbf{G}^T$ is itself unknown, hence we do not as yet have a representation for the singular bases of $\mathbf{G}^T$, in terms of absolute bases for $\mathbb{R}^K$ and $\mathbb{R}^L$. We now derive a representation for the basis vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_L\}$ for $\mathbb{R}^L$ with respect to an absolute basis i.e., the eigenvectors of $\mathbf{R}_{\mathbf{YY}}$. We will at this point assume that $\mathbf{R}_{\mathbf{ZZ}} = (N_0/2)\mathbf{I}$.

Theorem 3.6: $\exists$ an optimum decoder $\mathbf{G}^T$ for which the basis vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_L\}$ are the eigenvectors of $\mathbf{R}_{\mathbf{YY}}$, the covariance matrix of $\mathbf{Y}$.

Proof: Using the fact that $\mathbf{R}_{\mathbf{ZZ}} = (N_0/2)\mathbf{I}$, rewrite (III.45) as,

$$\mathbf{G}^T(\mathbf{\Gamma}\mathbf{R}_{\mathbf{YY}}\mathbf{\Gamma}^T + \frac{N_0}{2}\mathbf{I}) = \mathbf{R}_{\mathbf{YS}}. \tag{III.56}$$

Postmultiply (III.42) by $\mathbf{R}_{\mathbf{YY}}\mathbf{\Gamma}^T$, premultiply (III.56) by $\mathbf{G}$ and use the fact that $\mathbf{R}_{\mathbf{YS}} = \mathbf{R}_{\mathbf{YY}}\mathbf{\Gamma}^T$ to get

$$\psi\mathbf{\Gamma}\mathbf{R}_{\mathbf{YY}}\mathbf{\Gamma}^T = \frac{N_0}{2}\mathbf{G}\mathbf{G}^T . \tag{III.57}$$

Substitute (III.57) in (III.56) and postmultiply by $\mathbf{e}_i$ to get (assuming $\psi \neq 0$),

$$\mathbf{G}^T \left(\frac{N_0}{2\psi} \mathbf{G}\mathbf{G}^T + \frac{N_0}{2} \mathbf{I} \right) \mathbf{e}_i = \mathbf{R}_{\mathbf{YS}} \mathbf{e}_i = \mathbf{R}_{\mathbf{YY}} \mathbf{\Gamma}^T \mathbf{e}_i \ . \qquad \text{(III.58)}$$

Now use the facts that $\mathbf{\Gamma}^T \mathbf{e}_i = \gamma_i \mathbf{v}_i$, $\mathbf{G}\mathbf{G}^T \mathbf{e}_i = g_i^2 \mathbf{e}_i$ and $\mathbf{G}^T \mathbf{e}_i = g_i \mathbf{v}_i$, $1 \leq i \leq t$, to get

$$\mathbf{R}_{\mathbf{YY}} \mathbf{v}_i = \frac{N_0}{2\gamma_i} \left(\frac{g_i^2}{\psi} + 1 \right) g_i \mathbf{v}_i, \quad 1 \leq i \leq t \ . \qquad \text{(III.59)}$$

Hence, $\{\mathbf{v}_i,\ 1 \leq i \leq t\}$ are the eigenvectors of $\mathbf{R}_{\mathbf{YY}}$. For $t < i \leq L$, $\mathbf{v}_i$ was chosen arbitrarily so that $\{\mathbf{v}_i,\ i = 1, 2, \ldots,\ L\}$ spanned $\mathbb{R}^L$. Hence it suffices to choose $\mathbf{v}_i$, $t < i \leq L$ as the remaining eigenvectors of $\mathbf{R}_{\mathbf{YY}}$, since $\mathbf{R}_{\mathbf{YY}}$ is positive semi-definite and thus has an orthogonal set of eigenvectors that span $\mathbb{R}^L$.

To recapitulate, we have proved so far that if we use the eigenvectors of $\mathbf{R}_{\mathbf{YY}}$ to represent $\mathbb{R}^L$, $\exists$ a basis for $\mathbb{R}^K$ such that the optimum $\mathbf{G}^T$ and $\mathbf{\Gamma}$ have ZOD representations. Note that the covariance matrix of the noise process is unaffected by the choice of a basis for $\mathbb{R}^K$ because of the white noise assumption. Even though we do know that it suffices to choose the set of vectors $\{\mathbf{v}_1, \ldots,\ \mathbf{v}_L\}$ as the eigenvectors of $\mathbf{R}_{\mathbf{YY}}$, we do not know how the correspondence between this set and the set of eigenvectors is to be made. This is equivalent to determining how the eigenvalues of $\mathbf{R}_{\mathbf{YY}}$ should be indexed. For the moment we will assume that the eigenvalues of $\mathbf{R}_{\mathbf{YY}}$ have been indexed arbitrarily. It is not possible to deduce the best indexing of the eigenvalues of $\mathbf{R}_{\mathbf{YY}}$ based solely upon the necessary conditions of optimality. The reason is that there exist locally optimal solutions for *every* possible indexing of the eigenvalues. In order to determine the best indexing it is therefore necessary to evaluate the objective function at the locally optimum solutions and then choose the indexing that results in the *globally* optimum solution.

Since we have chosen the eigenvectors of $\mathbf{R}_{\mathbf{YY}}$ as the basis for $\mathbb{R}^L$, the matrix representation of $\mathbf{R_{YY}}$ in this basis is

$$\mathbf{R_{YY}} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_L \end{bmatrix}. \tag{III.60}$$

Further, we use the fact that $K \leq L$ and that $\mathbf{G}^T$ and $\mathbf{\Gamma}$ have ZOD representations in the chosen bases for $\mathbb{R}^K$ and $\mathbb{R}^L$ to write

$$\mathbf{G}^T = \left[\begin{array}{ccc} g_1 & & \\ & \ddots & \\ & & g_K \\ \hline & 0 & \end{array}\right], \tag{III.61}$$

and

$$\mathbf{\Gamma} = \left[\begin{array}{ccc|c} \gamma_1 & & & \\ & \ddots & & 0 \\ & & \gamma_K & \end{array}\right]. \tag{III.62}$$

Now use the fact that $\mathbf{R_{YS}} = \mathbf{R_{YY}}\mathbf{\Gamma}^T$ and substitute (III.60)–(III.62) in (III.45) (the optimum decoder condition) to obtain,

$$g_i = \frac{\lambda_i \gamma_i}{\lambda_i \gamma_i^2 + N_0/2}, \quad i = 1, 2, \ldots, K. \tag{III.63}$$

Use (III.63) in the expression for the distortion (III.39) to get

$$LD' = \sum_{i=1}^{L} \lambda_i - \sum_{i=1}^{K} \frac{\lambda_i^2 \gamma_i^2}{\lambda_i \gamma_i^2 + N_0/2}. \tag{III.64}$$

The energy constraint states that

$$\sum_{i=1}^{K} \lambda_i \gamma_i^2 \leq L\mathcal{E}_a, \tag{III.65}$$

and we wish to minimize (III.64) subject to (III.65), by suitably selecting γ_i, $1 \leq i \leq K$. Upon defining $\mathcal{E}_i = \lambda_i \gamma_i^2$ (the average energy in dimension i), (III.64) becomes

$$\begin{aligned} LD' &= \sum_{i=1}^{L} \lambda_i - \sum_{i=1}^{K} \frac{\mathcal{E}_i \lambda_i}{\mathcal{E}_i + N_0/2} \\ &= \sum_{i=1}^{K} \frac{\lambda_i N_0/2}{\mathcal{E}_i + N_0/2} + \sum_{i=K+1}^{L} \lambda_i, \end{aligned} \tag{III.66}$$

and (III.65) becomes

$$\sum_{i=1}^{K} \mathcal{E}_i \leq L\mathcal{E}_a \ . \qquad \text{(III.67)}$$

The problem has thus been reduced to an *energy allocation* problem. The optimum solution can be arrived at in a very intuitive way by using an incremental energy allocation technique quite similar to the incremental bit allocation techniques used in block transform coding systems. The derivatives, $\partial LD'/\partial\mathcal{E}_i$, given by,

$$\frac{\partial LD'}{\partial \mathcal{E}_i} = \frac{-\lambda_i N_0/2}{(\mathcal{E}_i + N_0/2)^2} \ , \ i = 1, 2, \ldots, \ K, \qquad \text{(III.68)}$$

are sketched in Fig. III.5 and are convex $\cap$, nonpositive, increasing functions of $\mathcal{E}_i$. If we start off with a zero energy assignment to every signal (or channel) dimension

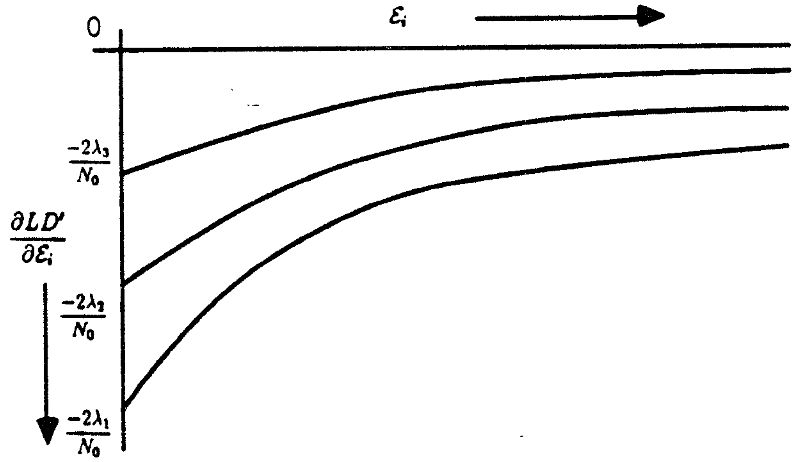

Figure III.5: $\partial LD'/\partial\mathcal{E}_i$ vs. $\mathcal{E}_i$.

and assign an infinitesimal amount of energy along the coordinate i for which the derivative $\frac{\partial LD'}{\partial \mathcal{E}_i}$ has the largest magnitude until the total allocated energy equals $L\mathcal{E}_a$, we should arrive at an optimum energy allocation. At this point the following conditions would be satisfied by $\mathcal{E}_i$, $1 \leq i \leq K$ for some $\psi \geq 0$:

$$\frac{\partial LD'}{\partial \mathcal{E}_i} = -\psi, \quad i \ : \ \mathcal{E}_i > 0, \qquad \text{(III.69)}$$

$$\frac{\partial LD'}{\partial \mathcal{E}_i} \geq -\psi, \quad i \, : \, \mathcal{E}_i = 0, \tag{III.70}$$

and

$$\frac{1}{L}\sum_{i=1}^{K} \mathcal{E}_i = \mathcal{E}_a. \tag{III.71}$$

Theorem 3.7: An energy allocation $\mathcal{E}_i$, $i = 1, 2, \ldots,$ K, that satisfies (III.69)-(III.71) minimizes (III.66) subject to (III.67).

Proof: The proof follows from the fact that $\mathcal{E}_i\lambda_i/(\mathcal{E}_i + N_0/2)$ is a convex $\cap$ function of $\mathcal{E}_i$. Let D' be the average distortion corresponding to the energy allocation that satisfies (III.69)-(III.71), and let $\tilde{D}'$ be the average distortion corresponding to some other energy allocation $\tilde{\mathcal{E}}_i$, $i = 1, 2, \ldots,$ K, such that $\sum_{i=1}^{K} \tilde{\mathcal{E}}_i \leq L\mathcal{E}_a$. Then

$$L(\tilde{D}' - D') \geq \sum_{i=1}^{K} \frac{\partial LD'}{\partial \mathcal{E}_i}(\tilde{\mathcal{E}}_i - \mathcal{E}_i) \geq -\psi \sum_{i=1}^{K}(\tilde{\mathcal{E}}_i - \mathcal{E}_i) \geq 0, \tag{III.72}$$

thus proving our claim.

The solution that satisfies (III.69)-(III.71) is an optimal solution for a *given* indexing of the eigenvalues and is only a locally optimal solution for the problem we set out to solve. We now prove that the *globally* optimum solution is obtained by computing the optimum energy allocation using the K largest eigenvalues of $\mathbf{R_{YY}}$.

Theorem 3.8: Let the eigenvalues be indexed such that the first K eigenvalues are the K largest eigenvalues of $R_{\mathbf{YY}}$. The globally optimum distortion is obtained by determining the optimum energy allocation for this indexing of the eigenvalues.

Proof: It suffices to prove that if the eigenvalues are arranged in such a way that $\lambda_i < \lambda_j$ for $i \leq K$ and $j > K$, the optimum distortion will decrease by exchanging λ_i and λ_j. Without loss of generality and to avoid generating unnecessary notation assume that $\lambda_1 < \lambda_{K+1}$. Let D' be the corresponding optimum distortion and $\mathcal{E}_i$, $i = 1, 2, \ldots,$ K, the corresponding energy allocation. Define the reordering, $j(i)$, by $j(i) = i$, $i \neq 1, K+1$; $j(1) = K+1$ and $j(K+1) = 1$. Consider the reordered eigenvalues, $\lambda_{j(i)}$ and let D'' be the corresponding distortion. From

(III.66) it follows that

$$\begin{aligned} L(D'-D'') &= \sum_{i=1}^{K} \frac{\lambda_i N_0/2}{\mathcal{E}_i + N_0/2} + \sum_{i=K+1}^{L} \lambda_i \\ &\quad - \sum_{i=1}^{K} \frac{\lambda_{j(i)} N_0/2}{\mathcal{E}_i + N_0/2} - \sum_{i=K+1}^{L} \lambda_{j(i)} \\ &= \frac{\lambda_1 N_0/2}{\mathcal{E}_1 + N_0/2} + \lambda_{K+1} - \frac{\lambda_{K+1} N_0/2}{\mathcal{E}_1 + N_0/2} - \lambda_1 \\ &= \frac{(\lambda_{K+1} - \lambda_1)\mathcal{E}_1}{\mathcal{E}_1 + N_0/2} > 0, \end{aligned} \tag{III.73}$$

thereby proving our claim.

An explicit solution for the optimization problem can now be developed as follows. The first step is to arrange the eigenvalues of $\mathbf{R_{YY}}$ in descending order, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_L \geq 0$. Let t be the number of dimensions to which strictly positive energies have been allocated. From (III.69) and (III.70) it follows that if $\mathcal{E}_j > 0$ then $\mathcal{E}_i > 0$ for $i < j$, which in turn implies that for some integer $t > 0$, $\mathcal{E}_i > 0,\ 1 \leq i \leq t$. Further, from (III.70) it follows that if $\mathcal{E}_{t+1} = 0$, then $-\psi \leq \partial LD'/\partial\mathcal{E}_{t+1}\ |_{\mathcal{E}_{t+1}=0}$ and since $\mathcal{E}_a$ increases as $-\psi$ increases, it follows that the value of $\mathcal{E}_a$ that corresponds to $-\psi = \partial LD'/\partial\mathcal{E}_{t+1}\ |_{\mathcal{E}_{t+1}=0}$ is the largest value of the energy constraint for which $\mathcal{E}_{t+1}$ is zero. Define this value of the energy constraint to be the t-th energy breakpoint denoted by $\mathcal{E}^t$.

Since $\partial LD'/\partial\mathcal{E}_{t+1}\ |_{\mathcal{E}_{t+1}=0} = -\lambda_{t+1}/(N_0/2)$, we can use (III.68) to determine $\mathcal{E}_j^t$, $1 \leq j \leq t$, where $\mathcal{E}_j^t$ is the energy $\mathcal{E}_j$ corresponding to the slope $-\psi = -\lambda_{t+1}/(N_0/2)$. Hence

$$\mathcal{E}_j^t = \frac{N_0}{2}\left(\sqrt{\frac{\lambda_j}{\lambda_{t+1}}} - 1\right), \quad 1 \leq j \leq t\,. \tag{III.74}$$

By summing (III.74) over j, $1 \leq j \leq t$ we then obtain

$$\mathcal{E}^t = \frac{N_0/2}{\sqrt{\lambda_{t+1}}} \sum_{j=1}^{t} \left(\sqrt{\lambda_j} - \sqrt{\lambda_{t+1}}\right). \tag{III.75}$$

For $\mathcal{E}^{t-1} < L\mathcal{E}_a \leq \mathcal{E}^t$, exactly t channel dimensions are allocated positive energies and we have

$$\frac{\partial LD'}{\partial \mathcal{E}_i} = -\psi, \quad 1 \leq i \leq t\,. \tag{III.76}$$

Use (III.76) in (III.68), multiply the result by $-(\mathcal{E}_i + N_0/2)^2$ for each i, compute the square-root of both sides and sum over i, $1 \le i \le t$, using $\sum_{i=1}^{t} \mathcal{E}_i = L\mathcal{E}_a$, to get

$$\psi = \frac{N_0}{2} \left(\frac{\sum_{i=1}^{t} \sqrt{\lambda_i}}{L\mathcal{E}_a + tN_0/2} \right)^2 , \tag{III.77}$$

from which it easily follows by using (III.68) that

$$\gamma_i = \sqrt{\frac{N_0}{2\lambda_i}} \left(\frac{\sqrt{\lambda_i}\,(1 + 2L\mathcal{E}_a/tN_0)}{\sum_{j=1}^{t} \sqrt{\lambda_j}/t} - 1 \right)^{1/2} , \; 1 \le i \le t. \tag{III.78}$$

The optimum decoder can then be determined using (III.63). Finally for $\mathcal{E}^{t-1} \le L\mathcal{E}_a < \mathcal{E}^t$, the minimum MSE is obtained by using (III.78) in (III.66) to get

$$D' = \frac{N_0}{2L} \frac{\left(\sum_{i=1}^{t} \sqrt{\lambda_i}\right)^2}{(tN_0/2 + L\mathcal{E}_a)} + \frac{1}{L} \sum_{i=t+1}^{L} \lambda_i. \tag{III.79}$$

Note that the MSE for the overall system (VQ, encoder, decoder) is given by (III.40).

Unfortunately it is not possible to jointly determine the optimum VQ partition along with the optimum signal set and the optimum decoder and we must therefore use an iterative algorithm for obtaining a *locally optimum* solution. The algorithm is presented in what follows.

III.5 The Design Algorithm

The basic idea is to iteratively solve the necessary conditions for optimality, successively decreasing the MSE at each step until the algorithm reaches a stationary point. The algorithm is stated below.

1. *Initialization:* Set $D^{(0)} = \Delta$ (a large constant). Select an initial encoder partition $\mathcal{A}^{(0)}$. Set the termination threshold $\delta > 0$. Set iteration index $j = 0$.

2. Compute the partition set probabilities $P_i^{(j)}$, $i = 1, 2, \ldots, N$, and the centroids $\mathbf{y}_i^{(j)}$, $i = 1, 2, \ldots, N$, and the covariance matrix $\mathbf{R}_{\mathbf{YY}}^{(j)}$.

3. Compute the optimum linear encoder $\mathbf{\Gamma}^{(j)}$ and the optimum linear decoder $\mathbf{G}^{(j)^T}$ using (III.78) and (III.63), respectively.

4. Set $j \leftarrow j + 1$. Compute the average overall distortion $D^{(j)}$ using (III.79) and (III.40). If $(D^{(j-1)} - D^{(j)})/D^{(j-1)} < \delta$, then stop, else continue.

5. Compute the optimum partition $\mathcal{A}^{(j)}$ using (III.15). Return to Step 2.

The algorithm converges since it produces a sequence of nonincreasing distortions, which are bounded below by zero. Due to computational problems in integrating functions over irregular regions of multidimensional Euclidean space, it is necessary to use a training sequence approach in order to compute the optimum partition using (III.15) and its associated probabilities and centroids. This is a commonly used technique in vector quantizer design [23] and we will not elaborate on the details here. We will refer to the system designed using the above algorithm as the BPLE system (Block encoder-, Projection demodulation-, Linear estimator decoder-based system).

IV A Bound on the System Performance

Our main goal in this section is to study the performance of the BPLE system as a function of the encoding rate (or equivalently, the signal set cardinality) while maintaining a bound on the average transmitted energy, for fixed K and L. Further, we wish to make comparisons against a simple analog modulation scheme, which we call BPAM (Block Pulse Amplitude Modulation). This system is illustrated in Fig. IV.1 and is described below. A source vector $\mathbf{X}$ of dimension L is mapped by the linear encoder $\mathbf{\Gamma} : \mathbb{R}^L \rightarrow \mathbb{R}^K$ to a K-dimensional signal $\mathbf{S}$. The received vector $\mathbf{U} = \mathbf{S} + \mathbf{Z}$ is mapped by the decoder $\mathbf{G}^T$, to form an estimate $\widehat{\mathbf{X}}$ of $\mathbf{X}$. An optimum BPAM system is a pair $(\mathbf{\Gamma}, \mathbf{G}^T)$ that minimizes the MSE, $E(\|\mathbf{X} - \widehat{\mathbf{X}}\|^2)/L$ subject to the energy constraint $E(\mathbf{S}^T\mathbf{S}) \leq L\mathcal{E}_a$. The MSE of the optimum BPAM system will be denoted by $D_B(\mathcal{E}_a, L, K)$. We denote the performance of the BPLE system with energy constraint $\mathcal{E}_a$, source dimension L, channel dimension K and signal set cardinality N, by $D(\mathcal{E}_a, N, L, K)$. It turns out that the MSE obtained by the BPAM system forms a lower bound on the MSE of any BPLE system. This result is proved in the next theorem.

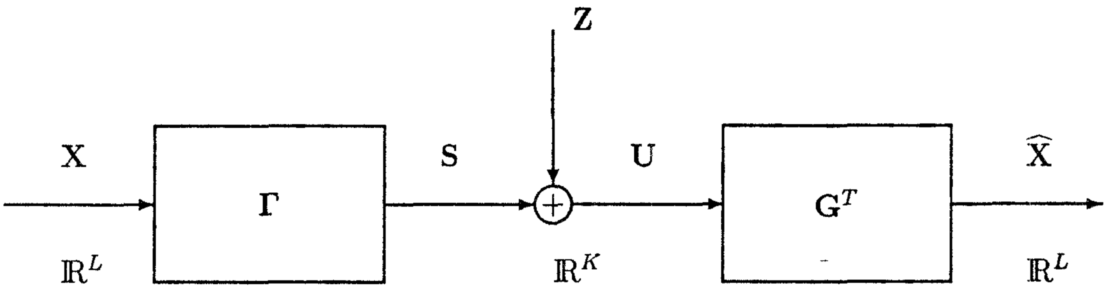

Figure IV.1: BPAM System

Theorem 4.1: For a given and fixed source dimension L, channel dimension K and energy constraint $\mathcal{E}_a$, the following results are true:

(i) $D(\mathcal{E}_a, N, L, K) \geq D_B(\mathcal{E}_a, L, K)$, $\forall N$,

(ii) $D(\mathcal{E}_a, N, L, K) \geq D(\mathcal{E}_a, M, L, K)$, $N \leq M$,

and

(iii) $\lim_{N\to\infty} D(\mathcal{E}_a, N, L, K) = D_B(\mathcal{E}_a, L, K)$.

Proof: (i) The main idea behind the proof is the fact that the signal set cardinality imposes a constraint on the signal set. Hence, if we remove this constraint while assuming that the decoder is linear and determine the optimum encoder, the performance so obtained will form a lower bound to $D(\mathcal{E}_a, N, L, K)$. As we will see, the system obtained by removing the cardinality constraint is a BPAM system, and, assuming for the moment that this result is true, it follows that,

$$D_B(\mathcal{E}_a, L, K) \leq D(\mathcal{E}_a, N, L, K), \ \forall N. \tag{IV.1}$$

It now remains to prove that for a given linear decoder, $\mathbf{G}^T$, the optimum signal set obtained after removing the cardinality constraint is linearly related to the source output. This is equivalent to proving that the optimum map, $\gamma(\cdot)$, from the source output space to the channel space, is linear. The MSE for a system with a decoder $\mathbf{G}^T$ and encoder $\gamma(\cdot)$ is given by

$$\begin{aligned} D &= \frac{1}{L} tr\left(\int_{\mathbb{R}^L} E\left((\mathbf{x}-\widehat{\mathbf{X}})(\mathbf{x}-\widehat{\mathbf{X}})^T | \mathbf{X}=\mathbf{x}\right) dP_{\mathbf{X}}\right) \\ &= \frac{1}{L} tr\left(\int_{\mathbb{R}^L} \left[\mathbf{x}\mathbf{x}^T - 2\mathbf{x}\gamma^T(\mathbf{x})\mathbf{G} + \mathbf{G}^T\gamma(\mathbf{x})\gamma^T(\mathbf{x})\mathbf{G} + \tfrac{N_0}{2}\mathbf{G}^T\mathbf{G}\right] dP_{\mathbf{X}}\right). \end{aligned} \tag{IV.2}$$

We wish to determine the encoder map $\gamma(\cdot)$, that minimizes (IV.2) subject to

$$E(\gamma^T(\mathbf{X})\gamma(\mathbf{X})) \leq L\mathcal{E}_a. \tag{IV.3}$$

The Lagrangian for the optimization problem, $L(\gamma, \psi)$, is given by

$$L(\gamma, \psi) = \frac{1}{L} \int_{\mathbb{R}^L} \left[tr\left(\mathbf{x}\mathbf{x}^T - 2\mathbf{x}\gamma^T(\mathbf{x})\mathbf{G} + \mathbf{G}^T\gamma(\mathbf{x})\gamma^T(\mathbf{x})\mathbf{G} + \tfrac{N_0}{2}\mathbf{G}^T\mathbf{G}\right)\right.$$

$$+\psi\left(\gamma^T(\mathbf{x})\gamma(\mathbf{x})-\mathcal{E}_a\right)\Big]\,dP_{\mathbf{X}}, \tag{IV.4}$$

and it suffices to minimize

$$F(\gamma(\mathbf{x}),\psi)\triangleq tr\left(\mathbf{x}\mathbf{x}^T-2\mathbf{x}\gamma^T(\mathbf{x})\mathbf{G}+\mathbf{G}^T\gamma(\mathbf{x})\gamma^T(\mathbf{x})\mathbf{G}\right)+\psi\left(\gamma^T(\mathbf{x})\gamma(\mathbf{x})-\mathcal{E}_a\right), \tag{IV.5}$$

for all $\mathbf{x}\in\mathbb{R}^L$. Differentiate (IV.5) with respect to $\gamma(\mathbf{x})$ and set the derivative to zero in order to arrive at the result,

$$\left(\mathbf{G}\mathbf{G}^T+\psi\mathbf{I}\right)\gamma(\mathbf{x})=\mathbf{G}\mathbf{x},\quad\forall\mathbf{x}, \tag{IV.6}$$

thus proving our claim that $\gamma(\mathbf{x})$ is linear in $\mathbf{x}$.

(ii) The proof of this statement is straightforward. The class of cardinality M linear estimator decoder-based systems includes the class of cardinality N linear estimator decoder-based systems, for $N\le M$. Since the minimum distortion over a given class of systems cannot exceed the minimum distortion over a subset of this class, it follows that $D(\mathcal{E}_a,N,L,K)\ge D(\mathcal{E}_a,M,L,K),\ N\le M$.

(iii) Let $(\mathbf{\Gamma},\mathbf{G}^T)$ constitute an optimum BPAM system. The MSE for this system, $D_B(\mathcal{E}_a,L,K)$, is given by,

$$D_B(\mathcal{E}_a,L,K)=\frac{1}{L}\int_{\mathbb{R}^L}tr\left(\mathbf{x}\mathbf{x}^T-2\mathbf{x}\mathbf{x}^T\mathbf{\Gamma}^T\mathbf{G}+\mathbf{G}^T\mathbf{\Gamma}\mathbf{x}\mathbf{x}^T\mathbf{\Gamma}^T\mathbf{G}+\frac{N_0}{2}\mathbf{G}^T\mathbf{G}\right)dP_{\mathbf{X}}. \tag{IV.7}$$

Define the set E by

$$E=\{\mathbf{x}\ :\ \|\mathbf{x}\|\le r\}. \tag{IV.8}$$

Since $D_B(\mathcal{E}_a)<\infty$, it follows that for any $\epsilon>0,\exists\ r$ such that

$$\frac{1}{L}\int_{E^c}tr(\mathbf{x}\mathbf{x}^T-2\mathbf{x}\mathbf{x}^T\mathbf{\Gamma}^T\mathbf{G}+\mathbf{G}^T\mathbf{\Gamma}\mathbf{x}\mathbf{x}^T\mathbf{\Gamma}^T\mathbf{G}+\frac{N_0}{2}\mathbf{G}^T\mathbf{G})dP_{\mathbf{X}}\le\frac{\epsilon}{2}\,. \tag{IV.9}$$

Each of the four terms of the integrand in (IV.7) is a continuous function, E is a bounded set; hence the integral of the last three terms over E can be approximated by the integral of a step function on E. In other words, for N sufficiently large, $\exists$

a grid $\mathcal{A} = \{A_i,\ i = 1,2,\ldots,\ N\}$, where $A_i \subset E$, that partitions E such that the step functions $tr\left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T \mathbf{I}_{A_i}(\mathbf{x})\mathbf{\Gamma}^T\mathbf{G}\right)$ and $tr\left(\mathbf{G}^T\mathbf{\Gamma}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T \mathbf{I}_{A_i}(\mathbf{x})\mathbf{\Gamma}^T\mathbf{G}\right)$, $\mathbf{x}_i \in A_i$ satisfy,

$$\left|\int_E tr\left(\mathbf{x}\mathbf{x}^T - 2\mathbf{x}\mathbf{x}^T\mathbf{\Gamma}^T\mathbf{G} + \mathbf{G}^T\mathbf{\Gamma}\mathbf{x}\mathbf{x}^T\mathbf{\Gamma}^T\mathbf{G} + \frac{N_0}{2}\mathbf{G}^T\mathbf{G}\right) dP_{\mathbf{X}} \right.$$
$$- \int_E tr\left(\mathbf{x}\mathbf{x}^T - 2\left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T\mathbf{I}_{A_i}(\mathbf{x})\right)\mathbf{\Gamma}^T\mathbf{G} + \mathbf{G}^T\mathbf{\Gamma}\left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T\mathbf{I}_{A_i}(\mathbf{x})\right)\mathbf{\Gamma}^T\mathbf{G}\right.$$
$$\left.\left. + \tfrac{N_0}{2}\mathbf{G}^T\mathbf{G}\right) dP_{\mathbf{X}}\right| < \frac{L\epsilon}{2}. \tag{IV.10}$$

Now consider a system in which the encoder imposes a partition $\{E^c,\ A_i,\ i = 1,2,\ldots,\ N\}$ on $\mathbb{R}^L$ and in which if $\mathbf{x} \in E^c$, it is transmitted using signal $\mathbf{s}_{N+1} = 0$. If we set $\mathbf{x}_i = E(\mathbf{X}|\mathbf{X} \in A_i)$ [6], the average distortion for this system will be given by the second integral on the left hand side of (IV.10) [7]. Further, this system satisfies the energy constraint as we now prove. Since $\mathbf{s}_{N+1} = 0$ it suffices to prove that

$$tr\int_E \left(\mathbf{\Gamma}\mathbf{x}\mathbf{x}^T\mathbf{\Gamma}^T - \mathbf{\Gamma}\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T\mathbf{I}_{A_i}(\mathbf{x})\right)\mathbf{\Gamma}^T\right) dP_{\mathbf{X}} \geq 0\ , \tag{IV.11}$$

where $\mathbf{x}_i = E(\mathbf{X}|\mathbf{X} \in A_i)$, $i = 1,2,\ldots,\ N$. To do this, consider the inequality,

$$tr\left(\int_E \mathbf{\Gamma}\left(\sum_{i=1}^{N}(\mathbf{x}-\mathbf{x}_i)(\mathbf{x}-\mathbf{x}_i)^T\mathbf{I}_{A_i}(\mathbf{x})\right)\mathbf{\Gamma}^T dP_{\mathbf{X}}\right) \geq 0, \tag{IV.12}$$

which holds, because the integrand is always non-negative. Equation (IV.12) can be written as

$$tr\left(\int_E \mathbf{\Gamma}\sum_{i=1}^{N}\mathbf{x}(\mathbf{x}-\mathbf{x}_i)^T\mathbf{I}_{A_i}(\mathbf{x})\mathbf{\Gamma}^T dP_{\mathbf{X}} - \int_E \mathbf{\Gamma}\left(\sum_{i=1}^{N}\mathbf{x}_i(\mathbf{x}-\mathbf{x}_i)^T\mathbf{I}_{A_i}(\mathbf{x})\right)\mathbf{\Gamma}^T dP_{\mathbf{X}}\right) \geq 0. \tag{IV.13}$$

[6] We have used the definition of the Reimann integral over a bounded region here, rather than the definition of the Lebesgue integral. The reason is that in the definition of the Lebesgue integral the sets of the partition of $\mathbb{R}^L$ may not be convex in which case $E(\mathbf{X}|\mathbf{X} \in A_i)$ need not lie in A_i due to which the approximation in (IV.10) need not be true with $\mathbf{x}_i = E(\mathbf{X}|\mathbf{X} \in A_i)$.

[7] It is necessary to prove that $||E(\mathbf{X}|\mathbf{X} \in E^c)||$ is finite. But $E(\mathbf{X}) = \sum_{i=1}^{N} E(\mathbf{X}|\mathbf{X} \in A_i)Pr(A_i) + E(\mathbf{X}|\mathbf{X} \in E^c)Pr(E^c)$ and since $E(\mathbf{X}) < \infty$ and so is $E(\mathbf{X}|\mathbf{X} \in A_i)$, $i = 1,2,\ldots,\ N$, it follows that $||E(\mathbf{X}|\mathbf{X} \in E^c)|| < \infty$.

It is easily verified that the second term in (IV.13) is zero and that the first term is identical to (IV.11) which proves our claim. Hence we have now determined a (possibly non-optimal) system which satisfies the energy constraint and whose distortion, D, satisfies,

$$|D - D_B(\mathcal{E}_a, L, K)| \leq \epsilon\ . \tag{IV.14}$$

But by part (i) of this theorem we know that $D \geq D_B(\mathcal{E}_a, L, K)$, hence by using the fact that the distortion for the BPLE system, $D(\mathcal{E}_a, N+1, L, K)$, satisfies $D_B(\mathcal{E}_a, L, K) \leq D(\mathcal{E}_a, N+1, L, K) \leq D$, it follows that,

$$0 \leq D(\mathcal{E}_a, N+1, L, K) - D_B(\mathcal{E}_a, L, K) \leq \epsilon\ , \tag{IV.15}$$

which proves (iii).

The proof of part (iii) indicates that when the encoding rate is high enough, the precise selection of the encoder partition is relatively unimportant. It also demonstrates clearly that the BPLE systems that we design are finite rate approximations of the BPAM system.

The second result (ii), though straightforward, is somewhat surprising, for, by studying simple detection-based demodulation systems one might have expected otherwise. To be specific, consider a system in which source samples are encoded by a scalar quantizer which has a rate R_s bits/sample and transmitted using a 2^{R_s}-PAM modulation system. Let T be the modulation symbol duration and assume we have a transmitter with an average power constraint of P_a. Symbol error probabilities for this signal constellation are given in terms of energy per bit E_b, ([19], Fig. 4.2.24) which in this case is given by $E_b = P_aT/R_s$. For fixed P_a and T, as R_s increases, E_b decreases. Further since the symbol error probability *increases* with increasing number of signals in the signal constellation *and* with a decrease in CSNR [19], it follows that the average symbol error probability will have increased. Hence we can expect the contribution of the channel error to the overall mean squared-error to increase as the encoding rate increases. In fact, it is fairly simple to verify this fact by direct computation. Based on this one might have expected the MSE of our system to increase, or at least eventually increase, as the encoding rate is increased. For the BPLE system, since the signal set is optimally chosen, the rate of increase of the channel MSE can be made smaller than the rate of decrease of the source encoder MSE. This explains why the MSE does not increase as the encoding rate is increased.

V Performance Results

We present performance results for the BPLE system. Results are presented for a first-order Gauss-Markov source for a variety of correlation coefficients, channel-signal-to-noise ratios, encoding rates and bandwidths. Comparisons are made against the BPAM bound as described in Section IV, against a Linde-Buzo-Gray vector quantizer (LBG VQ)-based system in which a QAM signal set is used in the modulator and against the OPTA obtained by evaluating the distortion-rate function of the Gaussian source at the channel capacity. Tabulated performance results corresponding to the graphs presented in this section are provided in Appendix C.

V.1 Source and Channel Description

The source is assumed to be modeled by a zero-mean, unit-variance, stationary, first-order, Gauss-Markov random process with correlation coefficient ρ, described by

$$X_n = \rho X_{n-1} + W_n, \quad n \in \mathbb{Z}, \tag{V.1}$$

where W_n is an i.i.d. Gaussian process with variance $EW_n^2 = (1-\rho^2)$. For the above source the $L \times L$ covariance matrix, $\mathbf{R}_{\mathbf{XX}}^L$, is given by

$$\mathbf{R}_{\mathbf{XX}}^L = \begin{bmatrix} 1 & \rho & \cdots & \rho^{L-1} \\ \rho & 1 & \cdots & \rho^{L-2} \\ \vdots & & & \\ \rho^{L-1} & \rho^{L-2} & \cdots & 1 \end{bmatrix}, \tag{V.2}$$

where the entry in the (i,j) position is EX_iX_j. We will denote by $\{\mathbf{X}_n^L\}$ a Gaussian *independent* vector process of dimension L, with covariance matrix $\mathbf{R}_{\mathbf{XX}}^L$, and will simply refer to this source as the vector source. The dimension of the vector source will be obvious from the context.

The channel is modeled by a zero-mean, stationary, independent, vector Gaussian random process of dimension K with a covariance matrix $(N_0/2)\mathbf{I}$. Performance results are presented for various values of the CSNR, which is described by

$$\mathrm{CSNR} = 10\log_{10}(E\mathbf{S}^T\mathbf{S}/LN_0). \tag{V.3}$$

V.2 BPAM bound

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_L > 0$, be the eigenvalues of $\mathbf{R}^L_{\mathbf{XX}}$. Let $D_B(\mathcal{E}_B, L, K)$ denote the MSE per source sample of a BPAM system that maps source vectors of dimension L to channel vectors of dimension K and uses an average energy $\mathcal{E}_B$ per source sample. From (III.79) the MSE is given by

$$D_B(\mathcal{E}_B, L, K) = \frac{1}{L}\left(\frac{N_0}{2}\frac{\left(\sum\limits_{i=1}^{t}\sqrt{\lambda_i}\right)^2}{tN_0/2 + L\mathcal{E}_B} + \sum_{i=t+1}^{L}\lambda_i\right), \tag{V.4}$$

where, from (III.75), $t = \min(K, t')$ and t' is the smallest integer which satisfies

$$\frac{N_0/2}{\sqrt{\lambda_{t'+1}}}\left(\sum_{j=1}^{t'}\left(\sqrt{\lambda_j} - \sqrt{\lambda_{t'+1}}\right)\right) \geq L\mathcal{E}_B. \tag{V.5}$$

In order to make comparisons against bounds from information theory it is more convenient to express the MSE and average energy parametrically. From (III.66), (III.68), (III.69) and (III.70) it follows that the distortion D_B and the average transmitted energy $\mathcal{E}_B$ can be expressed parametrically[8] in terms of θ as follows:

$$D_B(\theta, L, K) = \frac{1}{L}\left(\sum_{i=1}^{K}\min\left[\theta, \lambda_i\right] + \sum_{i=K+1}^{L}\lambda_i\right) \tag{V.6}$$

and

$$\mathcal{E}_B(\theta, L, K) = \frac{N_0 K}{2L}\left(\frac{1}{K}\sum_{i=1}^{K}\max\left[\lambda_i/\theta, 1\right] - 1\right). \tag{V.7}$$

V.3 Optimum Performance Theoretically Attainable (OPTA)

The OPTA is determined by evaluating the distortion-rate function of the source at a rate equal to the channel capacity. The channel capacity of a Gaussian vector

[8] A derivation is provided in Appendix B

channel of dimension K and covariance matrix $(N_0/2)\mathbf{I}$, for an average energy per source sample $\mathcal{E}_{opt}$, is given by [18],

$$C = \frac{K}{2L}\log_2\left(1 + \frac{2L\mathcal{E}_{opt}}{KN_0}\right), \text{ bits/source sample.} \tag{V.8}$$

The distortion-rate functions for the Gauss-Markov source described by (V.1) and for the Gaussian vector source $\{\mathbf{X}_n^L\}$ having covariance matrix $\mathbf{R}_{\mathbf{XX}}^L$ are denoted by $D_{opt}(R)$ and $D_{opt}(R, L)$, respectively. As is well-known [24],

$$\lim_{L\to\infty} D_{opt}(R, L) = D_{opt}(R). \tag{V.9}$$

Both $D_{opt}(C, L)$ and $D_{opt}(C)$ serve as upper bounds on the SNR performance of the BPLE system and the BPAM system where it is assumed that both systems operate on source vectors of length L.

The rate-distortion function for the vector source can be described parametrically in terms of θ by [24],

$$D_{opt}(\theta, L) = \frac{1}{L}\sum_{i=1}^{L}\min\left[\theta, \lambda_i\right], \tag{V.10}$$

and

$$R_{opt}(\theta, L) = \frac{1}{L}\sum_{i=1}^{L}\max\left[0, \frac{1}{2}\log_2\frac{\lambda_i}{\theta}\right], \text{ bits/source sample.} \tag{V.11}$$

By equating the channel capacity given by (V.8) to $R_{opt}(\theta, L)$ as given by (V.11), the OPTA for the vector source is given parametrically in terms of θ by,

$$D_{opt}(\theta, L) = \frac{1}{L}\sum_{i=1}^{L}\min\left[\theta, \lambda_i\right], \tag{V.12}$$

and

$$\mathcal{E}_{opt}(\theta, L, K) = \frac{N_0K}{2L}\left(\left(\prod_{i=1}^{L}\max\left[\lambda_i/\theta, 1\right]\right)^{1/K} - 1\right). \tag{V.13}$$

Let $\Phi(\omega)$ denote the power spectral density of the first-order Gauss-Markov source. The rate-distortion function for this source is evaluated by applying a theorem on the asymptotic distribution of the eigenvalues of a Toeplitz form [25] (hereafter referred to as the Toeplitz distribution theorem, following [26]) to evaluate the limits of

(V.10) and (V.11). The rate-distortion function for the Gauss-Markov source can also be described parametrically in terms of θ by [24],

$$D_{opt}(\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \min\left[\theta, \Phi(\omega)\right] d\omega, \tag{V.14}$$

and

$$R_{opt}(\theta) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \max\left[0, \log_2 \frac{\Phi(\omega)}{\theta}\right] d\omega, \text{ bits/source sample.} \tag{V.15}$$

Analogous to the procedure used for the vector source, the OPTA for the first-order Gauss-Markov source is obtained by equating the channel capacity given by (V.8) to $R_{opt}(\theta)$ as given by (V.15) and can be expressed parametrically as follows:

$$D_{opt}(\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \min\left[\theta, \Phi(\omega)\right] d\omega, \tag{V.16}$$

and

$$\mathcal{E}_{opt}(\theta, B) = \frac{N_0 B}{2} \left(2^{\frac{1}{2\pi B} \int_{-\pi}^{\pi} \max\left[0, \log_2 \frac{\Phi(\omega)}{\theta}\right] d\omega} - 1 \right), \tag{V.17}$$

where B, the bandwidth expansion factor, is the number of channel dimensions per source dimension.

V.4 The Linde-Buzo-Gray Vector Quantizer-Based System

For purposes of comparison, we consider a simple communication system in which the source encoder consists of a VQ designed optimally for the given source (and a noiseless channel) using the Linde-Buzo-Gray algorithm [23]. It is assumed that the source encoding rate and the source vector dimension are the same as for the BPLE system designed in Section III. The output of the VQ is then transmitted across a Gaussian waveform channel. Standard signal constellations as listed below in Table V.1 are used in the transmitter. Here $(M\ PAM)^K$ denotes the K-fold cartesian product of the M-ary PAM signal constellation with itself. The receiver consists of a conventional maximum-likelihood detection-based demodulator followed by a source decoder, that maps the decoded signal back to the centroid of the corresponding cell. It turns out that a critical part of this communication system is the mapping from the VQ codewords to the signals in the modulation signal set. We have developed a heuristic algorithm for determining this mapping, details for which can be found in [27]. Results for the LBG VQ-based system have been computed based on mappings obtained via this algorithm. In cases where not all the signals in the signal set have the same energy, the average transmitted energy depends on the probability of usage of the signals. In these cases the levels of the PAM signal set have been scaled so as to satisfy the energy constraint. In all cases the performance results presented are based on the actual average transmitted energy. We refer to the LBG VQ-based system as the LBGDC system (D for Detection, C for Centroids, which are used as the reconstruction vectors by the decoder).

K	8	4	2	1
K/L	1.0	0.5	0.25	0.125
Signal Constellation	$(2\text{ PAM})^8$	$(4\text{ PAM})^4$	$(16\text{ PAM})^2$	(256 PAM)

Table V.1: Modulation Signal Constellations for 1 bit/sample LBGDC Systems.

V.5 Description of Graphs.

BPLE systems have been designed for various combinations of the source correlation ρ = 0.0, 0.5, 0.9 and 0.95, source vector dimension $L = 1, 2, 4$ and 8, channel signal-to-noise ratio, CSNR = −6.0 dB to 15.0 dB in steps of 3 dB and various source encoding rates, with the restriction that the number of signals in the signal constellation given by $N = 2^{LR_s}$ satisfies $N \leq 256$. All results have been computed based on a training sequence of length 48,000 source vectors. The restriction $N \leq$ 256 was imposed because it was felt that larger constellation sizes would result in a loss of accuracy due to the restricted size of the training sequence. We have considered bandwidth expansions of 1.0, 0.5, 0.25 and 0.125.

We will use the term "performance" for the output SNR[9] of the communication system. Performance results are presented in graphical form in Figs. V.1–V.11. A bandwidth expansion of 1.0 is assumed for Figs. V.1–V.4. In Figs. V.5–V.8 various values of bandwidth expansion are considered, but the source block size is held fixed at $L = 8$. In Figs. V.9 and V.10 comparisons are made against LBGDC systems, for $L = 8$, $R_s = 1.0$ and various bandwidth expansions. In Fig. V.11 the source, channel and overall MSE[10] are plotted as a function of the source coding rate R_s. Finally, in Fig. V.12 optimum two-dimensional 16-ary signal constellations are displayed for $L = 2$, $K/L = 1.0$ and various source correlation values.

The curves in the performance graphs have been labeled in the following manner: OPTA and OPTA(L) correspond to the OPTA for the Gauss-Markov source and the vector source, respectively. BPAM stands for the performance of the BPAM system, BPLE for the performance of the BPLE system and LBGDC for the performance of the LBGDC system.

V.6 Discussion of the Results

We now discuss the performance of the BPAM system relative to the OPTA and the performance of the BPLE system relative to the BPAM system, the LBGDC system and the OPTA.

V.6.A Comparisons of BPAM Performance to OPTA

We begin by discussing the performance of the BPAM system for a bandwidth expansion of unity (Figs. V.1–V.4). If we compare the parametric representations

[9] The output SNR is given by $10 \log_{10} D/\sigma^2$, where D is the per-sample distortion and σ^2 is the source variance.

[10] The terms, source and channel MSE are defined after (III.40).

for the BPAM system performance (V.6–V.7) with those of the OPTA for the vector source with the same dimension (V.12-V.13) under the assumption that $K = L$, we notice that for a given value of the parameter θ, the average distortion achieved is the same in both cases. However, the average transmitted energy in the case of BPAM is a function of the *arithmetic* mean of the terms λ_i/θ, $i = 1, 2, \ldots, L$, whereas the OPTA has exactly the same functional form in terms of the *geometric* mean of the above terms. Hence, with $L = K$, the BPAM performance and the OPTA will coincide for all values of the CSNR only when all the eigenvalues of the source are equal or when $L = K = 1$. For all other cases with $K = L$ the energy required for a given distortion with a BPAM system is strictly greater than the energy that an optimum system would require.

Hence, when the source correlation, $\rho = 0$ and $K/L = 1.0$, the BPAM bound for any value of $L \geq 1$ coincides with the OPTA for the vector source for all values of CSNR. We remark that in this case the OPTA for the vector source is identical to the OPTA for the Gauss-Markov source for $L \geq 1$. As the source correlation increases the bounding curves begin to separate. For example, when $\rho = 0.95$ the BPAM curve lies approximately 6 dB below the OPTA for the Gauss-Markov source and approximately 5 dB below the OPTA for the vector source.

We now consider more restricted bandwidths, i.e., $K/L < 1.0$. The parametric representations for the BPAM system performance and the OPTA for the vector source point to yet another case in which the two coincide, though only over a range of CSNR values. This case occurs when $K = 1$ and the CSNR is low enough so that $\lambda_i/\theta \leq 1$, for all $i > 1$ or, equivalently, for CSNR$\leq 10\log_{10}((\lambda_1/\lambda_2 - 1)/(2L))$. In this case again the BPAM system provides an optimal code, in the sense that for the given value of CSNR its performance cannot be improved upon by any other block-structured system with identical values of L and K. For example, from Fig. V.8 ($\rho = 0.95$), when $K/L = 0.125$, we see that the BPAM system is optimum for CSNR values below 0 dB. As the source correlation decreases the value of the CSNR below which the BPAM system becomes optimal, decreases.

For $K/L < 1.0$ and high values of the CSNR, the BPAM curve saturates. This is because the BPAM system becomes dimension-limited. As is evident from (V.6), the second term in the parentheses represents the residual distortion due to the uncoded source dimensions. While the first term can be made arbitrarily small by increasing the transmitted energy (i.e., by increasing θ), the second term remains unchanged and determines the saturation value of the BPAM performance as the CSNR becomes large and is given by $10\log_{10}(\sum_{i=K+1}^{L}\lambda_i/(L\sigma^2))$, where σ^2 is the source variance. The dimension limitation is severe when the source correlation is low, since the residual term in (V.6) is large. This fact is also evident from Figs. V.5–V.8. For example, from Fig. V.5 ($\rho = 0.0$), when $K/L = 0.25$, the BPAM performance curve is saturated for CSNR values above 6.0 dB. However, from Fig. V.8 ($\rho = 0.95$), and for the same value of K/L, the BPAM curve begins to saturate for CSNR values above 9 dB.

V.6.B Performance of the BPLE System

Even though the BPAM system performance serves as an upper bound on the performance of the BPLE system, our results indicate that when the channel is noisy, we can design BPLE systems with a performance close to that of the BPAM system. For example, when $\rho = 0{:}0$, $L = 1$ and $K/L = 1.0$, it is possible to design for any given value of the CSNR in the range that we consider, a scalar quantizer with $R_s = 8.0$, that performs within 0.15 dB of the BPAM bound (and in this case the OPTA as well). For a highly correlated source with $\rho = 0.95$, our computations indicate that we can design a BPLE system whose performance is within 0.5 dB of the BPAM bound for CSNR= 0.0 dB. The difference between the performance of a BPLE system and the BPAM system gets smaller as the block size is increased for a given R_s, CSNR and ρ. A similar trend can be observed as the source correlation is increased for a fixed block size L and encoding rate R_s and when the bandwidth is reduced for a fixed rate, block size and source correlation. For example at a CSNR of 0 dB, $L = 8$ and $R_s = 1.0$, there is a 2 dB difference in performance for $\rho = 0$ whereas for $\rho = 0.95$ this difference is less than 0.5 dB. At high values of the CSNR, the performance of a BPLE system becomes rate-limited, rather than capacity-limited and hence it saturates.

It is interesting to note that for small block sizes, significant improvements in performance can be obtained by increasing encoding rate beyond the capacity of the channel. For example, with $\rho = 0.0$ and $K = L = 1$, at a CSNR of 9.0 dB, the channel capacity is approximately 2 bits/source sample. From Fig. V.1a, we see that roughly 5.5 dB can be gained by increasing the encoding rate above 2 bits/source sample.

For $K/L < 1.0$, the BPLE system is subject to both a dimension limitation and a rate limitation and it is easy to tell, by inspection, which effect is the dominant one. For example, from Fig. V.7 ($\rho = 0.9$), we see that the $R_s = 1.0$, $L = 8$, BPLE system becomes rate-limited before it becomes dimension limited, when $K/L = 0.5$, since it saturates at a lower SNR than the BPAM system does. However, when $K/L = 0.25$ it is the dimension limitation, rather than the rate limitation, that determines the saturation performance of the BPLE system, since the performance of the two systems stay close together over the entire range of CSNR values considered.

Relative to the LBGDC system, the BPLE system can result in significant performance improvements for low CSNR values. For example, at a CSNR of 0 dB, $K/L = 0.5$, $L = 8$, $\rho = 0.0$ and $\rho = 0.9$, the BPLE system results in gains of 2.0 dB and 3.7 dB, respectively. For intermediate values of the CSNR, we note that the LBGDC system outperforms the BPLE system. Clearly, for these CSNR values, the linearity assumption is a poor one, especially for the i.i.d. source and $K/L < 1.0$. We would like to note that when $L = K$ and the channel is relatively noiseless, the performance of the BPLE system is very close to that obtained by the LBGDC system.

The optimum signal constellations displayed in Fig. V.12 As the source correlation is increased for fixed N, greater amounts of energy are transmitted along one channel dimension at the expense of the other. This is a direct consequence of (III.68–III.71) and the fact that there is a greater spread in the eigenvalues as the correlation coefficient is increased.

V.6.C Open Issues

We have demonstrated that the OPTA for the vector source can be achieved by an analog block communication system in certain cases. It would be interesting to determine whether this is true in more general situations. However, it will be necessary for the encoder to be a nonlinear map from the source space to the channel space and for the decoder to be a nonlinear map from the channel space to the source space. If such a system can be designed, the next issue would be that of determining good finite rate approximations to this system, in order to get close to the OPTA using digital communication systems. It would also be useful to design such nonlinear analog communication systems when the source is non-Gaussian, because the BPAM performance for any source is determined by the second-order statistics of the source and hence is no different from the BPAM performance for a Gaussian source. On the other hand the OPTA for a non-Gaussian source will increase, and hence so will the gap between the OPTA and the BPAM performance. Finally, even in the linear case the encoding complexity places an upper bound on the rate for which a BPLE system can be designed or implemented. In order to achieve the performance promised by the BPAM system for high values of CSNR, a more structured approach to VQ design would be useful.

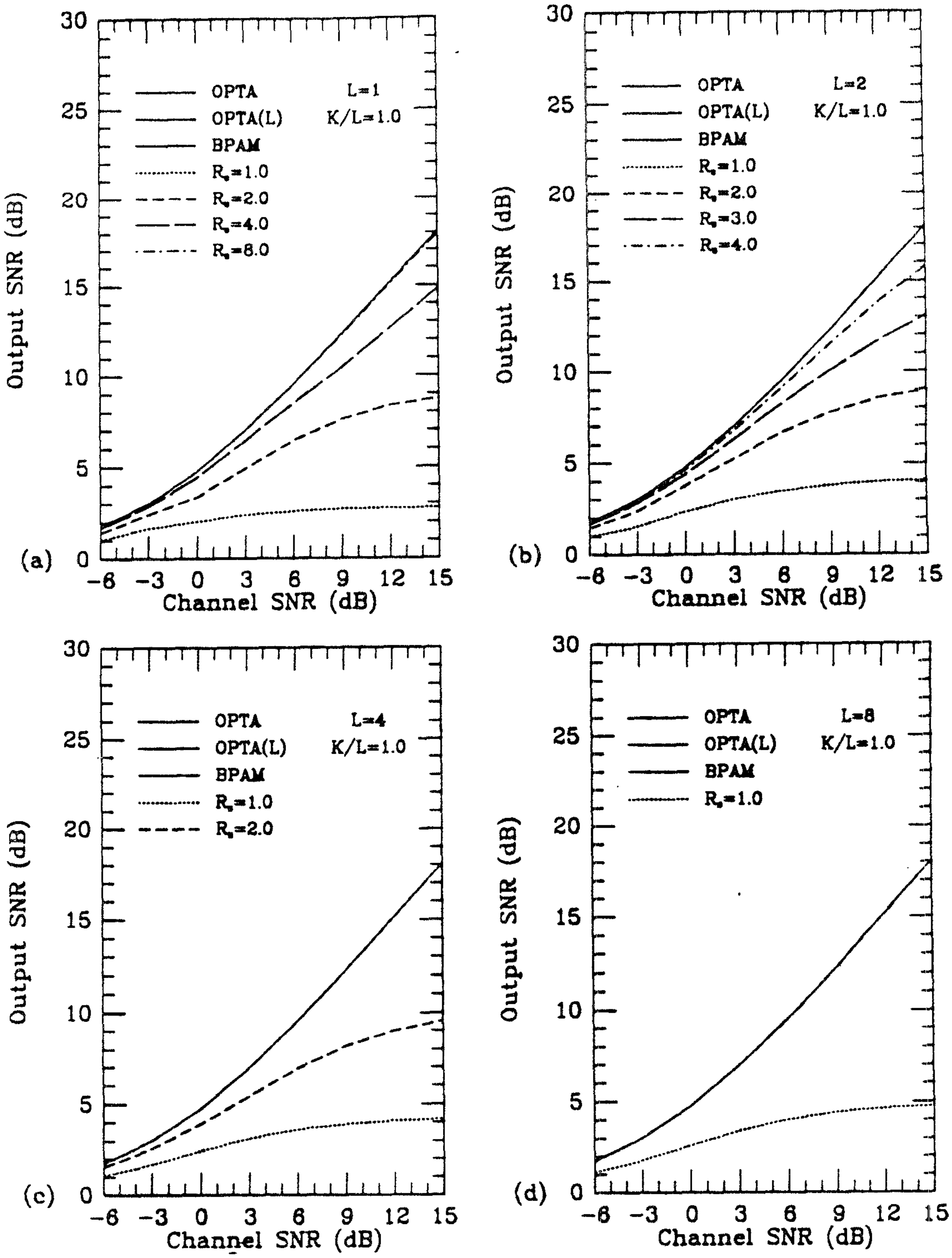

Figure V.1: Performance Results for the BPLE System; Memoryless Gaussian Source, $K/L = 1.0$; R_s: Encoding Rate of the BPLE System (bits/sample).

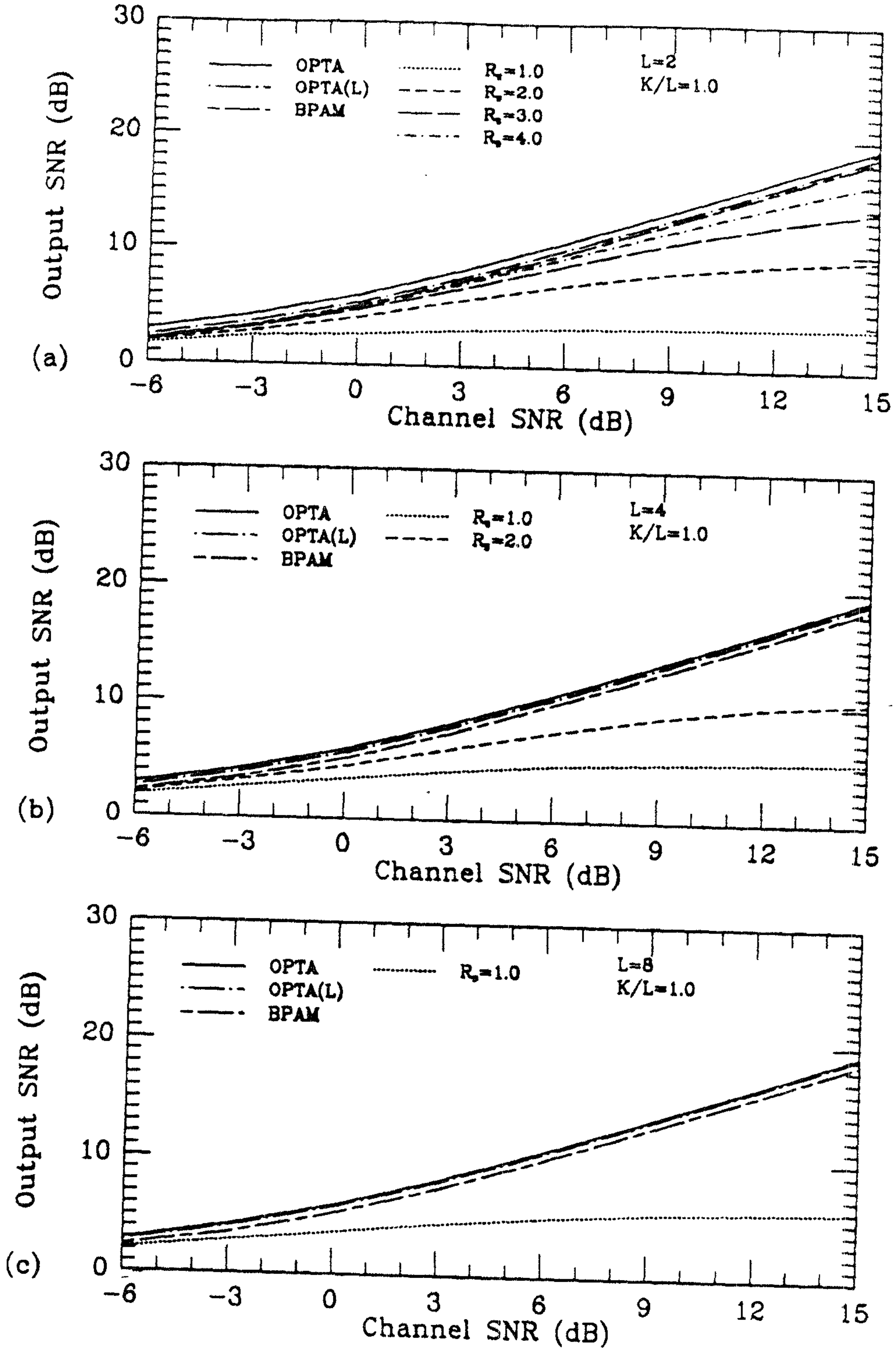

Figure V.2: Performance Results for the BPLE System; First-Order Gauss-Markov Source, $\rho = 0.5$; R_s: Encoding Rate of the BPLE System (bits/ sample).

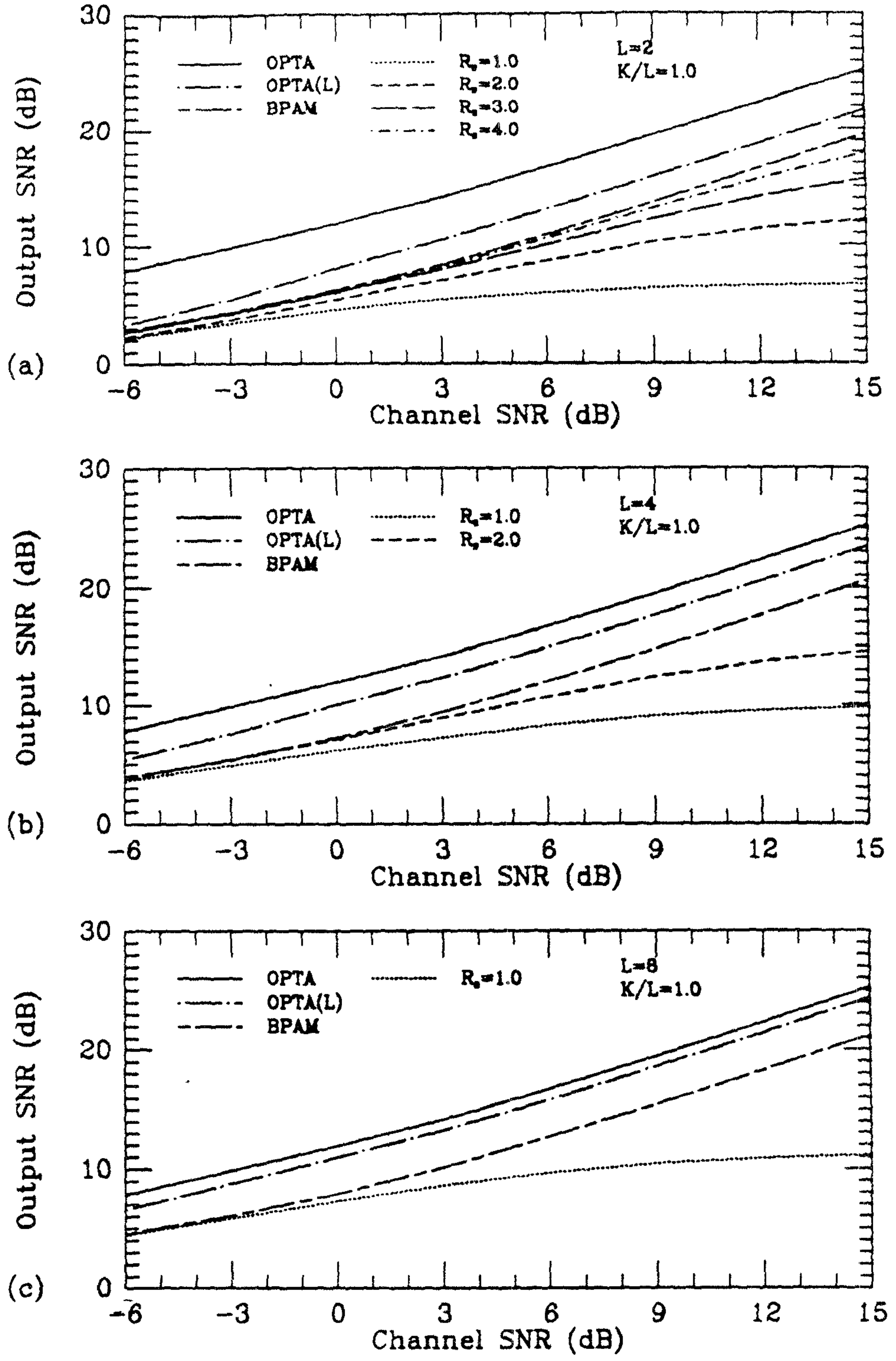

Figure V.3: Performance Results for the BPLE System; First-Order Gauss-Markov Source, $\rho = 0.9$; R_s: Encoding Rate of the BPLE System (bits/sample).

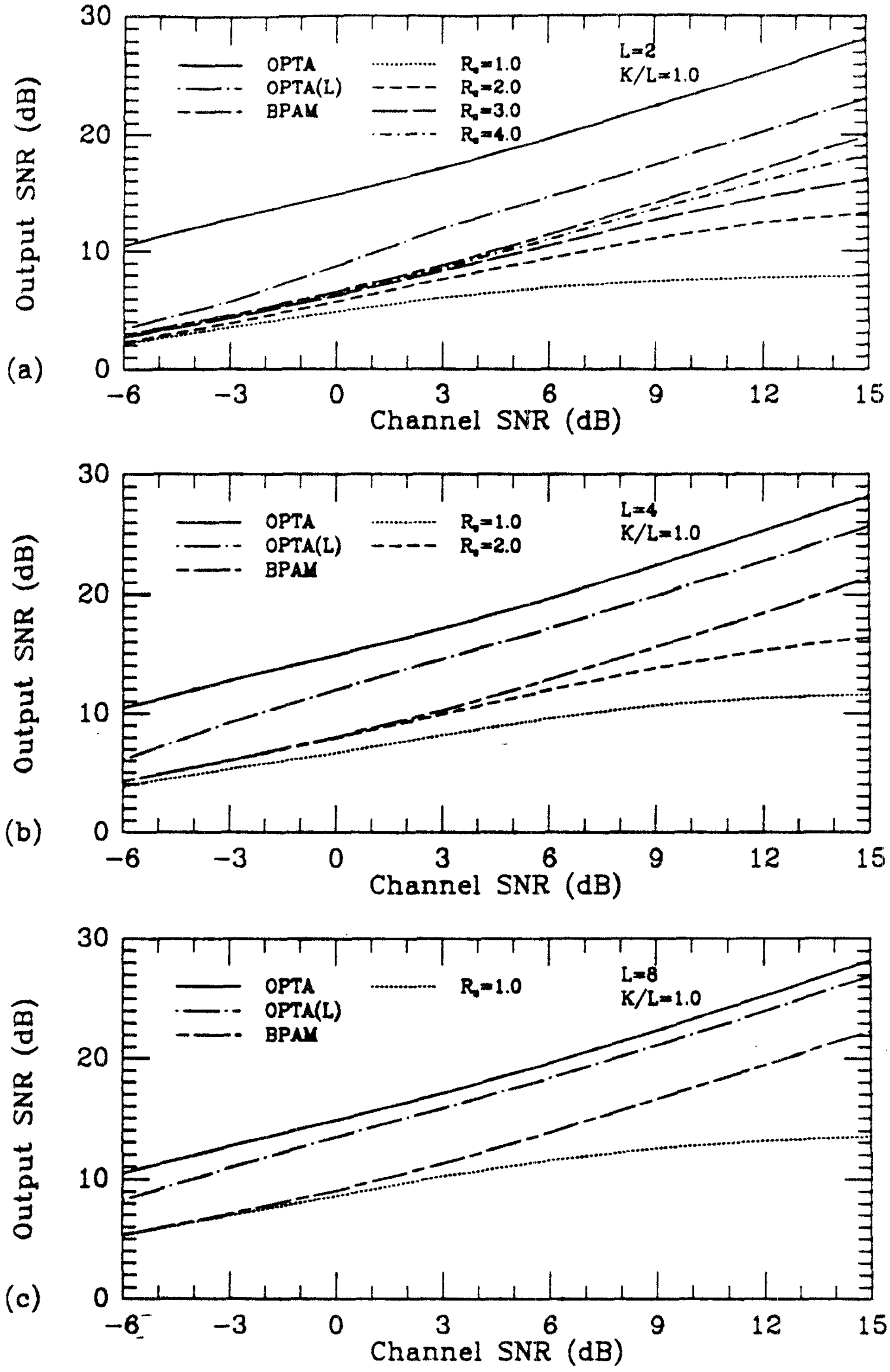

Figure V.4: Performance Results for the BPLE System; First-Order Gauss-Markov Source, $\rho = 0.95$; R_s: Encoding Rate of the BPLE System (bits/sample).

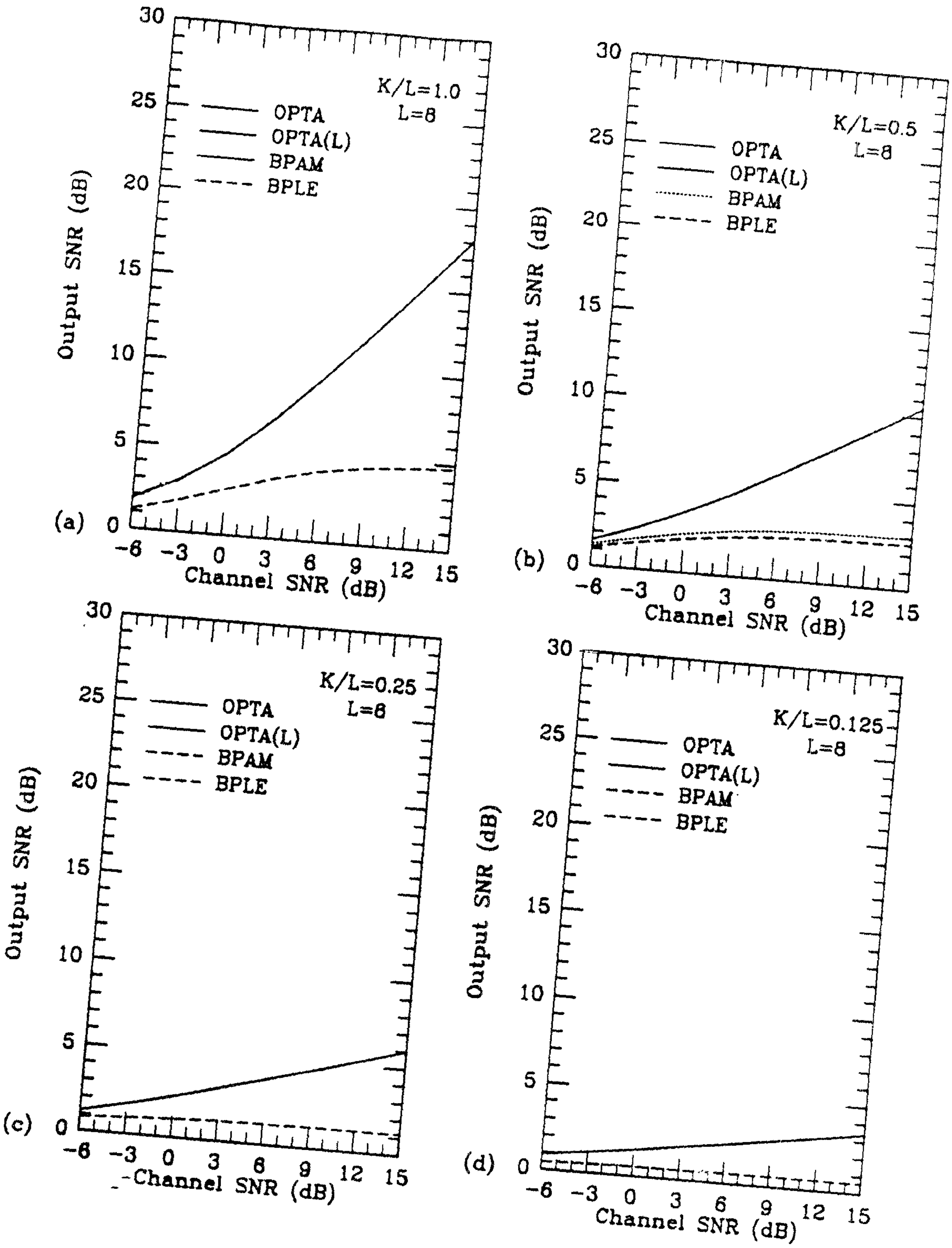

Figure V.5: Performance Results for the BPLE System; Memoryless Gaussian Source, $L = 8$, $R_s = 1.0$. (a) $K/L = 1.0$. (b) $K/L = 0.5$. (c) $K/L = 0.25$. (d) $K/L = 0.125$.

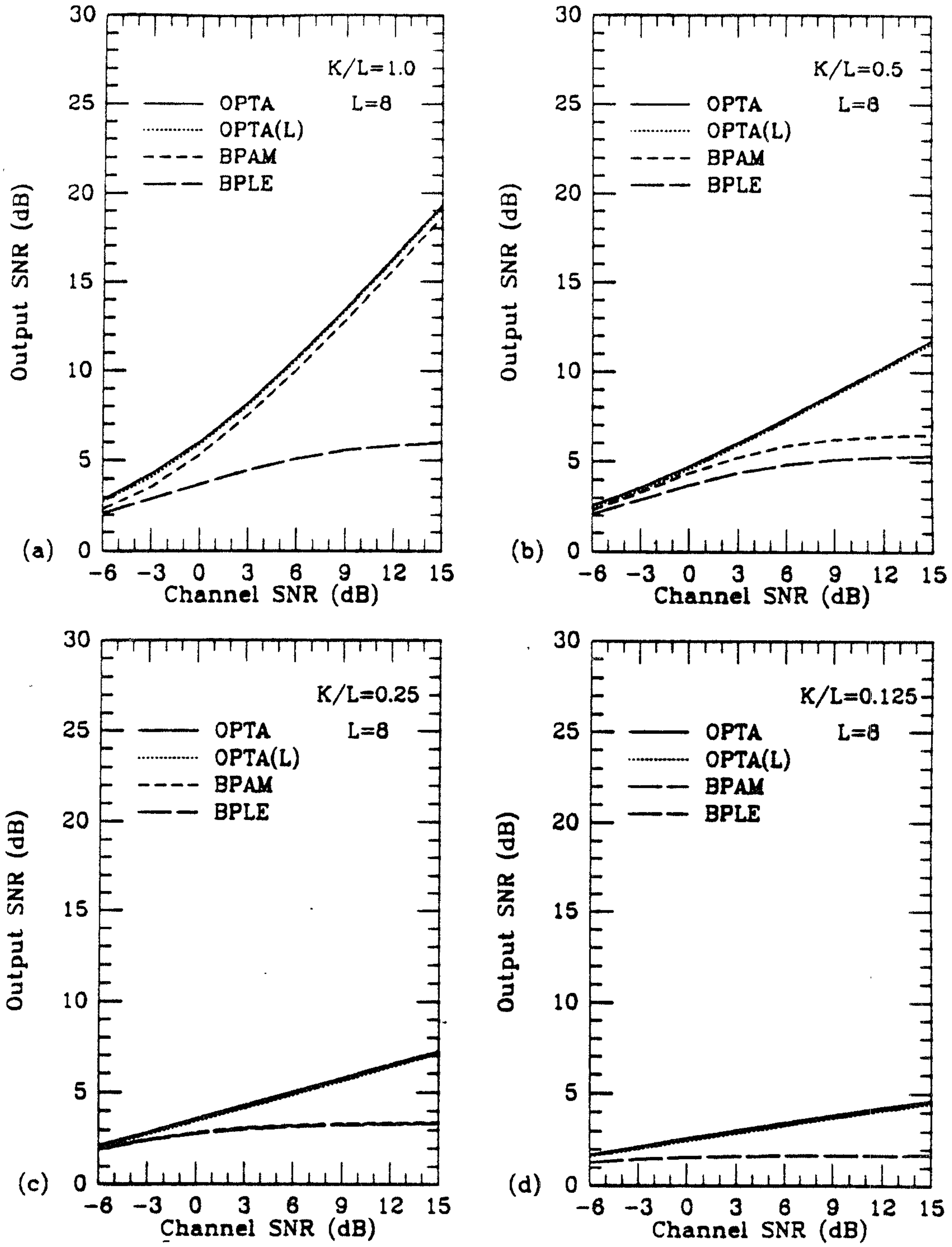

Figure V.6: Performance Results for the BPLE System; First-Order Gauss-Markov Source, $\rho = 0.5$, $L = 8$, $R_s = 1.0$. (a) $K/L = 1.0$. (b) $K/L = 0.5$. (c) $K/L = 0.25$. (d) $K/L = 0.125$.

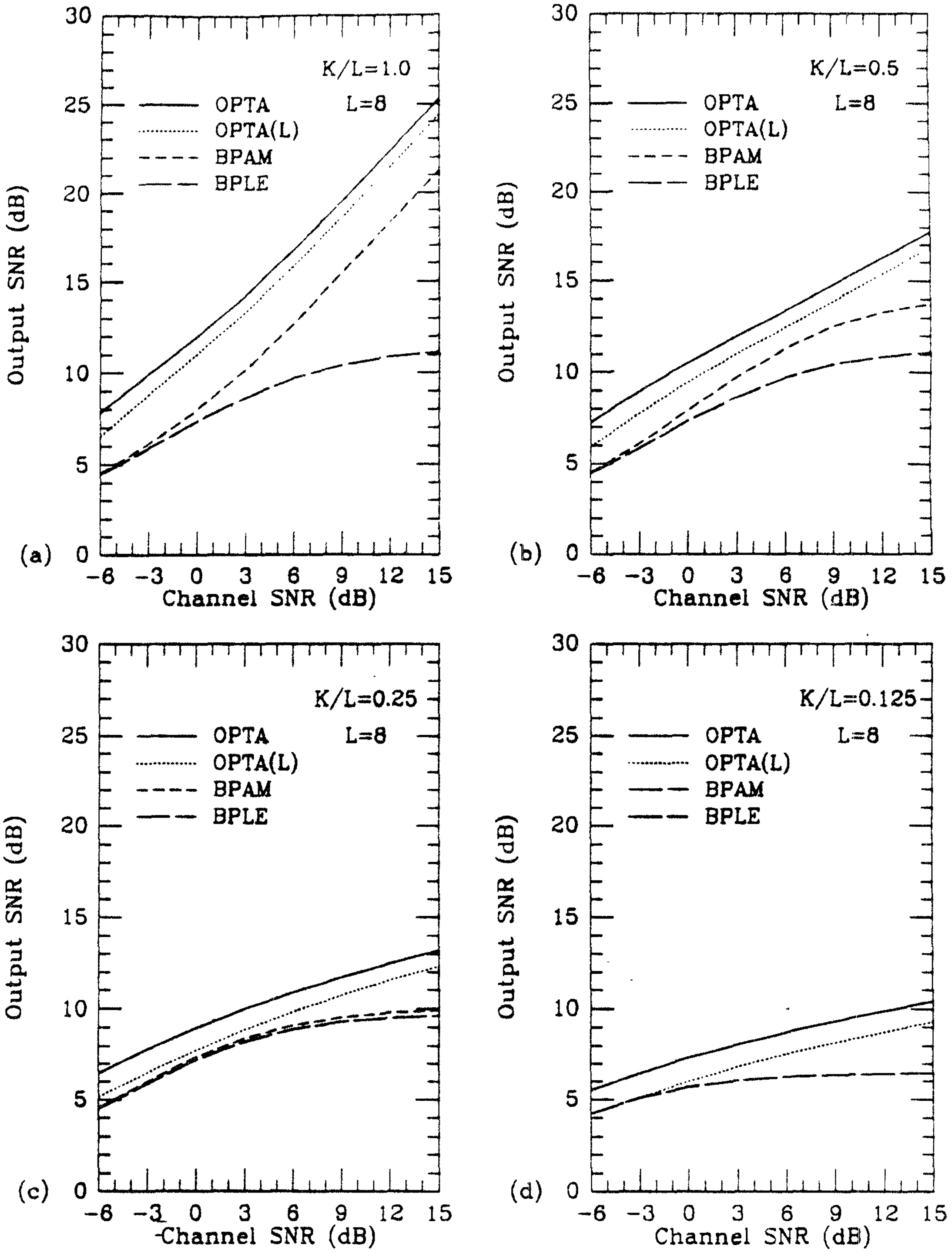

Figure V.7: Performance Results for the BPLE System; First-Order Gauss-Markov Source, $\rho = 0.9$, $L = 8$, $R_s = 1.0$. (a) $K/L = 1.0$. (b) $K/L = 0.5$. (c) $K/L = 0.25$. (d) $K/L = 0.125$.

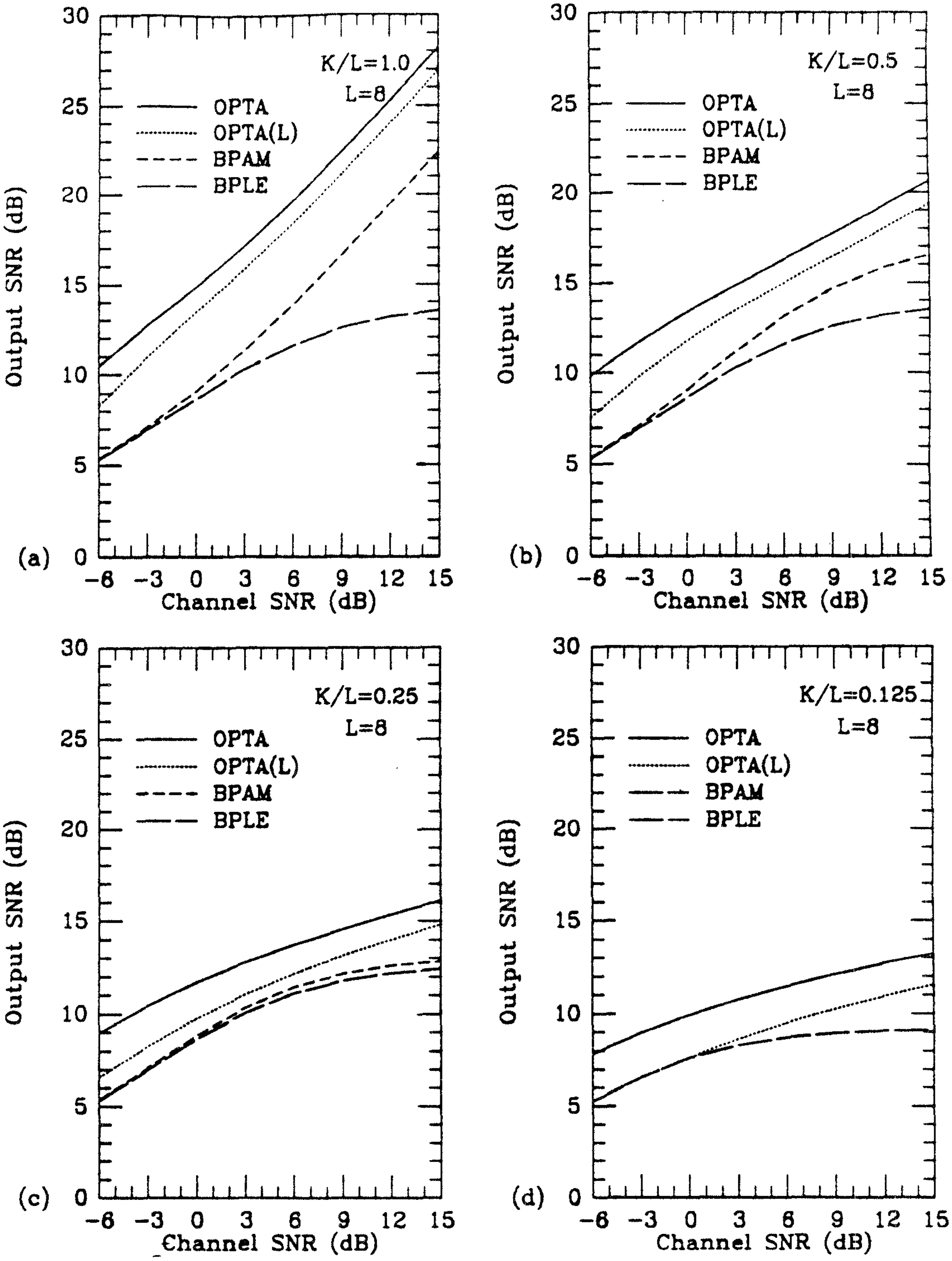

Figure V.8: Performance Results for the BPLE System; First-Order Gauss-Markov Source, $\rho = 0.95$, $L = 8$, $R_s = 1.0$. (a) $K/L = 1.0$. (b) $K/L = 0.5$. (c) $K/L = 0.25$. (d) $K/L = 0.125$.

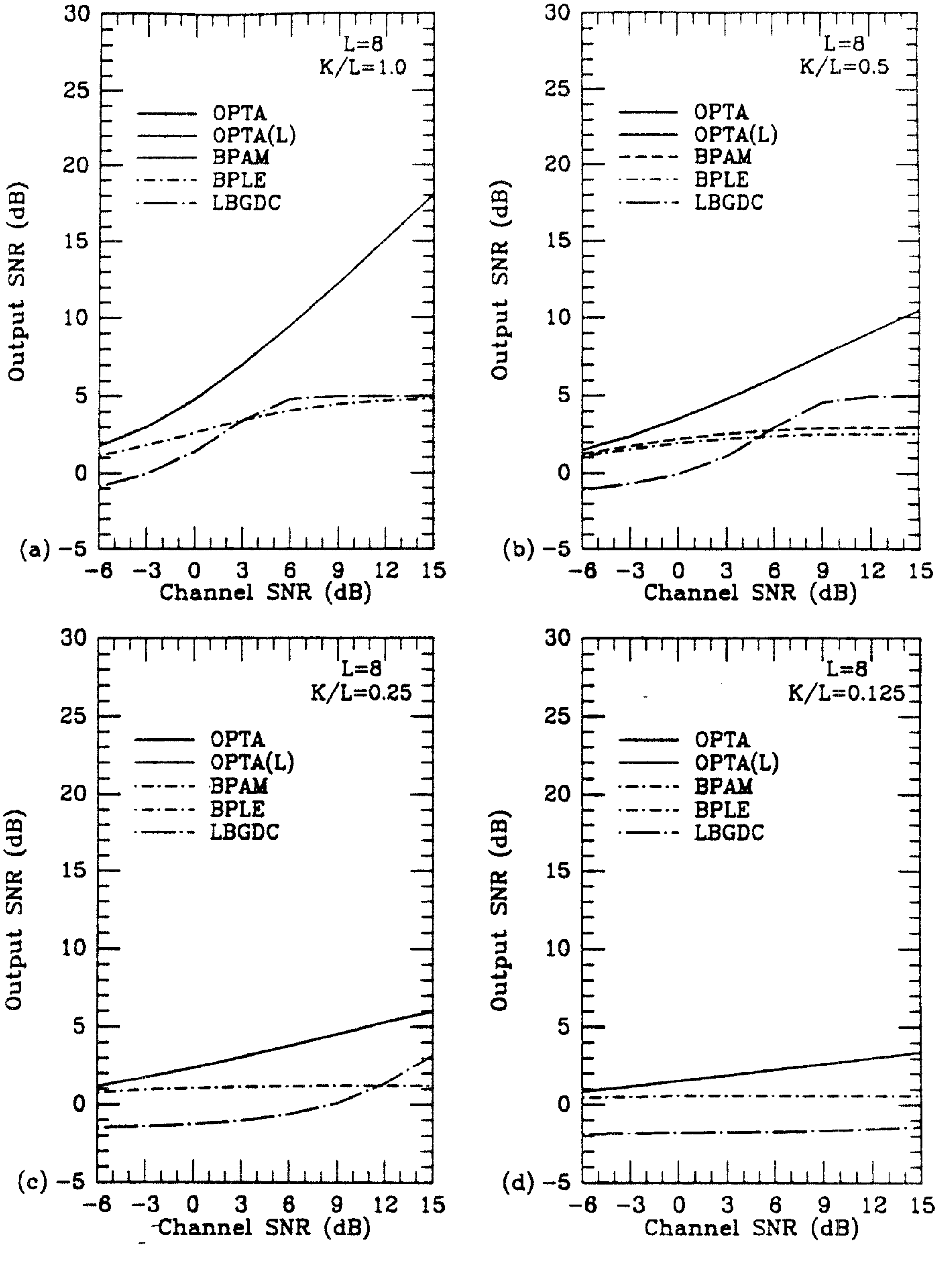

Figure V.9: Performance Comparisons, BPLE System vs. LBGDC System; Memoryless Gaussian Source, $L = 8$, $R_s = 1.0$. (a) $K/L = 1.0$. (b) $K/L = 0.5$. (c) $K/L = 0.25$. (d) $K/L = 0.125$.

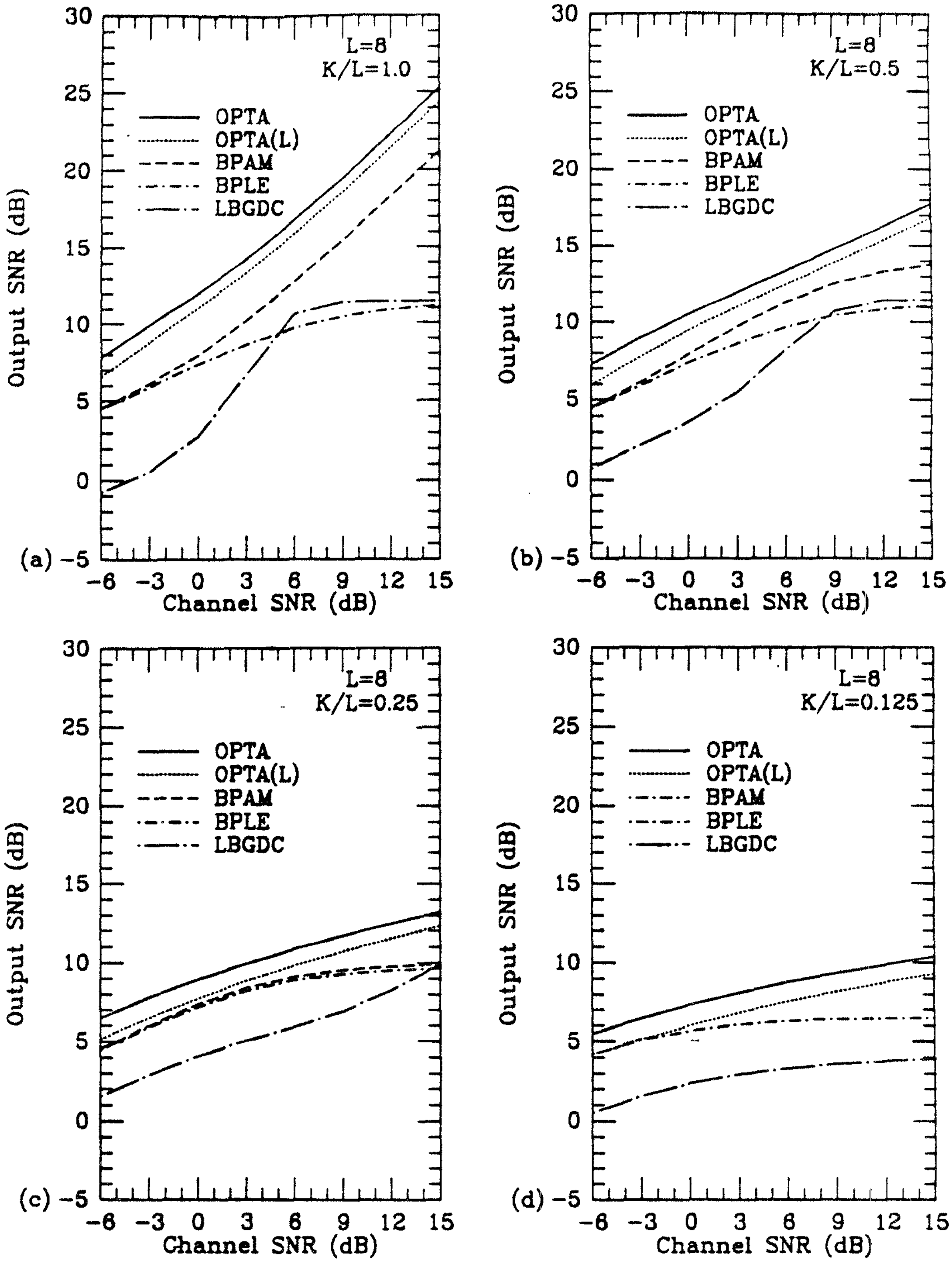

Figure V.10: Performance Comparisons, BPLE System vs. LBGDC System; First-Order Gauss-Markov Source, $\rho = 0.9$, $L = 8$, $R_s = 1.0$. (a) $K/L = 1.0$. (b) $K/L = 0.5$. (c) $K/L = 0.25$. (d) $K/L = 0.125$.

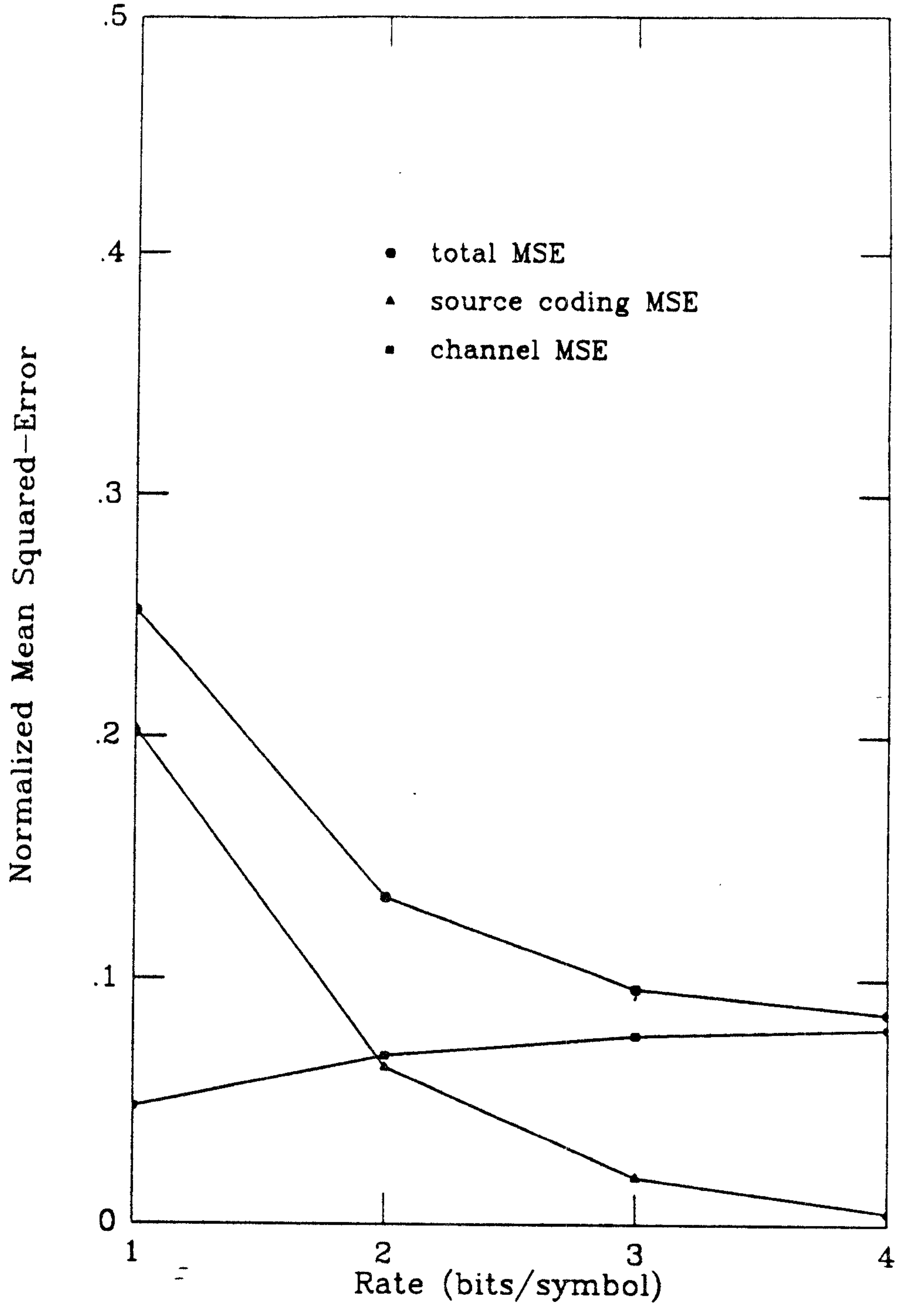

Figure V.11: Overall, Source and Channel MSE for the BPLE System as a function of the Encoding Rate; $L = 2$, $K/L = 1.0$, CSNR=6 dB, $\rho = 0.9$.

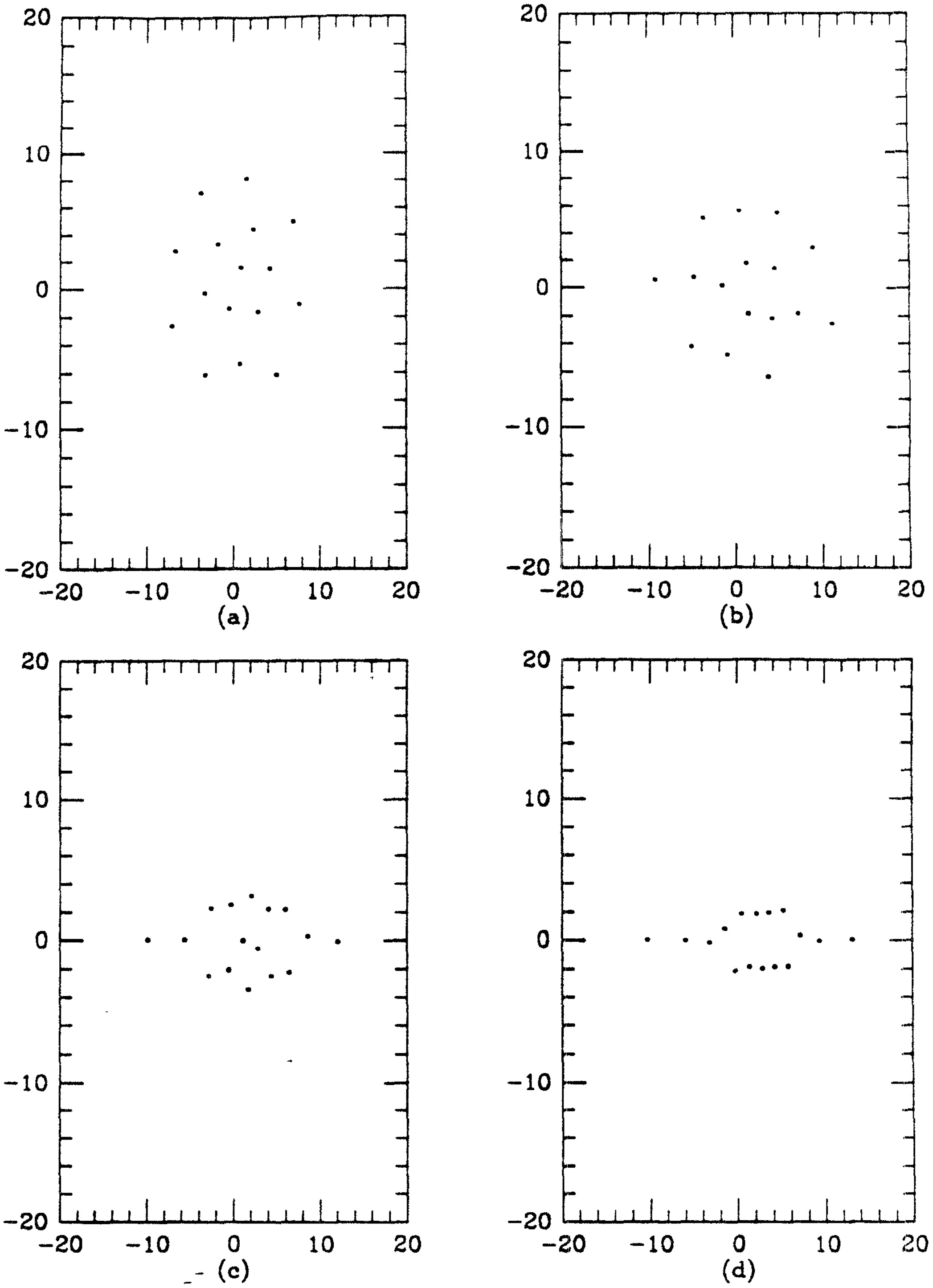

Figure V.12: 16-ary Signal Constellations; $L = 2$, $K/L = 1.0$, CSNR=9.0 dB, First-Order Gauss Markov Source. (a) $\rho = 0.0$. (b) $\rho = 0.5$. (c) $\rho = 0.9$. (d) $\rho = 0.95$.

VI Summary and Conclusions

We have considered the problem of designing block source codes and modulation signal sets that are both energy and bandwidth constrained. We have demonstrated that the class of linear estimator based decoders is asymptotically optimal in the limit of low CSNR. Based on this fact, we have derived necessary conditions for optimality for the encoder, decoder and modulation signal set. An algorithm that iteratively solves these necessary conditions to converge to a locally optimum solution has been developed.

By studying the performance of the above class of digital communication systems in the limit of infinite encoding rates, we have demonstrated that the MSE of a bandwidth and energy constrained digital system is bounded from below by that of a block pulse amplitude modulation system. This bound is readily computable in terms of the eigenvalues of the source and channel covariance matrices.

Our performance calculations indicate that it is possible to design digital systems whose performance is close to that of the BPAM bound in selected cases. We have demonstrated that in selected cases good performance can be achieved, as compared to the OPTA, and that significant improvements in performance can be achieved as compared to an LBGDC system.

A Final Step of Theorem 2.1

Our goal is to prove the last step of Theorem 2.1. Specifically, we wish to prove that (see II.19).

$$\lim_{\mathcal{E}\to 0} \frac{\|g(\mathbf{u}) - 2\sum_{i=1}^{N} P_i E(\mathbf{X}|i)\mathbf{s}_i^T\mathbf{u}/N_0\|}{\sqrt{\mathcal{E}}} = 0 \tag{A.1}$$

Proof: From (II.19) it follows that

$$\frac{\|g(\mathbf{u}) - 2\sum_{i=1}^{N} P_i E(\mathbf{X}|i)\mathbf{s}_i^T\mathbf{u}/N_0\|}{\sqrt{\mathcal{E}}} = \tag{A.2}$$

$$= \frac{\|\sum_{i=1}^{N} P_i E(\mathbf{X}|i)o(\mathbf{s}_i,\mathbf{u})/p_z(\mathbf{u}) - 2(\sum_{i=1}^{N} P_i E(\mathbf{X}|i)\mathbf{s}_i^T\mathbf{u})(\sum_{i=1}^{N} P_i o(\mathbf{s}_i,\mathbf{u}))/N_0 p_z(\mathbf{u})\|}{|1 + \sum_{i=1}^{N} P_i o(\mathbf{s}_i,\mathbf{u})/p_z(\mathbf{u})|\sqrt{\mathcal{E}}} \tag{A.3}$$

$$\leq \frac{\sum_{i=1}^{N} P_i\|E(\mathbf{X}|i)\|\ |o(\mathbf{s}_i,\mathbf{u})| + 2\left(\sum_{i=1}^{N} P_i\|E(\mathbf{X}|i)\mathbf{s}_i^T\mathbf{u}\|\right)\left(\sum_{i=1}^{N} P_i|o(\mathbf{s}_i,\mathbf{u})|\right)/N_0}{p_z(\mathbf{u})\sqrt{\mathcal{E}}|1 + \sum_{i=1}^{N} P_i o(\mathbf{s}_i,\mathbf{u})/p_z(\mathbf{u})|} \tag{A.4}$$

It suffices to prove that $|o(\mathbf{s}_i,\mathbf{u})|/\sqrt{\mathcal{E}} \to 0$ as $\mathcal{E} \to 0$, $\forall i$ and $\forall\mathbf{u}$. But

$$\frac{|o(\mathbf{s}_i,\mathbf{u})|}{\sqrt{\mathcal{E}}} = \frac{|o(\mathbf{s}_i,\mathbf{u})|}{\|\mathbf{s}_i\|}\frac{\|\mathbf{s}_i\|}{\sqrt{\mathcal{E}}} \tag{A.5}$$

$$\leq \frac{|o(\mathbf{s}_i,\mathbf{u})|}{\|\mathbf{s}_i\|}\frac{1}{\sqrt{P_i}}. \tag{A.6}$$

Hence $\lim_{\mathcal{E}\to 0}\frac{|o(\mathbf{s}_i,\mathbf{u})|}{\sqrt{\mathcal{E}}} = 0$, and this proves the last step of Theorem 2.1.

B Parametric Representation of BPAM Performance

We derive a parametric representation for the MSE of a BPAM system as a function of the transmitted energy. The MSE of a BPAM system, $D_B(\mathcal{E}_B, L, K)$, for average transmitted energy $\mathcal{E}_B$, is given by

$$D_B(\mathcal{E}_B, L, K) = \frac{1}{L}\left(\sum_{i=1}^{K} \frac{\lambda_i N_0/2}{\mathcal{E}_i + N_0/2} + \sum_{i=K+1}^{L} \lambda_i\right), \tag{B.1}$$

where $\sum_{i=1}^{K} \mathcal{E}_i/L = \mathcal{E}_B$ and $\mathcal{E}_i$ is the optimum value of the average energy transmitted along channel dimension i. Let $-\psi$, the value of $\partial LD/\partial\mathcal{E}_j$, for $\mathcal{E}_j > 0$ be a free parameter. Then, given a value of $\psi > 0$, $\mathcal{E}_i > 0$ iff $2\lambda_i/N_0 > \psi$, in which case $\psi = \lambda_i N_0/(2(\mathcal{E}_i + N_0/2)^2)$ holds. Equivalently, we can write

$$\mathcal{E}_i = \begin{cases} \sqrt{\frac{\lambda_i N_0}{2\psi}} - \frac{N_0}{2} & \text{if } 2\lambda_i/N_0 > \psi, \\ 0 & \text{otherwise.} \end{cases} \tag{B.2}$$

which in turn can be written more compactly as $\mathcal{E}_i = \max\left(\sqrt{\frac{\lambda_i N_0}{2\psi}} - \frac{N_0}{2}, 0\right)$. Substitute this expression for $\mathcal{E}_i$ in (B.1) and define parameter $\theta = \sqrt{\lambda_i N_0 \psi/2}$, in order to arrive at the desired parametric representation, namely,

$$D_B(\theta, L, K) = \frac{1}{L}\left(\sum_{i=1}^{K} \min\left[\theta, \lambda_i\right] + \sum_{i=K+1}^{L} \lambda_i\right), \tag{B.3}$$

and

$$\mathcal{E}_B(\theta, L, K) = \frac{N_0 K}{2L}\left(\frac{1}{K}\sum_{i=1}^{K} \max\left[\lambda_i/\theta, 1\right] - 1\right). \tag{B.4}$$

C Tables of Selected Performance Results

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BPLE($R_s = 8.0$)	1.77	3.01	4.77	6.97	9.50	12.23	15.07	17.96
BPAM	1.77	3.02	4.77	6.98	9.52	12.28	15.15	18.08
OPTA(L)	1.77	3.02	4.77	6.98	9.52	12.28	15.15	18.08
OPTA	1.77	3.02	4.77	6.98	9.52	12.28	15.15	18.08

Table C.1: Performance Results for the BPLE System; Memoryless Gaussian Source, $L = 1$, $K/L = 1.0$.

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BPLE($R_s = 4.0$)	1.73	2.95	4.64	6.75	9.10	11.50	13.81	15.80
BPAM	1.77	3.02	4.77	6.98	9.52	12.28	15.15	18.08
OPTA(L)	1.77	3.02	4.77	6.98	9.52	12.28	15.15	18.08
OPTA	1.77	3.02	4.77	6.98	9.52	12.28	15.15	18.08

Table C.2: Performance Results for the BPLE System; Memoryless Gaussian Source, $L = 2$, $K/L = 1.0$.

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BPLE($R_s = 2.0$)	1.54	2.58	3.95	5.47	6.93	8.16	9.01	9.56
BPAM	1.77	3.02	4.77	6.98	9.52	12.28	15.15	18.08
OPTA(L)	1.77	3.02	4.77	6.98	9.52	12.28	15.15	18.08
OPTA	1.77	3.02	4.77	6.98	9.52	12.28	15.15	18.08

Table C.3: Performance Results for the BPLE System; Memoryless Gaussian Source, $L = 4$, $K/L = 1.0$.

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BPLE($R_s = 1.0$)	1.12	1.80	2.62	3.42	4.04	4.45	4.67	4.79
LBGDC($R_s = 1.0$)	-0.88	-0.06	1.37	3.41	4.76	4.95	4.96	4.96
BPAM	1.77	3.02	4.77	6.98	9.52	12.28	15.15	18.08
OPTA(L)	1.77	3.02	4.77	6.98	9.52	12.28	15.15	18.08
OPTA	1.77	3.02	4.77	6.98	9.52	12.28	15.15	18.08

Table C.4: Performance Results for the BPLE and LBGDC Systems; Memoryless Gaussian Source, $L = 8$, $K/L = 1.0$.

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BPLE($R_s = 1.0$)	1.11	1.54	1.93	2.22	2.39	2.49	2.54	2.57
LBGDC($R_s = 1.0$)	-1.08	-0.70	-0.10	1.05	2.92	4.55	4.94	4.96
BPAM	1.25	1.76	2.22	2.55	2.76	2.88	2.94	2.98
OPTA(L)	1.51	2.39	3.49	4.77	6.14	7.58	9.04	10.53
OPTA	1.51	2.39	3.49	4.77	6.14	7.58	9.04	10.53

Table C.5: Performance Results for the BPLE and LBGDC Systems; Memoryless Gaussian Source, $L = 8$, $K/L = 0.5$.

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BDCE($R_s = 4.0$)	1.24	1.93	2.72	3.87	4.91	5.74	7.07	8.59
LBGDC($R_s = 4.0$)	-2.48	-1.73	-1.20	-0.86	-0.65	-0.53	-0.43	-0.31
BPAM	1.25	1.76	2.22	2.55	2.76	2.88	2.94	2.98
OPTA(L)	1.51	2.39	3.49	4.77	6.14	7.58	9.04	10.53
OPTA	1.51	2.39	3.49	4.77	6.14	7.58	9.04	10.53

Table C.6: Performance Results for the BDCE and LBGDC Systems; Memoryless Gaussian Source, $L = 2$, $K/L = 0.5$.

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BDCE($R_s = 2.0$)	0.81	1.17	1.69	2.25	2.86	3.48	3.96	4.69
BPAM	0.79	0.97	1.09	1.16	1.21	1.23	1.24	1.24
OPTA(L)	1.20	1.75	2.39	3.07	3.79	4.52	5.27	6.01
OPTA	1.20	1.75	2.39	3.07	3.79	4.52	5.27	6.01

Table C.7: Performance Results for the BDCE and LBGDC Systems; Memoryless Gaussian Source, $L = 4$, $K/L = 0.25$.

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BDCE($R_s = 1.0$)	0.57	0.78	1.02	1.30	1.61	1.91	2.21	2.52
LBGDC($R_s = 1.0$)	-1.86	-1.84	-1.81	-1.78	-1.73	-1.66	-1.56	-1.41
BPAM	0.46	0.51	0.54	0.56	0.57	0.58	0.58	0.58
OPTA(L)	0.88	1.19	1.54	1.90	2.26	2.63	3.01	3.38
OPTA	0.88	1.19	1.54	1.90	2.26	2.63	3.01	3.38

Table C.8: Performance Results for the BDCE and LBGDC Systems; Memoryless Gaussian Source, $L = 8$, $K/L = 0.125$.

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BPLE($R_s = 4.0$)	2.71	4.26	6.10	8.22	10.65	13.17	15.64	17.86
BDCE($R_s = 4.0$)	2.61	4.24	6.31	8.36	10.56	12.74	14.75	16.62
LBGDC($R_s = 4.0$)	-1.52	0.41	2.44	4.45	6.40	8.23	9.96	11.75
BPAM	2.81	4.36	6.21	8.42	10.96	13.71	16.58	19.52
OPTA(L)	3.27	5.42	8.08	10.59	13.13	15.88	18.75	21.68
OPTA	7.83	9.90	11.97	14.19	16.74	19.49	22.36	25.29

Table C.9: Performance Results for the BPLE, BDCE and LBGDC Systems; 1st-Order Gauss-Markov Source, $\rho = 0.9$, $L = 2$, $K/L = 1.0$.

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BDCE($R_s = 4.0$)	2.78	4.34	6.16	8.02	9.67	10.94	11.86	12.96
LBGDC($R_s = 4.0$)	-2.45	-0.43	1.53	3.38	5.00	6.31	7.31	8.04
BPAM	2.81	4.36	6.20	8.07	9.74	11.02	11.89	12.41
OPTA(L)	2.81	4.36	6.20	8.07	9.74	11.18	12.65	14.13
OPTA	7.27	8.96	10.53	11.97	13.35	14.79	16.26	17.74

Table C.10: Performance Results for the BDCE and LBGDC Systems; 1st-Order Gauss-Markov Source, $\rho = 0.9$, $L = 2$, $K/L = 0.5$.

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BPLE($R_s = 1.0$)	4.47	5.88	7.35	8.63	9.70	10.45	10.90	11.16
BDCE($R_s = 1.0$)	2.78	4.28	6.29	8.58	10.66	11.47	11.48	11.48
LBGDC($R_s = 1.0$)	-0.78	0.51	2.80	6.65	10.59	11.43	11.44	11.44
BPAM	4.53	6.11	7.94	10.15	12.70	15.45	18.32	21.25
OPTA(L)	6.55	8.78	11.02	13.29	15.84	18.59	21.46	24.39
OPTA	7.83	9.90	11.97	14.19	16.74	19.49	22.36	25.29

Table C.11: Performance Results for the BPLE, BDCE and LBGDC Systems; 1st-Order Gauss-Markov Source, $\rho = 0.9$, $L = 8$, $K/L = 1.0$.

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BPLE($R_s = 1.0$)	4.47	5.88	7.35	8.63	9.71	10.44	10.87	11.11
BDCE($R_s = 1.0$)	3.81	5.25	6.38	8.06	9.83	10.78	11.46	11.48
LBGDC($R_s = 1.0$)	0.71	2.17	3.63	5.49	8.21	10.75	11.42	11.44
BPAM	4.53	6.11	7.90	9.72	11.31	12.52	13.31	13.78
OPTA(L)	5.95	7.76	9.46	11.01	12.45	13.89	15.36	16.84
OPTA	7.27	8.96	10.53	11.97	13.35	14.79	16.26	17.74

Table C.12: Performance Results for the BPLE, BDCE and LBGDC Systems; 1st-Order Gauss-Markov Source, $\rho = 0.9$, $L = 8$, $K/L = 0.5$.

	CSNR(dB)							
	−6.0	−3.0	0.0	3.0	6.0	9.0	12.0	15.0
BPLE($R_s = 1.0$)	4.47	5.90	7.21	8.22	8.91	9.31	9.54	9.66
BDCE($R_s = 1.0$)	4.30	5.68	6.97	7.97	8.78	9.45	10.15	10.82
LBGDC($R_s = 1.0$)	1.55	2.93	4.09	5.06	5.95	6.93	8.27	9.94
BPAM	4.53	6.00	7.36	8.43	9.14	9.56	9.80	9.93
OPTA(L)	5.14	6.51	7.75	8.87	9.85	10.74	11.55	12.32
OPTA	6.49	7.81	8.96	9.98	10.89	11.71	12.48	13.22

Table C.13: Performance Results for the BPLE, BDCE and LBGDC Systems; 1st-Order Gauss-Markov Source, $\rho = 0.9$, $L = 8$, $K/L = 0.25$.

References

[1] J. W. Modestino and D. G. Daut, "Combined source-channel coding of images," *IEEE Trans. Commun.*, vol. COM-27, pp. 1644–1659, November 1979.

[2] T. Fine, "Properties of an optimum digital system and applications," *IEEE Trans. Inform. Theory*, vol. IT-10, pp. 287–296, October 1964.

[3] A. E. Kurtenbach and P. A. Wintz, "Quantizing for noisy channels," *IEEE Trans. Commun. Techn.*, vol. COM-17, pp. 291–302, April 1969.

[4] N. Farvardin and V. Vaishampayan, "Optimal quantizer design for noisy channels: An approach to combined source-channel coding," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 827–838, November 1987.

[5] N. Farvardin and V. Vaishampayan, "Some issues on vector quantization for noisy channels," in *Abstracts of Papers, IEEE Int. Symp. Inform. Theory*, Kobe, Japan, pp. 163, June 1988.

[6] E. Ayanoglu and R. M. Gray, "The design of joint source and channel trellis waveform coders," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 855–865, November 1987.

[7] C. M. Thomas, M. Y. Weidner, and S. H. Durrani, "Digital amplitude-phase keying with M-ary alphabets," *IEEE Trans. Commun.*, vol. COM-22, pp. 168–180, February 1974.

[8] G. J. Foschini, R. D. Gitlin, and S. B. Weinstein, "Optimization of two-dimensional signal constellations in the presence of Gaussian noise.," *IEEE Trans. Commun.*, vol. COM-22, pp. 28–38, January 1974.

[9] M. K. Simon and J. G. Smith, "Hexagonal multiple phase-and-amplitude-shift-keyed signal sets," *IEEE Trans. Commun.*, vol. COM-21, pp. 1108–1115, October 1973.

[10] E. Bedrosian, "Weighted PCM," *IRE Trans. Inform. Theory*, vol. IT-4, pp. 45–49, March 1958.

[11] C.-E. W. Sundberg, "Optimum weighted PCM for speech signals," *IEEE Trans. Commun.*, vol. COM-26, pp. 872–881, June 1978.

[12] R. Steele, C.-E. W. Sundberg, and W. C. Wong, "Transmission of log-PCM via QAM over Gaussian and Rayleigh fading channels," *IEE Proceedings*, vol. 134, Pt. F, pp. 539–556, October 1987.

[13] C.-E. W. Sundberg, W. C. Wong, and R. Steele, "Logarithmic PCM weighted QAM transmission over Gaussian and Rayleigh fading channels," *IEE Proceedings*, vol. 134, Pt. F, pp. 557–570, October 1987.

[14] K.-H. Lee and D. P. Petersen, "Optimal linear coding for vector channels," *IEEE Trans. Commun.*, vol. COM-24, pp. 1283–1290, December 1976.

[15] D. W. Tufts, "Nyquist's problem – the joint optimization of transmitter and receiver in pulse amplitude modulation," *Proc. IEEE*, vol. 53, pp. 248–259, March 1965.

[16] T. Berger and D. W. Tufts, "Optimum pulse amplitude modulation, Part I: Transmitter-receiver design and bounds from information theory," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 196–208, April 1967.

[17] C. L. Weber, *Elements of Detection and Signal Design.* New York: McGraw-Hill, 1968.

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

[19] J. G. Proakis, *Digital Communications.* New York: McGraw-Hill, 1983.

[20] W. A. Gardner, "Structural characterization of locally optimum detectors in terms of locally optimum estimators and correlators," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 924–932, November 1982.

[21] A. L. Tits, *Class notes, ENEE 664 (optimal control).* Electrical Engineering Department, University of Maryland, College Park, MD 20740, 1987.

[22] V. Voyevodin, *Linear Algebra.* Moscow: Mir Publishers, 1983.

[23] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, January 1980.

[24] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding.* New York: McGraw-Hill, 1979.

[25] U. Grenander and G. Szegö, *Toeplitz Forms And Their Applications.* Berkeley, California: University of California Press, 1958.

[26] T. Berger, *Rate Distortion Theory: A Mathematical Basis For Data Compression.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1971.

[27] V. Vaishampayan, "Combined source-channel coding for bandlimited waveform channels," Ph.D. Dissertation, Electrical Engineering Department, University of Maryland, College Park, MD 20742, 1989.

List of Figures

List of Tables