Quantization Strategies for Low-Power Communications

Riten Gupta

COMMUNICATIONS & SIGNAL PROCESSING LABORATORY
Department of Electrical Engineering and Computer Science
The University of Michigan
Ann Arbor, MI 48109-2122

May, 2001

Technical Report No. 326

TABLE OF CONTENTS

CHAPTER 1

Introduction

1.1 Overview of Dissertation

Communication systems engineers have for decades tackled problems inherent to the design of efficient and reliable systems. Key objectives have been: low bandwidth utilization, low transmitter power, and high data throughput, all while achieving a reliable communication link [59]. Advances in high-speed electronic devices and circuits have enabled communications engineers to realize significant progress toward these goals. This progress has been the impetus for the burgeoning wireless industry. Overlooked in the design process has been an analysis of the relation of system performance to the power consumption of the digital circuits that comprise the transceiver [16]. Lacking such an analytical power-performance relationship, designers have resorted to trial-and-error methods in efforts to minimize power consumption. The intent of this dissertation is to develop the aforementioned power-performance relationship as well as procedures that utilize the relationship for communication system design.

Any component of a digital communication system comprised of electronic circuitry – digital or analog – is a power-consuming device. The most obvious strategy for power reduction in the digital circuitry is the reduction of the wordlengths used to represent internal variables [52]. This dissertation will focus on the wordlength reduction approach to power minimization. It is well known that power consumption of digital circuits increases with the wordlength. Wordlength reduction, however, generally entails a degradation in

performance of the communication system. It is beneficial, therefore, to understand, quantitatively, the effect of wordlength reduction on system performance. Equally important is the development of procedures to compress data so that it may be stored in reduced-wordlength registers without significant performance degradation. Both of these issues will be investigated in this dissertation.

1.2 Register Length and Power

The importance of wordlength reduction can be illustrated by a simple upper bound on the power consumption of a B-bit register. It is well known that the power consumed by the operation of loading successive time samples of a random real-valued sequence into a binary B-bit register is proportional to the average number of bit flips per unit time in the register [63]. While for a uniform white sequence the average number of bit flips is $B/2$, in general this average can be much less than $B/2$ for a correlated sequence. The reduced activity can be explained by noting that for correlated sequences most significant bits (MSB's) have lower probability of transitioning than least significant bits (LSB's). Several models for the power consumption of register loading have been proposed [63]. The following formula, derived in Appendix A, is an approximate upper bound on the power consumption of a fixed-point, B-bit register into which is loaded a zero-mean, wide-sense stationary Gaussian random sequence:

$$\begin{aligned} P_B &\leq B\eta \cdot \left[1 - \frac{1}{2}\mathrm{erf}\left(\left[2^B\sqrt{2R(0) - 2R(1)}\right]^{-1}\right)\right] \\ &= P_{\max} \end{aligned} \tag{1.1}$$

where η is the power dissipation per bit, which depends on factors such as switching load capacitance and supply voltage, and $R(\tau)$ is the autocorrelation function of the random sequence.

A plot of $P_{\max}$ versus B is given in Figure 1.1 for an AR(1) sequence with real pole located at a_1. For $|a_1| < 0.9$, the power dissipation increases almost linearly as a function

of B. Note from Figure 1.1 that for more highly correlated sequences stored in registers with few bits, the power increase due to increasing bit width (wordlength) is less severe. This is due to the fact that only the LSB's flip with high probability in two successive samples of a highly correlated sequence, while for an uncorrelated uniform sequence, all bits have equal probability (0.5) of flipping. Note also that as the bit width becomes large, all sequences consume approximately the same power. This is attributable to the fact that as the wordlength becomes very large, each new register bit is less significant than all previous bits. Thus, for large wordlengths, a relatively small fraction of the total bits are significant and, although the sequence may be highly correlated, most of the register bits are insignificant and have probability of flipping equal to approximately one half.

Figure 1.1 provides ample support for the wordlength reduction strategy for low power design. Even for extremely correlated sequences – which, though they are not commonplace, do arise in digital communications – significant power savings may be achieved by wordlength reduction according to the bound (1.1). Thus, it is greatly beneficial to understand the complex relationship between wordlength and performance of digital communication systems. Furthermore, an investigation of optimal data compression techniques that may reduce the wordlength required to store data is warranted.

1.3 Components of a Digital Communication System

Figure 1.2 shows the components of a typical digital communication system. The message that is to be transmitted is first compressed by the source encoder so that it may be transmitted in an efficient manner. Next, the channel encoder adds redundant bits to the source-encoded data stream in an effort to protect the message from errors introduced by the channel. Finally, the modulator formats the data in a manner that is suitable for transmission across the physical channel.

At the receiver, the signal coming out of the channel is first demodulated. Next, the channel equalizer attempts to combat distortion arising from possible frequency-selective

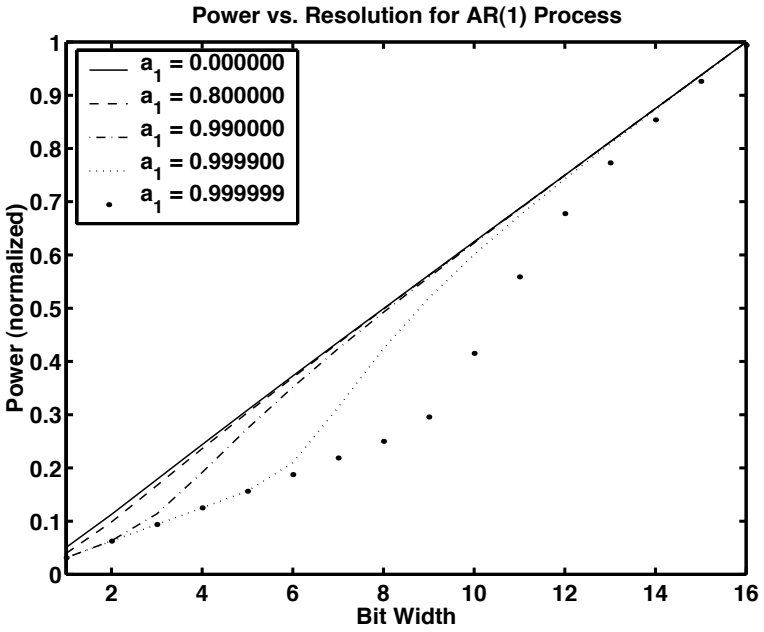

Figure 1.1: Power versus bit width B as function of AR parameter a_1 for loading a B-bit register with successive samples of an AR(1) process. Curves are normalized relative to power P_{16} consumed for a white sequence in a 16-bit register. In the plot, the bit width is reduced from right to left by eliminating LSB's.

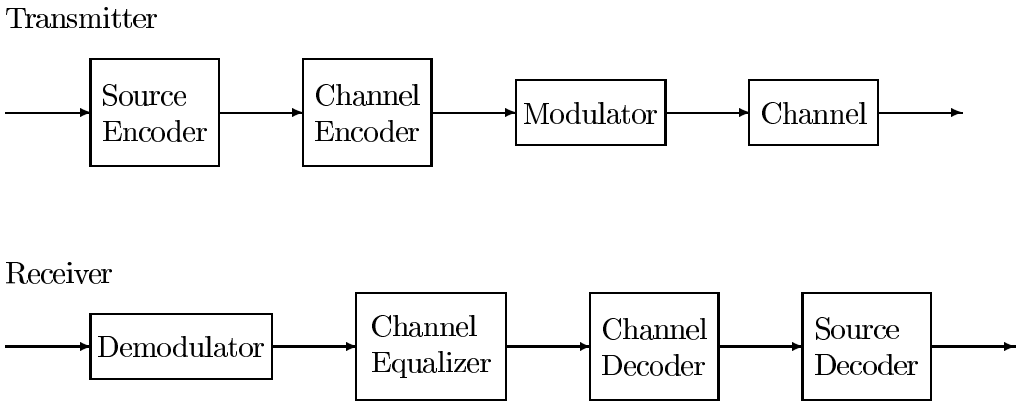

Figure 1.2: Digital communication system.

fading in the channel. The channel decoder then decodes the data stream in an attempt to faithfully reproduce the source-encoded message. Finally, the source decoder inverts the operation of the source encoder.

1.4 Lossless Source Coding and Quantization

Besides the transmitted source in Figure 1.2, virtually all signals in a digital communication system are source coded. These include the input signals to each digital subsystem, as well as the internal variables of these subsystems.

A source encoder takes an input signal, or source, and represents it in a form that requires fewer bits than the original representation. The function of the source decoder is to "reconstruct" the source. The average number of bits needed to represent one source sample is called the *rate* of the source encoder. A good source encoder has as low a rate as possible. Another performance measure of a source encoder is its *distortion*. The distortion is a measure of the difference between the source and reconstruction. Low distortion is always desirable. Some signals, such as the internal variables of a channel equalizer, are encoded, but – as their reconstructions are not required – never decoded.

Source encoders can be characterized as *lossless* or *lossy*. Lossless source encoding involves no loss of information. Equivalently, the distortion of a lossless source encoder is zero and the reconstruction is identical to the source. Lossless encoding can only be performed on sources that are discrete valued.

For continuous-valued sources, the source encoder is necessarily lossy. In lossy source coding, also known as *quantization*, a loss in information between the source and the reconstruction is permitted. A lossy source encoder is a many-to-one mapping. This renders the encoder noninvertible and the reconstruction nonidentical to the source. Consequently, the distortion of a lossy source encoder is greater than zero.

One of the simplest and most common forms of lossy source coding is known as *uniform scalar quantization*. A uniform scalar quantizer operates on scalar, real-valued samples

and has a "stair-step" input/output characteristic. Fixed-point registers perform uniform scalar quantization on their inputs. These registers are extremely popular for storage of variables in digital devices due to their simplicity, speed, and low power consumption. Thus, the majority of the digital subsystems of a communication system use uniform scalar quantization on their internal variables.

The source encoder of a communication system, as shown in Figure 1.2, usually utilizes methods much more sophisticated than uniform scalar quantization. For example, mobile telephones employ source coding algorithms that exploit the correlated nature of speech waveforms. Elaborate techniques are warranted for these source encoders as their rates significantly impact the system bit rate, bandwidth requirements, and power consumption.

1.5 Power Reduction by Quantization

All of the communication system components shown in Figure 1.2 (with the obvious exception of the channel) are power-consuming electronic devices. Power is consumed in these devices by both analog and digital circuitry. To counter the effects of noise introduced by the channel and increase the received signal-to-noise ratio (SNR), the modulator employs an analog power amplifier. The power amplifier usually renders the modulator the single biggest consumer of power in the communication system [70]. A wealth of research has been conducted into the reduction of power consumption in the amplifier. Since no digital component consumes as much power as the amplifier, less research has been performed concerning power reduction in digital circuitry. However, virtually all baseband processing in modern communication systems is done digitally [18] and the benefits of power reduction in digital subsystems could thus have far-reaching impact. Digital systems store their internal data in shift registers and power is consumed by the switching activity of these registers. Examples of digital subsystems include the channel equalizer, as well as parts of the demodulator in Figure 1.2.

1.5.1 A Simple Power Formula

In Section 1.2, a simple relation between register power consumption and wordlength was given. Figure 1.1 indicates that this relationship is approximately linear. Thus, the power consumption of a digital circuit appearing in any of the devices in Figure 1.2 that employs fixed-point registers of wordlength B bits may be expressed as

$$P = \frac{\alpha}{T_{\text{clock}}} B$$

where α is the consumed energy per bit (see Appendix A) and T_{clock} is the clock cycle of the digital circuit. The clock cycle is dependent on the rate at which the system transmits information (the bit rate). Faster rates require shorter clock cycles and it is reasonable to relate the clock cycle to the transmitted bit duration T_B linearly:

$$T_{\text{clock}} = \beta T_B.$$

The purpose of the source encoder in Figure 1.2 is to reduce the bit rate, or increase the bit duration. There are several benefits to such an operation, one of which is the reduction of power consumption throughout the communication system. The source that is fed to the input of the source encoder is modeled as a random sequence. This sequence enters the source encoder at a rate of R_S samples per second. The source encoder outputs a bit stream at rate R bits per second. The rate R_{SC} of the source encoder is defined to be the average number of bits output by the source encoder per input sample. Therefore, the rate of the output bit stream can be written

$$R = R_{SC} R_S.$$

The bit duration T_B is the inverse of the bit rate R. Thus, the power consumption can be written

$$P = \left(\frac{\alpha}{\beta}\right) R_S R_{SC} B. \tag{1.2}$$

In equation (1.2), the source encoder rate R_{SC} and the register wordlength B are design parameters. The equation clearly shows that both quantities must be minimized to achieve low power consumption. As may be expected, reduction of either design parameter has an adverse effect on system performance. The designer must therefore optimize performance subject to power consumption constraints. The fundamental objectives of this dissertation are to develop relations between system performance and the design parameters (wordlength and source code rate) and to develop techniques whereby these relations may be utilized to optimize performance subject to power constraints.

1.5.2 Wordlength Reduction

Many of the digital subsystems shown in Figure 1.2 are fixed-point units that perform signal processing algorithms in efforts to transmit or receive data reliably. The majority of the signal processing algorithms performed by the digital subsystems of a communication system are well understood and can be analyzed elegantly and rigorously. Such analysis, however, is often greatly complicated by introduction of fixed-point effects. It is always true that increased wordlengths correspond to improved performance, but the quantitative relationship between the two is almost never known. The designer, then, often resorts to trial-and-error methods to select the optimal wordlength to meet a given power constraint.

Low-Power LMS Adaptation

In Chapter 2, reduced wordlength effects are investigated for fixed-point adaptive channel equalizers employing the Least Mean Squares (LMS) algorithm. The relationship between the algorithm's performance, as measured by its transient behavior and steady-state mean square error (MSE), is determined. A design procedure is then developed whereby the algorithm can be optimized for minimum steady-state MSE subject to a constraint on its power consumption.

1.5.3 Source Encoding

Power consumption can be reduced in all digital processing units of a communication system by reducing the source encoder rate R_{SC}. The relationship between the rate and distortion of a source encoder has been studied extensively. Indeed an entire branch of information theory is devoted to so-called *rate-distortion theory*. When designing a lossy source encoder and decoder, the objective is to achieve a low rate while minimizing the distortion. The measure of distortion that is to be minimized must be selected carefully and intelligently based upon the intended application of the reconstructed signal. For instance, a speech encoder must be designed so that the reconstructed signal is perceived to be very similar to the source.

Vector Quantization for Distributed Hypothesis Testing

In Chapter 3, the problem of source coding for hypothesis testing applications is studied. Vector quantization is selected as the source coding technique and it is shown that, in some cases, a quantizer can achieve a very low rate while sacrificing no loss in performance of the hypothesis test for which the reconstructed data is intended. For the cases in which such quantizers do not exist, optimal quantizers are derived, for a given rate, that minimize the loss in area under an analog to the receiver operating characteristic (ROC) curve of the Neyman-Pearson hypothesis test.

CHAPTER 2

Low-Power LMS Adaptation

2.1 Introduction

The Least Mean Squares (LMS) algorithm, introduced by Widrow and Hoff [71, 74], finds itself in many components of a digital communication system. The relative simplicity of the algorithm renders it ideal for adaptive channel equalization and estimation [67], adaptive beamforming antenna arrays [72], and adaptive interference cancellation systems [25]. The near ubiquity of the algorithm in digital communications along with the ever-increasing demand for low-power communication systems warrants an understanding of the relationship between its performance and power consumption. Indeed, a great deal of research has been done on the LMS algorithm's behavior over the years [65], but very little of this work has focused on the power-performance relationship.

Insight into the LMS algorithm's power-performance relationship is of paramount importance for low-power LMS implementation. Without such insight, the system designer must resort to exhaustive trial-and-error methods to achieve low power consumption. Conversely, equipped with an understanding of LMS power-performance behavior, the designer may optimize an LMS system's performance given a power constraint, or vice-versa.

To illustrate more specifically the motivation for low-power LMS implementation, consider a battery-powered wireless receiver, for which the adaptive LMS equalization function consumes a significant portion of the total battery power. As an example, the SINCGARS combat radio used by the United States Army consumes on the average 7 Watts in receive

mode, of which more than 1 Watt (over 14% of total power) is consumed by the channel equalizer [44]. Clearly then, the equalization function is a prime target for power reduction in these handsets.

Many digital hardware design strategies have been proposed for power reduction including: reduction of supply voltage, reduction of clock speed and data rate, parallelization and pipelining of operations, using sign-magnitude arithmetic, and differential encoding of data [16, 51]. Another technique is the reduction of the number of bits (wordlength) used to represent the data and control variables in the digital circuit [52]. The wordlength reduction strategy is very highly leveraged since it reduces the power dissipation everywhere in the data and control flow paths. This strategy is also very versatile since it can be applied to any hardware architecture and can be easily adjusted in real time by dynamically disabling bus lines and register bits.

As mentioned previously, power reduction – by means of wordlength reduction or any other technique – must be carried out with foresight regarding algorithm performance. For example, LMS wordlength reduction generally entails a degradation in algorithm performance as measured by adaptive algorithm convergence rate and steady-state mean square error (MSE). This chapter provides an analysis of MSE degradation versus power reduction, by means of coefficient and data wordlength reduction, for the LMS algorithm. Additionally, optimization of the algorithm under a power constraint is described and illustrated for the case of channel equalization.

2.1.1 The LMS Algorithm

The LMS algorithm iteratively adapts the coefficients of an FIR filter in an attempt to minimize the mean-square value of an error signal formed by subtracting the filtered output from a primary, or desired, signal. In a digital implementation, all signals and coefficients are stored with finite precision. Most finite-precision implementations use fixed-point registers. The finite-precision LMS algorithm can be viewed as an infinite-precision LMS algorithm

implemented with separate quantizers in the data paths and in the filter coefficient paths.

2.1.2 Overview of Previous Work

Many aspects of finite-precision LMS implementations have been studied. The research has focused both on analysis of finite-precision effects (quantization of data and coefficients) as well as optimal finite-precision implementation strategies. Very little research has focused explicitly on power-performance relationships.

In [22] it was shown that uniform scalar quantization of the LMS error signal is close to optimal for the case of Gaussian inputs. All fixed-point implementations utilize uniform scalar quantization and this result thus provides motivation for optimization of fixed-point implementations. To obtain significant power reduction, it may be beneficial to apply different quantizer resolutions during the transient acquisition phase and the steady-state tracking phase of the algorithm. For example, "dynamic precision tuning" of filter coefficients has been studied for a low-power, 128-tap adaptive modem in [49]. In this chapter, however, we focus on steady-state analysis of finite-precision effects when the coefficient and data wordlengths are distinct but fixed over time.

Since quantization is a noninvertible and nondifferentiable operation, an exact statistical analysis of the finite-precision LMS algorithm is intractable. Several methods of approximate analysis have been proposed. Alexander [2] and Caraiscos and Liu [15] performed second-order statistical analyses of the real-valued, finite-precision LMS algorithm under the assumption of a linear white noise model for the quantization error. This type of analysis will be referred to as the *standard analysis*. Under a similar assumption, Cioffi [20] performed an analysis of the finite-precision LMS adaptive echo canceler, which was later extended to block LMS by Cioffi and Ho [19]. It has been observed that when the roundoff error is not too large, the standard analysis gives useful and accurate performance predictions [2, 17, 19, 20, 23]. However, the standard analysis does not account for the slowdown phenomenon, a feature unique to some finite-precision implementations that oc-

curs when the LMS weights do not get regularly updated due to excessive roundoff error. For the special case of the LMS algorithm implemented with finite-precision coefficients and infinite-precision data, Bermudez and Bershad [6, 7, 8] presented a recursive learning curve approximation that gives more accurate predictions of the steady-state MSE than the standard analysis, the standard analysis giving increasingly poor predictions as coefficient resolution decreases.

2.1.3 Overview of Contribution

In this chapter, we use the standard analysis methodology to derive expressions for optimal bit allocation factors for finite-precision LMS implementations that do not experience slowdown under total data + coefficient wordlength constraints and total power constraints. As a check against the presence of slowdown, we propose a simple and accurate criterion that predicts that slowdown will not occur when the number of coefficient bits exceeds the number of data bits by a quantity specified by the criterion. This conclusion is based on extensive numerical simulations of finite-precision LMS for channel equalization and system identification, as well as analytical studies. While the results are directly applicable to the generic finite-precision LMS algorithm in a variety of applications, for concreteness we concentrate on the case of channel equalization with training.

An outline of the chapter is as follows. We begin with a preliminary overview of the infinite-precision LMS algorithm along with a mathematical description of our model for the finite-precision algorithm in Sections 2.2.1 and 2.2.2. In Section 2.2.3, we derive a formula for the iteration power of the fixed-point, power-of-two step-size LMS algorithm under fixed-point complex arithmetic, cyclic updating of the data stack, and table lookup implementation of multiplication. Next, in Section 2.2.4, using the averaged system techniques of Solo and Kong [65], we obtain analytical expressions for the increase in steady-state mean square error due to finite precision in the absence of the slowdown phenomenon which generalize the formulas of Caraiscos and Liu [15] to the case of complex data and coefficients.

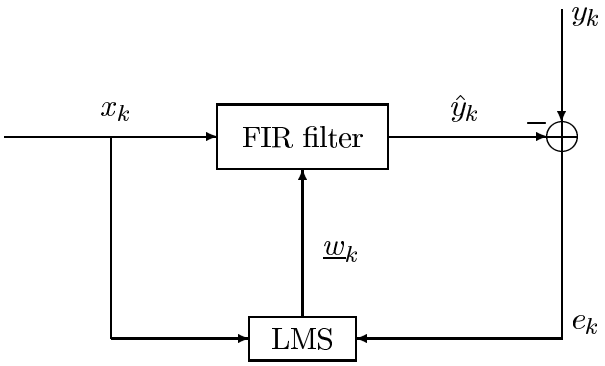

Figure 2.1: Infinite-precision LMS algorithm.

We also present a simple constraint on the data wordlength, coefficient wordlength, and adaptation step size μ that prevents the slowdown phenomenon. We then derive a pair of optimal bit allocation factors, in Section 2.3, that minimize the amount of increase in MSE subject to two constraints: 1) a total bit width constraint (total bit budget) and 2) a total power consumption constraint (total power budget). We conclude that assigning more bits to coefficients than data is necessary to avoid the slowdown phenomenon and in general gives better steady-state MSE performance under a total power budget constraint. This conclusion mirrors a similar result reported in [23] for the finite-precision sign LMS algorithm. Finally, in Section 2.4, we verify the accuracy of our theoretical predictions for the case of LMS equalization of a single-pole IIR channel.

2.2 Finite-Precision LMS Adaptation

2.2.1 Infinite-Precision LMS Algorithm

Figure 2.1 shows a block diagram of the infinite-precision LMS algorithm. Here y_k is a complex training signal, x_k is the complex FIR filter input, and $\hat{y}_k = \underline{w}_k^H \underline{x}_k$ is a linear estimate of y_k given the p samples $\underline{x}_k = [x_k, \ldots, x_{k-p+1}]^T$ and filter coefficients $\underline{w}_k = [w_{0,k}, \ldots, w_{p-1,k}]^H$. Note that the FIR filter has p taps. The signal $e_k = y_k - \hat{y}_k$ is the error signal whose mean-square value the algorithm attempts to minimize. The vehicle by which this minimization is accomplished is a recursive coefficient update, first derived

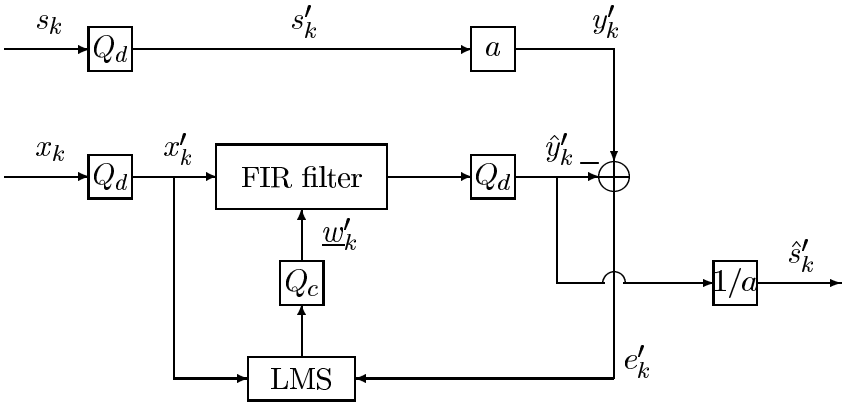

Figure 2.2: Finite-precision LMS algorithm.

by Widrow and Hoff [71, 73, 74]. The recursion seeks the minimum of the MSE surface $E[|e_k|^2] = E[|y_k - \underline{w}_k^H \underline{x}_k|^2]$ by means of gradient descent with an estimated gradient. The recursion is given by

$$\begin{aligned} \underline{w}_{k+1} &= \underline{w}_k + \mu \underline{x}_k e_k^* \\ e_k &= y_k - \hat{y}_k \end{aligned} \tag{2.1}$$

where μ is the adaptive gain parameter that controls the convergence properties of the algorithm and * denotes complex conjugation.

2.2.2 Finite-Precision LMS Algorithm

Figure 2.2 shows the LMS algorithm described in the previous section with the addition of two different quantizers, denoted Q_d and Q_c, applied to the complex data and to the complex filter coefficients used by the algorithm. In addition, the training sequence is now denoted s_k and is scaled by the factor a. The quantizers Q_d and Q_c are assumed to be uniform scalar quantizers, corresponding to fixed-point arithmetic. In fixed point, all signal magnitudes must lie below a threshold (which we assume is unity), and thus care must be taken to prevent register overflow. The scaling factor a is used to prevent overflow of the quantized weight vector and is usually implemented by a right shift of one or more bits.

The quantizers Q_d and Q_c are allocated B_d bits plus sign and B_c bits plus sign, respectively, to the real and imaginary parts of their inputs. We assume that the input sequences

have been scaled to lie between -1 and $+1$. As in [15], we assume that the effect of the operators Q_d and Q_c is to add to their inputs complex white noises of variance $2\sigma_d^2 = (1/6)2^{-2B_d}$ for Q_d and $2\sigma_c^2 = (1/6)2^{-2B_d}$ for Q_c.

The finite-precision LMS algorithm implements the recursion (2.1) with quantizers in all data paths as shown in Figure 2.2. The finite-precision recursion can be written as

$$\underline{w}'_{k+1} = \underline{w}'_k + Q_c\left(\mu\, \underline{x}'_k e'^*_k\right) \tag{2.2}$$

where

$$e'_k = y'_k - Q_d\left(\underline{w}'^H_k \underline{x}'_k\right) \tag{2.3}$$

is the quantized error signal. We assume that the gain parameter μ is chosen to be a power of two, thereby enabling the multiplication by μ to be performed by right shifts. We have used primed symbols to represent quantized values. For example, $\underline{x}'_k = Q_d(\underline{x}_k)$ is simply the quantized value of the FIR filter input. Similarly, $s'_k = Q_d(s_k)$. For the coefficients, $\underline{w}'_k$ represents the value of the weight vector used in the finite-precision algorithm at iteration k. Note that this should approximate $\underline{w}_k$, the weight vector at iteration k of the infinite-precision algorithm, if the input signals are the same. We assume that the computation of the inner product in (2.3) is accomplished by quantizing the partial sums. Therefore,

$$Q_d(\underline{w}'^H_k \underline{x}'_k) = \sum_{i=0}^{p-1} Q_d(w'_{i,k} x'_{k-i})$$

and the noise added to the inner product has variance $2p\sigma_d^2$.

2.2.3 Power Consumption of LMS Algorithm

The total iteration power of the finite-precision LMS algorithm is determined by power dissipation of shift, add, multiply, memory load, and memory store operations. This depends on the specific circuit implementation of the FIR filter and control circuitry. The bulk of the power dissipation usually comes from add and multiply operations. In Appendix B, we derive the following formula for iteration power of LMS considering only adds and multiplies

under the assumption that $\mu = 2^{-q}$ for $q \in \{0, \ldots, B_c - 1\}$:

$$P_T = 4p[(2B_d + B_c + 2)\eta_a + (3B_d + B_c)\eta_t]. \tag{2.4}$$

This expression is linear in the number of bits B_d and B_c and assumes fixed-point complex arithmetic, cyclic updating of the data stack $[x_k, \ldots, x_{k-p+1}]^T$ by overwriting x_{k-p+1} with x_{k+1}, and multiplication using table lookup as opposed to adding partial products. In (2.4), we have defined generic power coefficients η_a and η_t representing power consumption per add per bit and power consumption per table lookup operation per bit, respectively.

2.2.4 Statistical Performance of Finite-Precision LMS Algorithm

The performance of the LMS adaptive algorithm is typically characterized by two quantities: the speed of convergence and the excess MSE [65, 67, 74]. We assume that s_k and x_k are both wide-sense stationary random sequences. Caraiscos and Liu [15] analyzed the effects of finite wordlength on (real-valued) LMS filter performance under the following high resolution assumptions:

> The quantization errors of quantizers Q_d and Q_c are zero mean, white, with variances $2\sigma_d^2$ and $2\sigma_c^2$, respectively. Furthermore, these errors are independent of the quantizer inputs. The process x_k is circular Gaussian. The step size μ is greater than 0 and B_c, $B_d \geq 1$. Finally, the sequences s_k and x_k have been scaled so that they do not overflow. (2.5)

In the remainder of this section, we derive the mean convergence rate, steady-state weight-error covariance, and excess mean square error for complex-valued, finite-precision LMS under the above assumptions. As will be shown below, use of these assumptions yields accurate error predictions when slowdown does not occur. Later, we derive constraints on B_c, B_d, and μ that prevent the occurance of slowdown.

Mean Convergence

Define the $p \times p$ covariance matrix $R_{x'} = E[\underline{x}'_k \underline{x}'^H_k]$ of the quantized data and let this matrix have real, non-negative eigenvalues $\{\lambda'_i\}_{i=1}^p$. Further, define the cross correlation vector $R_{x'y'} = E[\underline{x}'_k y'^*_k]$. Assume the gain parameter μ satisfies the condition

$$0 < |1 - \mu\lambda'_i| < 1, \qquad i = 1, \ldots, p. \tag{2.6}$$

In Appendix C.1, we show that the filter coefficients of the finite-precision LMS algorithm converge in the mean to a set of optimal weights $\underline{w}'^o$ called the (finite-precision) *Wiener weights*:

$$\lim_{k\to\infty} E[\underline{w}'_k] = \underline{w}'^o \tag{2.7}$$

where

$$\underline{w}'^o = R_{x'}^{-1} R_{x'y'}.$$

When the finite-precision LMS algorithm converges, its MSE trajectory, or *learning curve*, converges as a decaying exponential with the $1/e$ time constant of the slowest mode equal to $\tau_{3dB} = 1/(-\max_i \ln(|1 - \mu\lambda'_i|))$, called the adaptation time constant. Note that the speed of convergence generally increases as μ increases.

Steady-State Weight-Error Covariance

In Appendix C.2, we derive an expression for the steady-state quantized weight-error covariance. Let $P_k = E[(\underline{w}'_k - \underline{w}_k)(\underline{w}'_k - \underline{w}_k)^H]$ where $\underline{w}_k$ is the weight vector at iteration k of the infinite-precision LMS algorithm. Then, assuming P_k converges as $k \to \infty$, it converges to the steady-state covariance matrix given by

$$P = \mu(p+1)\sigma_d^2 I + \frac{1}{\mu}\sigma_c^2 R_{x'}^{-1}. \tag{2.8}$$

Excess Mean Square Error

Define the infinite-precision covariance matrix $R_x = E[\underline{x}_k \underline{x}_k^H]$ and cross correlation vector $R_{xy} = E[\underline{x}_k y_k^*]$. Under the assumptions (2.5), the following asymptotic expression

for the steady-state mean square error is derived for small μ in Appendix C.2:

$$\xi = E[|s_k - \hat{s}'_k|^2] = \xi_{\min} + \xi_{\text{excess}} + \xi_q \tag{2.9}$$

where

$$\xi_{\min} = \frac{1}{a^2}\left(\sigma_y^2 - R_{xy}^H R_x^{-1} R_{xy}\right)$$

is the optimal mean square error with the infinite-precision Wiener weights $\underline{w}^o$,

$$\xi_{\text{excess}} = \frac{1}{2}\mu \text{tr}(R_x)\xi_{\min}$$

is the excess MSE due to misadjustment, and

$$\xi_q = \alpha_c \, 2^{-2B_c} + \alpha_d \, 2^{-2B_d} \tag{2.10}$$

where

$$\alpha_c = \frac{p}{12\mu a^2}, \qquad \alpha_d = \frac{||\underline{w}^o||^2 + p}{6a^2}$$

and $\underline{w}^o = R_x^{-1} R_{xy}$ is the optimal Wiener weight vector for the standard, infinite-precision, complex LMS algorithm.

The term ξ_q is the excess MSE due to quantization of the data and filter coefficients. The first term in the expression (2.10) is the excess MSE due only to quantization of the filter coefficients while the second term represents the MSE due to quantization of the data. Note that for small μ the term α_c dominates the excess MSE due to quantization unless B_c is made large. This implies that for small step sizes a high resolution is required for the filter coefficients. Also worth noting is that ξ_q increases in p at a linear rate, decreases in μ at an inverse linear rate, and decreases in B_d and B_c at an exponential rate. Therefore, the total number of bits allocated gives more leverage over excess MSE than any other of the design parameters.

With these relations, the increase in MSE due to finite precision ξ_q can be plotted as a function of B_d and B_c. A plot of the increase in MSE is given in Figures 2.3 and 2.4 for

the case of white s_k and for x_k generated by passing s_k through a single-pole IIR filter with pole at $a_1 = 0.8$ and with LMS gain parameter $\mu = 1/4$, scaling factor $a = 1/8$, and $p = 2$ taps. The vertical plane on the surface plot and the thick line on the contour plot divide the B_c, B_d plane into two regions. In the next section, we will show that the value ξ_q is valid only in the region to the left of the vertical plane in Figure 2.3 (right of the thick line in Figure 2.4) while the other region is the slowdown region.

Predicting Onset of Slowdown

The slowdown phenomenon occurs when one or more components of the input to the quantizer Q_c in (2.2) falls below the LSB of Q_c [6, 7, 15, 28]. The corresponding onset of slowdown can be defined as the minimum integer $k > 0$ such that

$$\left|\mathrm{Re}\{\mu x'_{k-i} e'^{*}_{k}\}\right| < \frac{\Delta_c}{2} \quad \text{or} \quad \left|\mathrm{Im}\{\mu x'_{k-i} e'^{*}_{k}\}\right| < \frac{\Delta_c}{2} \tag{2.11}$$

for some $i \in \{0, \ldots, p-1\}$. Here $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ denote real and imaginary parts and $\Delta_c = 2^{-B_c}$ is the granularity of the coefficient quantizer. During the initial stages of adaptation, before onset of slowdown, the finite-precision LMS algorithm's behavior does not differ significantly from that of the infinite-precision algorithm [6, 7]. When condition (2.11) is met at iteration k, at least one of the complex components of the ith element of the vector $\underline{w}'_k$ does not get updated and the finite-precision algorithm's learning curve begins to diverge from that of the infinite-precision algorithm. To determine the point at which this divergence occurs, called the *slowdown point*, consider the probability

$$\max_{i=0,\ldots,p-1} P\left(\left|\mathrm{Re}\{\mu x'_{k-i} e'^{*}_{k}\}\right| < \frac{\Delta_c}{2}\right) > 1 - \epsilon \tag{2.12}$$

for $0 < \epsilon \ll 1$. It is easily shown that the left-hand side of (2.12) is a lower bound on the joint probability of the event (2.11). Furthermore, the left-hand side completely specifies this joint probability when $\underline{x}_k$ is independent and identically distributed (i.i.d.) with i.i.d. real and imaginary parts. Thus, if for some k (2.12) is satisfied, then with probability at least $1 - \epsilon$ the iteration k is a slowdown point. Furthermore, we can assert that if no

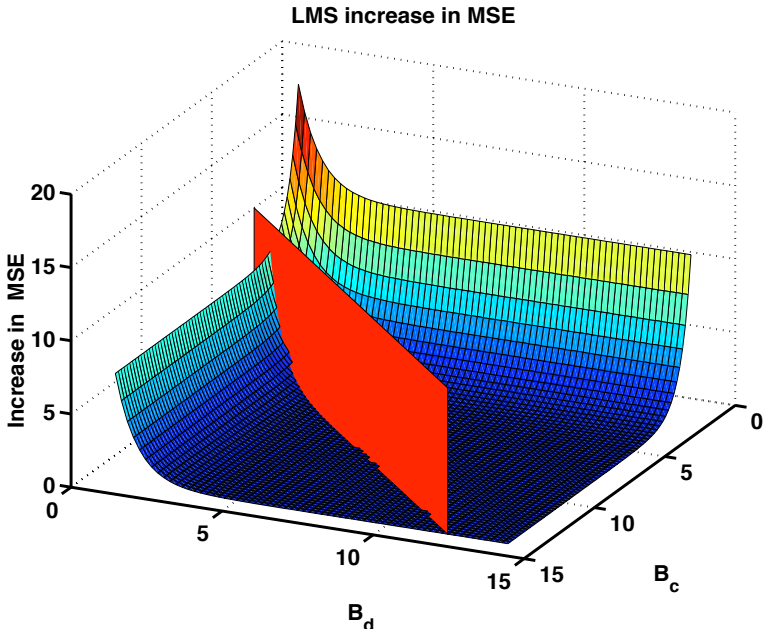

Figure 2.3: Surface plot of excess MSE due to finite precision as a function of B_d and B_c for single-pole IIR channel.

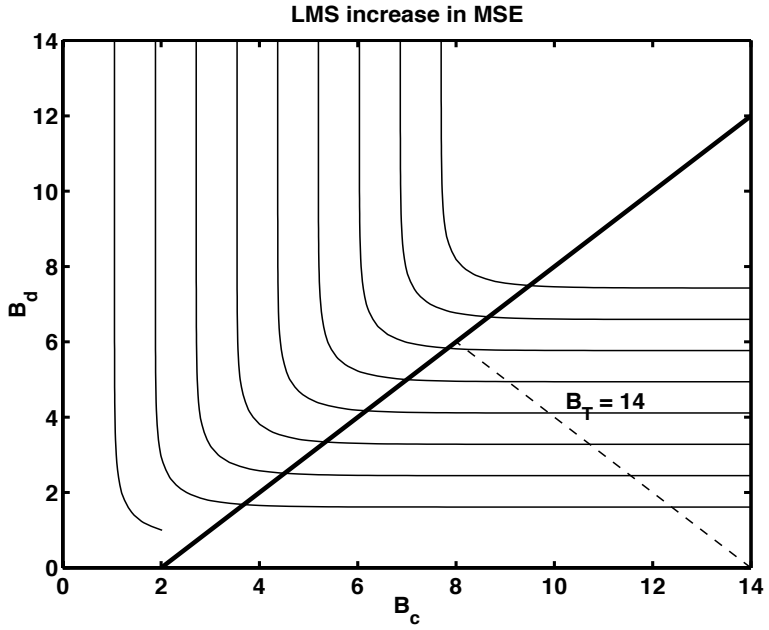

Figure 2.4: Contour plot of excess MSE due to finite precision as a function of B_d and B_c for single-pole IIR channel.

iteration k satisfies (2.12) before steady-state is reached, then slowdown will not occur with probability at least $1-\epsilon$. These observations are the basis for our approach to prevention of slowdown: determine the minimum value of Δ_c that ensures that (2.12) can only be satisfied after steady-state is reached. This value of Δ_c will then specify a range of admissible values for B_c, B_d, and μ for which slowdown does not occur. The choice of ϵ will be discussed in Section 2.4.

The condition (2.12) will be evaluated by invoking the fact that the finite-precision algorithm's behavior does not differ significantly from that of the infinite-precision algorithm before slowdown and therefore (2.12) can be replaced by

$$\max_{i=0,\ldots,p-1} P\left(|\mathrm{Re}\{\mu x_{k-i} e_k^*\}| < \frac{\Delta_c}{2}\right) > 1-\epsilon \tag{2.13}$$

where $e_k = y_k - \hat{y}_k$ is the error signal in the infinite-precision algorithm. Now define $g_k = x_{k-i}e_k^*$ and make the simplifying assumption that g_k is approximately circular Gaussian with mean zero and variance $\sigma_x^2\sigma_{e,k}^2$ where $\sigma_x^2 = E[|x_k|^2]$ and $\sigma_{e,k}^2 = E[|e_k|^2]$ is the mean square error of the infinite-precision algorithm. Before slowdown we have $\sigma_{e,k}^2 \approx \xi_k' = E[|e_k'|^2]$. Then (2.13) predicts that slowdown will begin (with probability at least $1-\epsilon$) when k satisfies

$$\mathrm{erf}\left(\frac{\Delta_c}{2\mu\sigma_x\sqrt{\xi_k'}}\right) = 1-\epsilon.$$

This is equivalent to

$$\begin{aligned} \xi_k' &= \frac{2^{-2(B_c+1-q)}}{\sigma_x^2[\mathrm{erf}^{-1}(1-\epsilon)]^2} \\ &= \xi_{\mathrm{slow}}' \end{aligned} \tag{2.14}$$

where $q = -\log_2 \mu$.

Now, from the definition of e_k' and the results of the previous section, we have $\xi_\infty' = a^2\xi + 2\sigma_d^2$ in the absence of slowdown. Therefore, assuming no slowdown, for a fixed B_d

$$\xi_\infty' > a^2(\xi_{\min} + \xi_{\mathrm{excess}} + \xi_q|_{B_c=\infty}) + 2\sigma_d^2$$

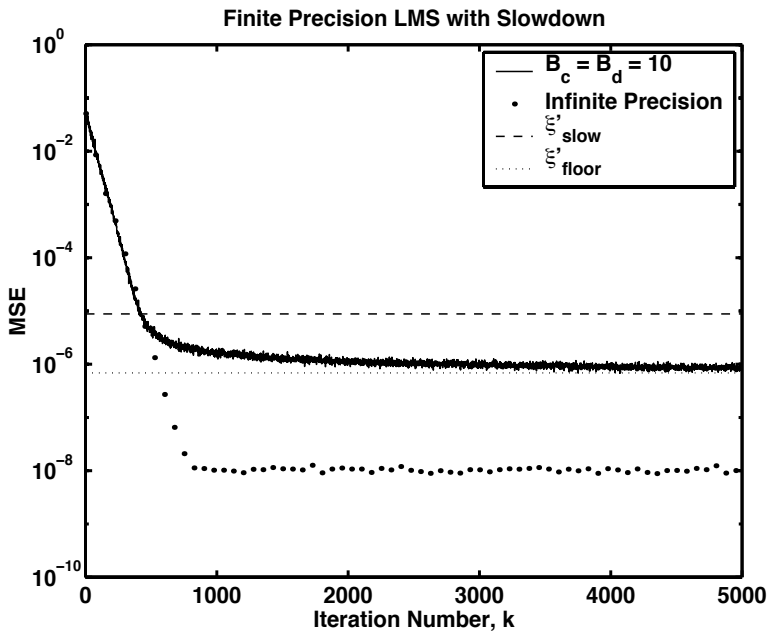

Figure 2.5: Example of slowdown.

and assuming μ is small

$$\begin{aligned} \xi'_\infty &> a^2\xi_{\min} + 2\sigma_d^2(\|\underline{w}^o\|^2 + p + 1) \\ &= \xi'_{\text{floor}}. \end{aligned} \tag{2.15}$$

Figure 2.5 shows a sample learning curve of a finite-precision LMS channel equalizer that experiences slowdown along with the values ξ'_{slow} and ξ'_{floor}.

Again, since the finite and infinite-precision algorithms agree closely before onset of slowdown, slowdown can be prevented by choosing B_c such that

$$\xi'_{\text{slow}} < \xi'_{\text{floor}}. \tag{2.16}$$

Note from the derivation of ξ'_{slow} that this quantity is the minimum value of ξ'_k achievable before the onset of slowdown with infinite-precision data. The MSE ξ'_{floor} is the steady-state MSE (in the absence of slowdown) with infinite-precision coefficients. Thus, if B_c and B_d are chosen such that the inequality (2.16) is satisfied, then with high probability the slowdown MSE ξ'_{slow} will not be reached and slowdown will not occur. Inequality (2.16) is

equivalent to

$$2^{-2(B_c+1-q)} < \sigma_x^2[\text{erf}^{-1}(1-\epsilon)]^2[a^2\xi_{\min} + 2\sigma_d^2(\|\underline{w}^o\|^2 + p + 1)].$$

If a lower bound ψ on $\|\underline{w}^o\|^2$ is available, we can use this lower bound and the lower bound $\xi_{\min} \geq 0$ to obtain the following sufficient condition in B_c, B_d for no slowdown:

$$B_c > B_d + \nu \tag{2.17}$$

where

$$\nu = q - 1 - \frac{1}{2}\log_2\left(\frac{\sigma_x^2}{6}(\psi + p + 1)[\text{erf}^{-1}(1-\epsilon)]^2\right). \tag{2.18}$$

2.3 Optimal Bit Allocation Strategies

We present expressions for the optimum allocation of bits to data versus filter coefficients under two constraints: total number of bits and total power consumption.

Assume that there are a total of B_T bits plus two sign bits that are available to allocate between data and coefficients, i.e. $B_T = B_d + B_c$. Further, define the data bit allocation factor $\rho = B_d/B_T$. Then we have the obvious relations

$$B_d = \rho B_T, \qquad\qquad B_c = (1-\rho)B_T. \tag{2.19}$$

2.3.1 Total Bit Budget Constraint

Under a constraint on B_T the objective is to minimize the increase ξ_q in MSE with respect to ρ. Graphically, this is the same as minimizing ξ_q along the diagonal line $B_T = B_d + B_c$ of slope -1 in the B_c, B_d plane shown on Figure 2.4. Note that the region to the left of the thick line in Figure 2.4 must be avoided. It is shown in Appendix D.1 that ξ_q is convex as a function of ρ with a single minimum occurring at the point $\rho = \rho^*$, where

$$\begin{aligned}\rho^* &= \frac{1}{4B_T}\log_2\left(\frac{\alpha_d}{\alpha_c}\right) + \frac{1}{2} \\ &= \frac{1}{4B_T}\log_2\left(2\mu\frac{\|\underline{w}^o\|^2 + p}{p}\right) + \frac{1}{2}.\end{aligned} \tag{2.20}$$

Observe that as B_T increases, ρ^* approaches the standard textbook allocation of 1/2. However, for low B_T the standard allocation is suboptimal.

Note that condition (2.17) is equivalent to

$$\rho < \frac{1}{2} - \frac{\nu}{2B_T} = \rho_{\text{slow}}. \tag{2.21}$$

Therefore, since ξ_q is convex as a function of ρ, the optimal value of ρ is

$$\rho^B = \min\{\rho^*, \rho_{\text{slow}}\}.$$

2.3.2 Total Power Budget Constraint

Under a constraint on total power budget P_T, we can use (2.4) and (2.19) to re-express the total combined number of bits B_T as a function of ρ and P_T:

$$B_T = \frac{P_T - 8p\eta_a}{4p[\rho(\eta_a + 2\eta_t) + \eta_a + \eta_t]}. \tag{2.22}$$

In Appendix D.2 we show that for $B_T \geq 2$, ξ_q is once again a convex function of ρ with unique minimum at $\rho = \rho^{**}$, where

$$\rho^{**} = \frac{P_T - 8p\eta_a + 2p(\eta_a + \eta_t)\log_2\left[\frac{\alpha_d}{\alpha_c} \cdot \frac{\eta_a + \eta_t}{2\eta_a + 3\eta_t}\right]}{2(P_T - 8p\eta_a) - 2p(\eta_a + 2\eta_t)\log_2\left[\frac{\alpha_d}{\alpha_c} \cdot \frac{\eta_a + \eta_t}{2\eta_a + 3\eta_t}\right]}. \tag{2.23}$$

Similar to the bit budget constraint, ρ^{**} converges to the standard 1/2 allocation as P_T becomes large while for low P_T the standard allocation is suboptimal.

Again applying constraint (2.17) the optimal allocation is

$$\rho^P = \min\{\rho^{**}, \rho_{\text{slow}}\}$$

where ρ_{slow} is calculated using (2.21) and (2.22).

2.4 Numerical Example

Here we briefly consider LMS equalization of an IIR channel with a single pole at $a_1 = 0.5$, a Gaussian transmitted signal s_k with variance $E[|s_k|^2] = \sigma_s^2 = 0.06$, a Gaussian-noise-corrupted training sequence $y_k = s_k + n_k$ with $E[|n_k|^2] = \sigma_n^2 = 10^{-8}$, and a two-tap

a_1	$B_c = B_d = 8$	$B_c = B_d = 10$
0.1	-6%	-4%
0.2	-8%	-8%
0.3	-4%	-6%
0.4	-7%	-8%
0.5	-8%	0%
0.6	-6%	-6%
0.7	-3%	-4%

Table 2.1: Percent difference between theoretical and measured slowdown points for finite-precision LMS channel equalizer with $\sigma_s^2 = 0.06$, $\sigma_n^2 = 10^{-8}$, $\mu = 1/4$, IIR channel with single pole at a_1, and $\epsilon = 10^{-3}$.

LMS filter with gain coefficient $\mu = 1/4$. With these parameters, automatic gain control (AGC) is unnecessary and the scale factor a can be set to unity as the probability of register overflow is small. The channel is a rather severe exponential memory channel with intersymbol interference (ISI) extending over approximately five data samples.

The first step in designing the finite-precision equalizer is to choose an appropriate confidence level ϵ in (2.12) that will give an accurate prediction of the actual onset point k of slowdown. Although an analytical solution is not yet available, we have strong experimental evidence [31, 32] that ϵ is strongly dependent on the number of taps p and adaptive gain parameter μ, but only weakly dependent on other parameters. In particular, as p becomes large slowdown is not likely to occur if only a few taps remain unchanged. To reflect this phenomenon, ϵ must decrease as p increases. The value of ϵ was chosen for the class of single-pole IIR channels by simulating a representative IIR channel and finite-precision, two-tap LMS algorithm with $B_d = B_c = 10$. By visual inspection the slowdown MSE ξ'_{slow} was determined and (2.14) was used to obtain $\epsilon = 10^{-3}$. Having chosen ϵ, we next calculate $\nu = 1.85$ from (2.18) using $\psi = ||\underline{w}^o||^2 = 1.25$. Thus, (2.17) becomes $B_c \geq B_d + 2$.

It was experimentally verified that the value $\epsilon = 10^{-3}$ accurately predicted the actual slowdown onset point for the entire range of two-tap LMS implementations and single-pole IIR channels. Table 2.1 shows the percent difference between the iteration number at which the MSE ξ'_{slow} is achieved and the measured slowdown point for several values

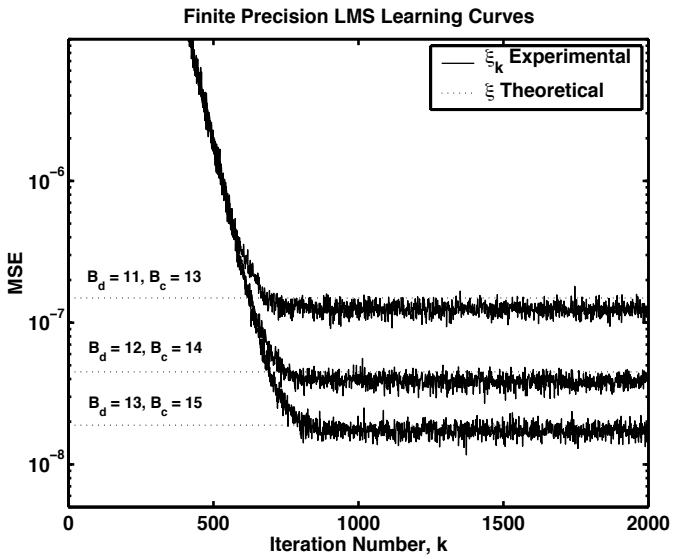

Figure 2.6: Learning curves of LMS filters equalizing IIR channel with Gaussian input without slowdown.

of the channel pole a_1. The table shows that using $\epsilon = 10^{-3}$ yields accurate slowdown point predictions for different wordlengths and channels. Note that since the differences are all non-positive, the estimate ξ'_{slow} and the constraint $B_c > B_d + \nu$ are both conservative. Figure 2.6 shows a representative sampling of the learning curves for different finite-precision equalizers satisfying (2.17) along with the predicted value of the steady-state MSE given by (2.9). Again, choice of $\epsilon = 10^{-3}$ has yielded a constraint preventing slowdown for various wordlengths.

To determine the power coefficients η_a and η_t for this example, several adders and multipliers with varying bit-width were simulated using the Epoch CAD package. The energy consumption of each adder and multiplier was determined and by using linear least squares fits to this data, the adder energy per bit and multiplier energy per bit were obtained. These figures were then multiplied by the assumed clock cycle of 50 MHz to give the coefficients $\eta_a \approx 1.4$ mW and $\eta_t \approx 6.8$ mW. The simulated multipliers were shift-add (partial product) multipliers. Although our analysis considers table lookup multipliers, it is clear from Figure B.1 that the chosen value of η_t gives a good power approximation for either multiplier.

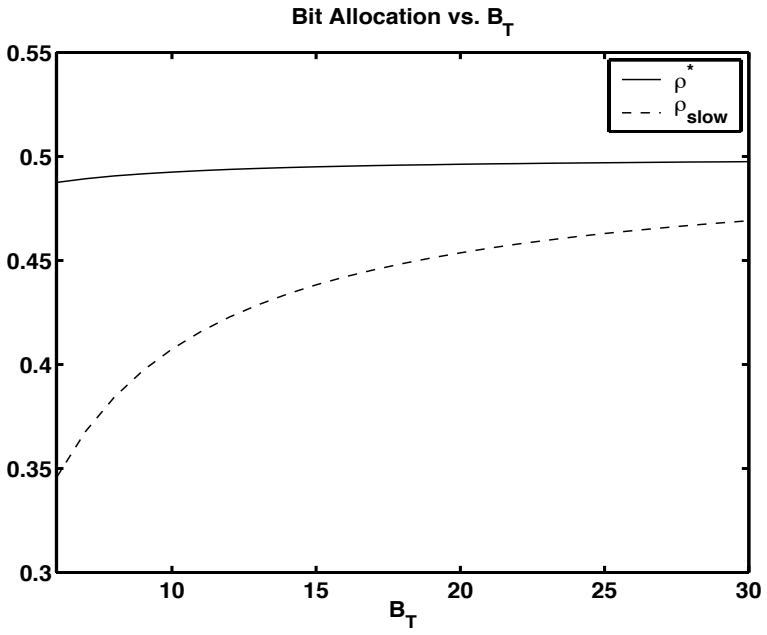

Figure 2.7: IIR channel with Gaussian input: data bit allocation factors under B_T constraint as functions of B_T.

Figure 2.7 shows the B_T-constrained data bit allocation factor ρ^* as a function of B_T as well as ρ_{slow}, the maximum allowable allocation satisfying (2.17). It is clear that for all B_T, $\rho_{\text{slow}} < \rho^*$ and therefore the optimal allocation factor is $\rho^B = \rho_{\text{slow}}$. Figure 2.8 shows the P_T-constrained data bit allocation factor ρ^{**} as a function of P_T as well ρ_{slow}. Again note that for all P_T of interest, $\rho_{\text{slow}} < \rho^{**}$ and the optimal allocation factor is $\rho^P = \rho_{\text{slow}}$. Also shown on this figure are two suboptimal allocations, each satisfying the no-slowdown constraint. Finally, Figure 2.9 shows the resultant MSE ξ as a function of P_T using the optimal allocation $\rho = \rho^P = \rho_{\text{slow}}$ as well as the suboptimal allocations plotted on a log scale.

While these results show that the bit allocation ρ_{slow} is always optimal for this example, it should be noted that this is not always the case. For example, consider the design of a finite-precision LMS equalizer with the following parameters: $p = 32$, $\|\underline{w}^o\|^2 = 1$, $a = 1$, $\mu = 1/8$, $\sigma_s^2 = 0.06$, $\sigma_x^2 = 0.09$. Using $\epsilon = 10^{-5}$ (the correct value for these parameters), Figures 2.10 and 2.11 show that for this case the optimal bit allocation factors are ρ^* and

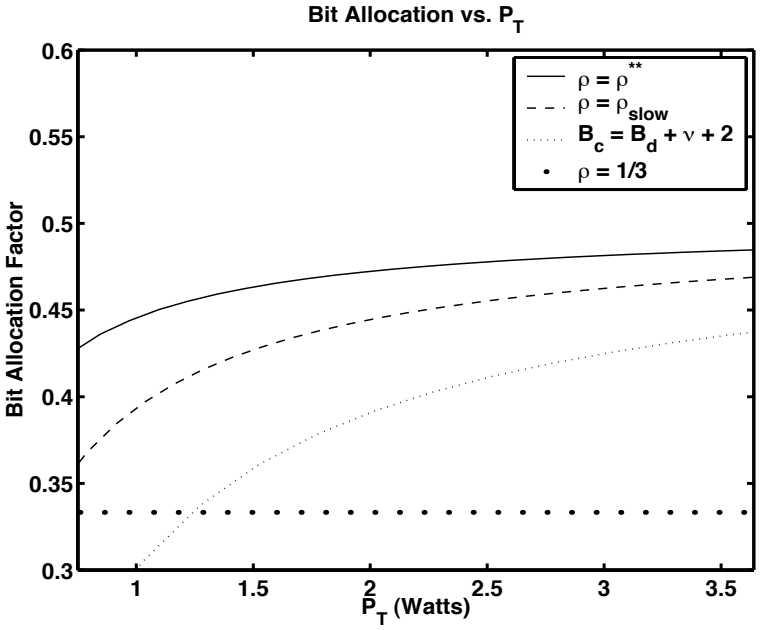

Figure 2.8: IIR channel with Gaussian input: data bit allocation factors under P_T constraint as functions of P_T.

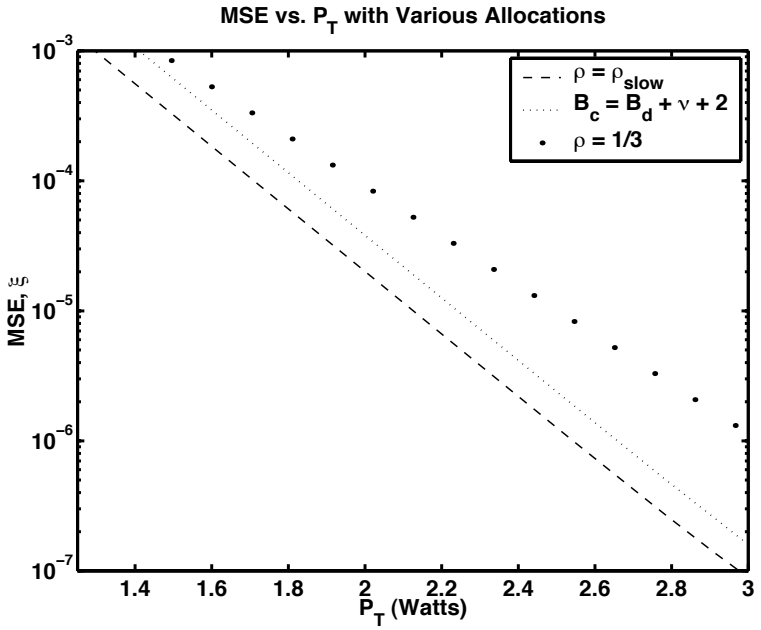

Figure 2.9: IIR channel with Gaussian input: MSE as function of P_T for various bit allocations.

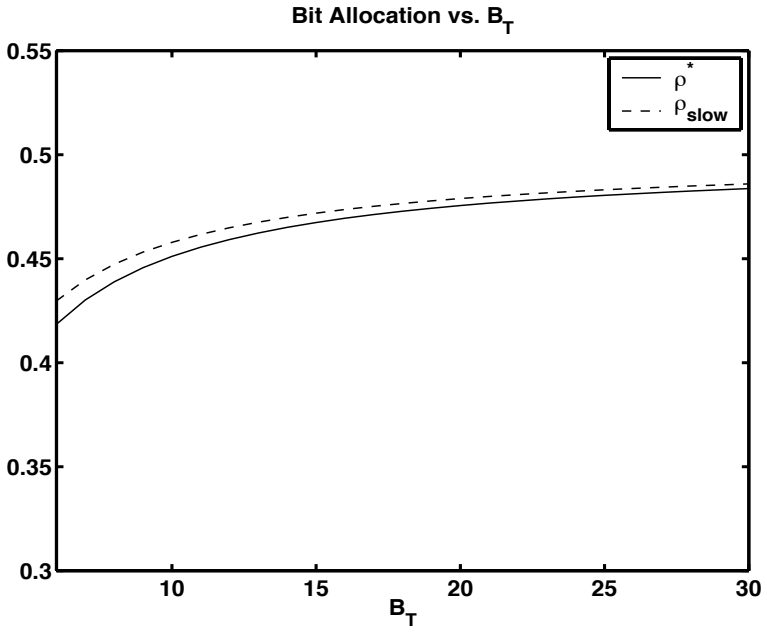

Figure 2.10: 32-tap LMS channel equalizer: data bit allocation factors under B_T constraint as functions of B_T.

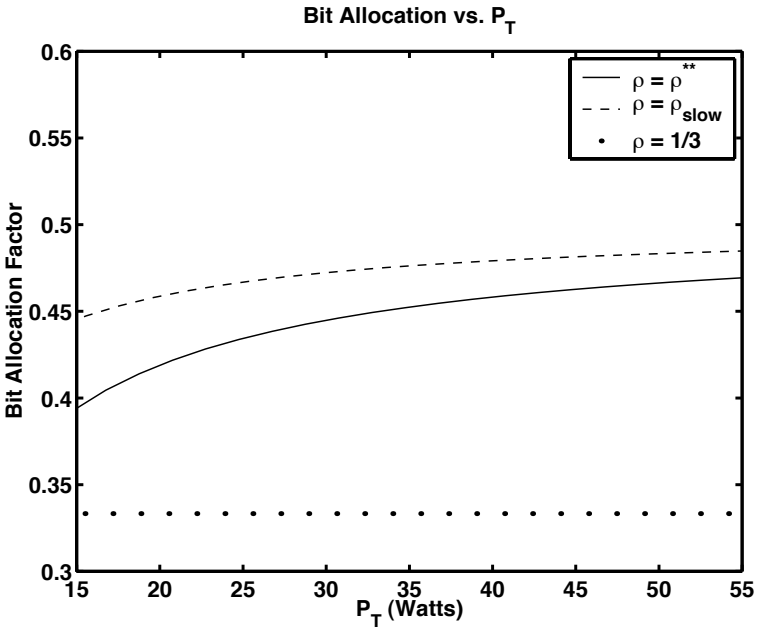

Figure 2.11: 32-tap LMS channel equalizer: data bit allocation factors under P_T constraint as functions of P_T.

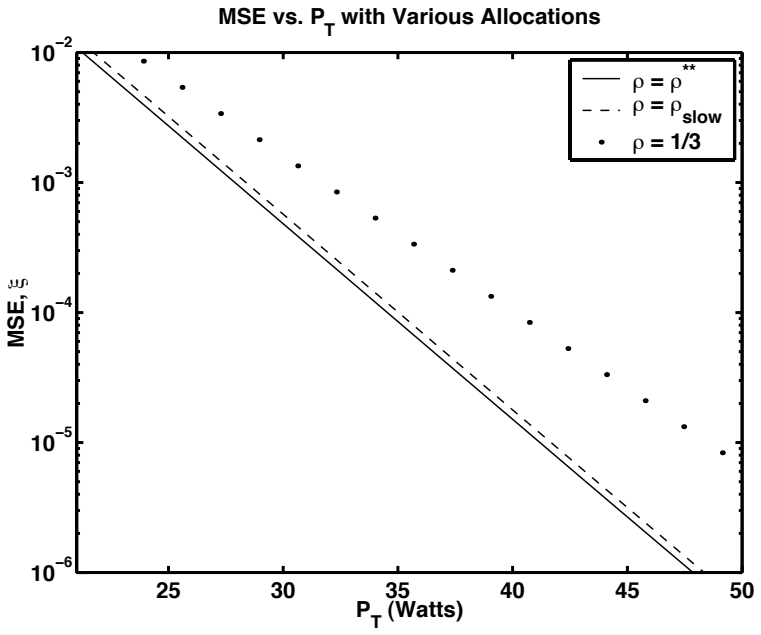

Figure 2.12: 32-tap LMS channel equalizer: MSE as function of P_T for various bit allocations.

ρ^{**}. Figure 2.12 shows the MSE as a function of P_T for this case with $\rho = \rho_{\text{slow}}$ and $\rho = \rho^{**}$. Observe that, as expected, the MSE with ρ^{**} attains the minimum.

2.5 Conclusion

In this chapter, a design methodology has been developed by which low-power implementation of LMS adaptive channel equalizers can be achieved. Expressions have been derived for optimal bit allocation under combined register length constraints and total power constraints while avoiding the slowdown phenomenon. These expressions can easily be specialized to a specific hardware implementation for computation of the number of bits to allocate to data and filter coefficients. A general conclusion is that the standard design strategy of allocating an equal number of bits to the data and filter coefficients is optimal only as the power or register length constraints become very large. Furthermore, this 50% allocation can yield undesired slowdown in the transient phase of adaptation. For most LMS implementations, it is optimal to allocate more bits to the filter coefficients than to the data.

It is important to emphasize that the linear steady-state analysis presented in this chapter is relevant only for implementations of LMS for which slowdown does not occur, i.e. for the cases that wordlength satisfies the condition (2.17). To optimize performance over all possible choices of wordlengths, including those for which slowdown occurs, a full, nonlinear, finite-precision analysis of LMS must be performed. For example, the methods of [6, 7, 8] might be applied if they could be extended to cover the case where both data and coefficients are finite precision.

CHAPTER 3

Vector Quantization for Distributed Hypothesis Testing

3.1 Introduction

In a myriad of applications, the intent of data transmission is for a user to make a decision, or hypothesis test, based upon the received data. For example, a radar sensor must transmit information to a user for determination of a target's presence. The optimality criterion by which the source encoder is designed must be directly related to the performance of the receiver's decision rule in these cases. In most communication systems, the source encoder objective function is the mean square error (MSE). This criterion is suitable when it is desired that the received data be an accurate estimate of the transmitted data. Thus, a source encoder designed for minimum MSE can be considered an *estimation-optimal* source encoder. Mean square error, however, is not the most suitable criterion for hypothesis testing. Type I and type II error probabilities are more conventional gauges of performance for hypothesis tests. A source encoder for hypothesis testing applications should therefore be designed with the error probabilities as minimization criteria. As hypothesis testing often consists of detection of a target, such a source encoder could be considered a *detection-optimal* source encoder. Detection performance can never improve with source encoding and the loss in detection performance will certainly be small when a source encoder with excellent estimation performance (low MSE) is used. However, as design of an estimation-optimal source encoder is in no way influenced by the hypothesis test for which the reconstructed data is intended, such an encoder may unnecessarily sacrifice rate, in the form of transmitted

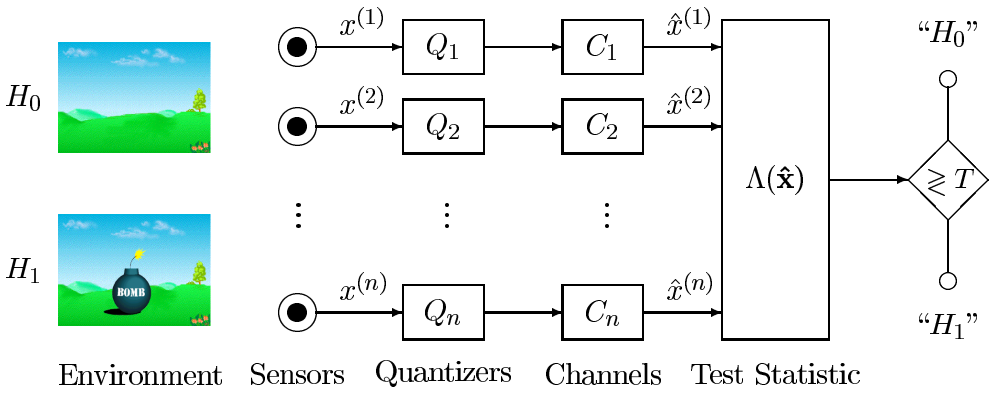

Figure 3.1: Sensor network.

bits, for information deemed useless by the hypothesis tester. It is therefore beneficial to investigate source encoding procedures that may discard this unnecessary information and transmit only the information useful for making a correct decision.

3.1.1 Distributed Hypothesis Testing

Detection-optimal source encoding is directly applicable to distributed hypothesis testing environments. In such an environment, a decision must be made based on a set of n observations, each of which is received from a node in a network. For example, these nodes could be sensors in a sensor network as in Figure 3.1. The Neyman-Pearson theorem [10, 39, 69] provides an optimal hypothesis test: the likelihood, or log-likelihood, ratio test. It is intuitively clear, and it will be shown, that the performance of the Neyman-Pearson test improves as the number of data sources (nodes) increases. Since the nodes are physically separated, they must communicate their observations to a central decision device as in Figure 3.1. From the channel coding theorem [10, 21, 61], it may be assumed that the channels over which the nodes communicate their observations are errorless as long as the data rate is below the channel capacity. Thus, the observations must be source encoded to achieve a data rate below capacity.

3.1.2 Vector Quantization

It is well known that vector quantization [27] is a powerful source coding technique that can achieve low data rates at little expense in fidelity. Analytical information-theoretic formulations indicate that improved performance may be obtained by quantizers that encode vectors rather than scalars. Further, the major drawback of vector quantization, its complexity in comparison to scalar quantization, has become increasingly less burdensome with the introduction of proper design methods [42]. Vector quantization is therefore an appealing choice for source coding in distributed hypothesis testing environments. In this chapter, the problem of vector quantizer design for optimal performance of hypothesis tests utilizing quantized data is investigated.

3.1.3 Overview of Previous Work

Quantization has been studied for many decades. Its rich history is traced in [30]. Early research on asymptotic vector quantization was done by Zador [75] and Gersho [26]. In [46], Na and Neuhoff derived a formula for the asymptotic MSE of a vector quantizer in terms of two functions that characterize the quantizer, known as the *point density* and *inertial profile*. The work in this chapter is an extension of [46] to asymptotic detectability of vector-quantized data.

The problem of optimal quantization for hypothesis testing has been analyzed for various quantization schemes and various optimality criteria. In [37], Kassam considered the composite hypotheses $\theta = 0$ versus $\theta > 0$ and used the efficacy of the sufficient statistic as the objective function. In [58], Poor and Thomas use various Ali-Silvey distances as optimality criteria and investigate non-asymptotic quantizer effects. Poor, in [56, 57], uses a generalized "f-divergence", of which the Kullback-Leibler distance is a special case, as an optimality criterion and studies asymptotic quantization effects. From this work, it can be seen that the loss in Kullback-Leibler distance due to quantization is a functional of a quantity called *discriminability*, that is defined in Section 3.5.3. Benitz and Bucklew [5]

proposed the alpha entropy as their optimality criterion as it gives the exponential decay to zero of the total probability of error of a binary hypothesis test with equal priors, according to a theorem of Chernoff. Asymptotically optimal companding functions for scalar quantizers were then derived. Picinbono and Duvaut [54] considered a deflection criterion similar to a signal-to-noise ratio (SNR) under one of two simple hypotheses. It was found that maximization of this deflection criterion is achieved by a vector quantizer that quantizes the likelihood ratio rather than the observation itself. In [68], Tsitsiklis explores some properties of these so-called likelihood ratio quantizers and investigates their optimality with respect to statistical divergences. Applications to distributed hypothesis testing are investigated as well. In [24], Flynn and Gray consider estimation and detection of correlated observations in distributed sensing environments. Achievable rate-distortion regions are obtained for the case of two sensors which extend the lossless source coding analysis of Slepian and Wolf [64] to lossy source coding. Non-asymptotic quantizer design for optimum detection performance via iterative maximization of a distributional distance called *Chernoff distance*, which is similar to the alpha entropy considered in [5], is also presented. The distributed hypothesis testing problem with quantized observations is directly addressed in [41] by Longo, Lookabaugh, and Gray. Optimal scalar quantizers are derived with the Bhattacharyya distance as the objective function and an iterative design algorithm is developed. In [1, 33, 34, 62], the effects of communication constraints, such as source encoding, on the performance of distributed hypothesis testing is investigated. In particular, the decay rate to zero of the probability of type II error is determined when the probability of type I error is constrained to be below a prescribed threshold. Oehler and Gray [50] and Perlmutter *et al.* [53] developed a method of combining vector quantization and classification by defining an objective function that incorporates MSE and Bayes risk. A vector quantizer design algorithm was derived to minimize this objective function. A summary of this work can be found in [29].

3.1.4 Overview of Contribution

While many of the studies listed above relate quantization effects to error probabilities of hypothesis tests through analytical determination of losses in statistical divergences, none directly tackles the problem of optimization of the receiver operating characteristic (ROC) curve. All of the statistical divergences used previously as optimization criteria are asymmetric functions of the hypothesized source densities. Consequently, these divergences are related to only one type (I or II) of error probability. In [5], the alpha entropy can be related to the total probability of error with equal priors, but the ROC curve is still unrelated.

In this chapter, we consider an optimality criterion, based on a statistical divergence known as Kullback-Leibler distance or *discrimination*, that is directly related to the area under the ROC curve. This criterion is a symmetric functional of the hypothesized source densities and is derived using large deviations error exponents [10]. The optimal vector quantizer that minimizes this criterion, and thus maximizes the area under the ROC curve, is derived and numerical studies are presented.

The formulation in this chapter draws heavily from the approach taken in [46]. In this framework, many-point quantizers with small cells are characterized by their point density and inertial profile functions. These functions describe a quantizer's distribution of points and cell shapes, respectively. In addition, we define a third function, similar to the inertial profile, called the *covariation profile*, that is used to characterize a quantizer's cell shapes. This framework permits a lucid analysis of the merits of various quantizers as they may be evaluated entirely by their point densities and covariation profiles. In addition, the problem of optimal quantizer design thus becomes a problem of point density and covariation profile optimization.

An outline of the chapter is as follows. We begin with some preliminary background on hypothesis testing and vector quantization in Sections 3.2.1 and 3.2.2. In Section 3.3, we

determine conditions under which quantizers can be derived that do not affect hypothesis testing performance and show that these conditions are rather restrictive. Consequently, in the sequel we focus on the majority of cases: those for which such "lossless" quantizers do not exist. In Section 3.4, we introduce the concepts and techniques that are used in subsequent sections to analyze quantizers for hypothesis testing performance. In particular, we discuss the "sequence approach" for analysis of small-cell quantization effects. We also introduce the log-likelihood ratio quantizer and discuss its merits and drawbacks. Next, in Section 3.5, certain discrimination losses, due to quantization by a many-point, small-cell quantizer, are derived. The formulas for these losses are analyzed and related to hypothesis testing performance of small-cell vector quantizers. Two important functions, the *Fisher covariation profile* and the *discriminability*, are defined. This analysis is then used in Section 3.6 to derive optimal small-cell quantizers for several objective functions. Comments are made regarding the detection and estimation performance of the various optimal quantizers. The performance of these quantizers is then compared in Section 3.7 for several specific numerical examples. Finally, in Section 3.8, we summarize the main results of the chapter. It is concluded that the best quantizer for detection performance is a log-likelihood ratio quantizer whose scalar constituent quantizer is optimized for ROC area. However, this quantizer often yields very poor estimation performance. For hypothesis testing applications in which some degree of estimation performance is desired, the various small-cell quantizers derived in Section 3.6 are optimal.

3.2 Preliminaries

3.2.1 Hypothesis Testing

In hypothesis testing problems, a random observation x or a set of random observations $\mathbf{x} = [x^{(1)}, \ldots, x^{(n)}]$ must be processed so as to decide which of a set of hypotheses is true. Each hypothesis is characterized by a family of probability distributions on the observed data. When each family contains a single distribution, the hypotheses are said to be *simple*.

In this chapter, we consider simple binary hypothesis testing of continuous-valued observations. Thus, the problem consists of deciding between two hypotheses H_0 and H_1 based on a set of observed random vectors $x^{(1)}, \ldots, x^{(n)}$. In general, we consider k-dimensional observations. Thus $x^{(i)} \in \mathbb{R}^k$ for $i = 1, \ldots, n$. Hypotheses H_0 and H_1 are sometimes referred to as the *null* and *alternate hypotheses*, respectively. Since the hypotheses are simple, we can associate probability densities with them:

$$\begin{aligned} H_0 &: x^{(i)} \sim q_{0,i}(x^{(i)}),\ i = 1, \ldots, n \\ H_1 &: x^{(i)} \sim q_{1,i}(x^{(i)}),\ i = 1, \ldots, n. \end{aligned} \tag{3.1}$$

In the case of independent and identically distributed (i.i.d.) observations, (3.1) can be simplified:

$$\begin{aligned} H_0 &: \mathbf{x} \sim q_0^{(n)}(\mathbf{x}) \\ H_1 &: \mathbf{x} \sim q_1^{(n)}(\mathbf{x}) \end{aligned} \tag{3.2}$$

where $\mathbf{x} = [x^{(i)}, \ldots, x^{(n)}]$,

$$q_0^{(n)}(\mathbf{x}) = \prod_{i=1}^{n} q_0(x^{(i)}), \quad q_1^{(n)}(\mathbf{x}) = \prod_{i=1}^{n} q_1(x^{(i)}), \tag{3.3}$$

and q_0 and q_1 are the common densities for each observation under hypotheses H_0 and H_1, respectively.

The hypothesis test consists of a decision rule that partitions the space of possible observations (called the *observation space*) into two disjoint regions $\mathcal{U}_0^{(n)}$ and $\mathcal{U}_1^{(n)}$. The decision is then based on which of these two regions contains the observation. If $\mathbf{x} \in \mathcal{U}_0^{(n)}$, then the decision is that hypothesis H_0 is true. Similarly, if $\mathbf{x} \in \mathcal{U}_1^{(n)}$, then hypothesis H_1 is chosen.

Type I and II Error Probabilities

The performance of any decision rule is dictated by two quantities: the probability of false alarm and the probability of miss. These probabilities are also referred to as *type I*

and type II probabilities of error, respectively. A type I error occurs when the decision is H_1 and hypothesis H_0 is true. Similarly, a type II error corresponds to a decision of H_0 when hypothesis H_1 is true. Let α denote the probability of false alarm and β the probability of miss. Then

$$\begin{aligned} \alpha &= P(\text{Decide } H_1 | H_0 \text{ is true}) = \int_{\mathcal{U}_1^{(n)}} q_0(\mathbf{x}) d\mathbf{x} \\ \beta &= P(\text{Decide } H_0 | H_1 \text{ is true}) = \int_{\mathcal{U}_0^{(n)}} q_1(\mathbf{x}) d\mathbf{x}. \end{aligned} \tag{3.4}$$

Bayes Test

If the probabilities of the two hypotheses (known as *prior probabilities*, or simply *priors*) are known, then a decision rule, known as a Bayes test, can be derived that minimizes a criterion known as *Bayes risk* [69]. A special case of the Bayes risk is the probability of error, or the probability of making an incorrect decision. To state this concretely, let the null and alternate hypotheses be, respectively

$$\begin{aligned} H_0 &: x \sim q_0 \\ H_1 &: x \sim q_1 \end{aligned} \tag{3.5}$$

and let the priors be

$$\begin{aligned} P_0 &= P(H_0 \text{ is true}) \\ P_1 &= P(H_1 \text{ is true}). \end{aligned}$$

The probability of error given by

$$P_e = \alpha P_0 + \beta P_1 \tag{3.6}$$

is minimized by the Bayes test [69]:

$$\frac{q_0(x)}{q_1(x)} \underset{H_1}{\overset{H_0}{\gtrless}} \frac{P_1}{P_0}. \tag{3.7}$$

Neyman-Pearson Test

In many cases, the prior probabilities are unknown and thus the Bayes test (3.7) can not be derived. In such cases, the Neyman-Pearson theorem and the receiver operating characteristic, described below, offer assistance in determining optimal decision rules.

For the hypotheses given by (3.5), the well-known Neyman-Pearson Theorem [10, 39, 48, 69] states that under the constraint $\beta \leq \beta^*$, the minimum α is achieved by a likelihood ratio or log-likelihood ratio test:

$$\Lambda(x) = \log \frac{q_0(x)}{q_1(x)} \underset{H_1}{\overset{H_0}{\gtrless}} T. \tag{3.8}$$

The threshold T will depend on the maximum allowable probability of miss β^*. The Neyman-Pearson theorem also states that under a constraint on false alarm probability $\alpha \leq \alpha^*$, the minimum β is achieved by the log-likelihood ratio test (3.8). Note that the Neyman-Pearson test (3.8) is equivalent to a Bayes test for some particular values of the priors. Specifically, if $\log(P_1/P_0) = T$, then the Neyman-Pearson test is a Bayes test. However, the utility of the Neyman-Pearson theorem is exhibited most when the priors are unknown.

Note that in the case of n i.i.d. observations, the Neyman-Pearson test becomes

$$\frac{1}{n} \sum_{i=1}^{n} \Lambda(x^{(i)}) \underset{H_1}{\overset{H_0}{\gtrless}} T \tag{3.9}$$

where the log-likelihood ratio is written in normalized form.

Receiver Operating Characteristic

A powerful illustrative tool for understanding hypothesis testing performance is the *receiver operating characteristic* (ROC) curve [9, 55, 69]. For a given decision rule, many pairs (α, β) of type I and type II error probabilities are achievable. The specific values of α and β depend on some parameter of the rule. For Neyman-Pearson tests, this parameter is the decision threshold T. Clearly, it is desirable to minimize both α and β, but in general there is a tradeoff between these two quantities. The ROC curve provides a visual

interpretation of this tradeoff. The ROC curve is a graph of the probability of detection $1-\beta$, also known as the *power* of the test, versus the probability of false alarm α for a given decision rule. For Neyman-Pearson tests, this graph is plotted parametrically as α and β are both functions of the threshold T.

Optimality Criteria for Hypothesis Tests

There are several criteria by which a decision rule may be considered optimal. Two of the most commonly used criteria are explained here. Interestingly, the Neyman-Pearson theorem shows that likelihood ratio tests are optimal with respect to both criteria.

Often hypothesis tests are designed so that the probability of miss is minimized subject to the constraint that the probability of false alarm does not exceed a pre-specified maximum tolerable value [55, 69]. That is: minimize the function $\beta(\alpha)$ on the set $\{\alpha \in (0,1) : \alpha \leq \alpha^*\}$. The Neyman-Pearson theorem indicates that for this criterion, the optimal decision rule is a likelihood or log-likelihood ratio test. The specific threshold that satisfies the false alarm constraint while minimizing the probability of miss is not given by the theorem, but can usually be determined numerically or analytically [10].

Another optimality criterion for a hypothesis test is the area under the ROC curve. Note that this is actually an optimality criterion for a *family* of hypothesis tests, since each test corresponds to only one point on the ROC curve. This criterion is similar to the previous criterion in an "average" sense, since the objective is now to minimize the integral

$$\int_0^1 \beta(\alpha) d\alpha.$$

Further motivation for selection of the area under the ROC curve as an optimality criterion comes from the fact that when the likelihood ratio is Gaussian, the area is monotonically related to the signal-to-noise ratio (SNR) by [3]

$$\text{Area} = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\text{SNR}}{2}\right)$$

where the SNR is defined as [4]

$$\text{SNR} = \frac{(E[\Lambda|H_0] - E[\Lambda|H_1])^2}{\frac{1}{2}(\text{var}[\Lambda|H_0] + \text{var}[\Lambda|H_1])}.$$

It can be shown that, like the previous criterion, the Neyman-Pearson test provides the optimal decision rule [4]. This fact follows directly from the Neyman-Pearson theorem.

Discrimination and Stein's Lemma

The *discrimination* (also known as *information discrimination*, *relative entropy*, *Kullback-Leibler distance*, and *divergence*) [10, 21] between two discrete sources with probability mass functions (pmf's) $q_a(x)$ and $q_b(x)$ is defined as

$$L(q_a||q_b) = \sum_i q_a(x_i) \log \frac{q_a(x_i)}{q_b(x_i)}. \tag{3.10}$$

Similarly, the discrimination between two continuous sources with densities $q_a(x)$ and $q_b(x)$ is defined as

$$L(q_a||q_b) = \int q_a(x) \log \frac{q_a(x)}{q_b(x)} dx. \tag{3.11}$$

Throughout this chapter, we assume that all logarithms are natural logarithms, unless otherwise noted. The discriminations defined in (3.10) and (3.11) are thus given in nats. The discrimination function finds use in many areas of information theory. Worth noting is the fact that discrimination is never negative [10, 21]. Its importance in simple binary hypothesis testing is evidenced by Stein's lemma. This lemma states that for the hypotheses given by (3.2) and (3.3), if the probability of false alarm is constrained to be less than or equal to α^*, then the minimum probability of miss β_n^*, over all decision rules with n i.i.d. observations, satisfies [10]

$$\lim_{n \to +\infty} (\beta_n^*)^{1/n} = e^{-L(q_0||q_1)}. \tag{3.12}$$

Stein's lemma indicates that for the simple binary hypothesis testing problem, the discrimination between the two source densities is an important measure of performance as the

number of observations increases. The lemma states that for any $\alpha \in (0,1)$, the minimum possible β is an exponentially decreasing function of $L(q_0||q_1)$. Another version of the lemma shows that for any $\beta \in (0,1)$, the minimum possible α is exponentially decreasing in $L(q_1||q_0)$.

Error Exponents and the Tilted Density

The asymptotic values of *both* the type I and II error probabilities as functions of the Neyman-Pearson threshold can be obtained by using large deviations arguments and Sanov's theorem [10, 12]. The essential result is that for simple binary hypotheses with n i.i.d. observations, the probability of false alarm α and the probability of miss β are both exponentially decreasing functions of discriminations. These discriminations, however, involve a third density known as the *tilted density* [10]. To state this result concretely, consider the hypotheses given by (3.2) and (3.3) and the Neyman-Pearson test (3.9). The tilted density given in [10] is

$$q_\lambda(x) = \frac{q_0(x)^{1-\lambda} q_1(x)^\lambda}{\int q_0(y)^{1-\lambda} q_1(y)^\lambda dy} \tag{3.13}$$

where $\lambda \in [0,1]$ is defined implicitly in terms of the threshold T of the Neyman-Pearson test (3.9):

$$T = \int q_{\lambda(x)} \log \frac{q_0(x)}{q_1(x)} dx = L(q_\lambda||q_1) - L(q_\lambda||q_0). \tag{3.14}$$

Then from Sanov's theorem, it can be shown that [10]

$$\begin{aligned} e^{-nL(q_\lambda||q_0)-o(n)} \leq \quad & \alpha \quad \leq e^{-nL(q_\lambda||q_0)} \\ e^{-nL(q_\lambda||q_1)-o(n)} \leq \quad & \beta \quad \leq e^{-nL(q_\lambda||q_1)}. \end{aligned}$$

Thus, for large n

$$\begin{aligned} \alpha \quad &\approx \quad e^{-nL(q_\lambda||q_0)} \\ \beta \quad &\approx \quad e^{-nL(q_\lambda||q_1)}. \end{aligned} \tag{3.15}$$

Note that in the Neyman-Pearson test (3.9), the threshold T may take any value in $\mathbb{R}$, but as $\lambda \in [0,1]$ is varied in the definition (3.14) of T, we see that $-L(q_1||q_0) \leq T \leq L(q_0||q_1)$. However, as $n \to +\infty$, by the weak law of large numbers, the normalized log-likelihood ratio is with high probability close to its conditional mean under hypothesis H_0 or H_1. The conditional means under H_0 and H_1 are $L(q_0||q_1)$ and $-L(q_1||q_0)$, respectively. Consequently, thresholds outside the range $[-L(q_1||q_0), L(q_0||q_1)]$ correspond to regions on the ROC curve where $\alpha \approx 0$ or $\beta \approx 0$. The set of thresholds within this range maps to the entire ROC curve as $n \to +\infty$.

The formulas (3.15) indicate that for a given value of the threshold T, the discriminations $L(q_\lambda||q_0)$ and $L(q_\lambda||q_1)$ are crucial in determining the performance of the Neyman-Pearson test. In Section 3.6.3, we shall utilize this fact for vector quantizer design.

Chernoff Information

The asymptotic behavior of the total probability of error in a Bayes test can be determined using Chernoff's theorem [21]. This theorem states that for n large, the probability of error in the Bayes test (3.7) is exponentially decreasing in n and the greatest possible exponent in the probability of error is

$$C(q_0, q_1) = L(q_{\lambda^*}||q_0) = L(q_{\lambda^*}||q_1) \tag{3.16}$$

where λ^* is chosen such that the two discriminations $L(q_{\lambda^*}||q_0)$ and $L(q_{\lambda^*}||q_1)$ are equal. The quantity $C(q_0, q_1)$ in (3.16) is called the *Chernoff information*.

3.2.2 Vector Quantization

Vector quantization [26, 27, 30, 46, 60, 75, 76] is an effective source coding technique for random vectors. Like a scalar quantizer, a vector quantizer consists of a set of cells and codebook points. Unlike a scalar quantizer, whose set of cells consists of intervals or unions of intervals, a vector quantizer's cells are regions in multidimensional Euclidean space. The advantage of vector quantization over scalar quantization is the ability to achieve non-

rectangular cell shapes. Often this results in improved performance, usually measured by mean square error.

Mathematical Description

A vector quantizer [27, 46] (VQ) $Q = (\mathcal{S}, \mathcal{C})$ consists of a codebook $\mathcal{C} = \{x_1, \ldots, x_N\}$ and a set of cells $\mathcal{S} = \{S_1, \ldots, S_N\}$ that partition $\mathbb{R}^k$. For each i, the codebook point x_i lies in cell S_i. The VQ operator can be written as

$$Q(x) = x_i, \text{ for } x \in S_i$$

where the input $x \in \mathbb{R}^k$ and the VQ is said to be k-dimensional. For a vector quantizer Q, let $V_i = \int_{S_i} dx$ denote the volume of the ith cell. The *specific point density* [46] of Q is defined as

$$\zeta_s(x) = \frac{1}{NV_i}, \text{ for } x \in S_i.$$

For large N, as its name suggests, this function is a density of points. When integrated over a small region A, it gives the approximate fraction of codebook points contained in A. Next, define the *diameter function* of the VQ as

$$d(x) = \sup\{\|u - v\| : u, v \in S_i\}, \text{ for } x \in S_i.$$

The (scalar) *specific inertial profile* function $m_s(x)$ is defined as [26]

$$m_s(x) = \frac{\int_{S_i} \|y - x_i\|^2 dy}{V_i^{1+2/k}}, \text{ for } x \in S_i. \tag{3.17}$$

Note that $m_s(x)$ is invariant to a scaling of S_i. This function contains information about the shapes of the cells of the VQ. Similarly, we define the following matrix function $M_s(x)$ called the *specific covariation profile*

$$M_s(x) = \frac{\int_{S_i} (y - x_i)(y - x_i)^T dy}{V_i^{1+2/k}}, \text{ for } x \in S_i.$$

This function is also scale invariant.

Role of Vector Quantization in Communication Systems

In a communication system, a message must be transmitted from one location to another across a channel that usually causes degradation of the transmitted signal. The objective of the communication system designer is to transmit the message with minimal degradation and at the maximum data rate [59].

In his seminal 1948 paper [61], Shannon showed that, through channel coding, a channel may be considered errorless when the rate of the transmitted data does not exceed a threshold known as *channel capacity*. However, meeting this rate constraint usually requires lossy compression of the message, especially when the message is continuously distributed.

Compression, or source coding, is therefore an important step in the transmission process. The *rate* of a source encoder, defined to be the average number of bits output per input sample, must be made as low as possible so that transmission above channel capacity is not attempted, and so that transmission is carried out efficiently. The rate of a vector quantizer is $\log_2 N$ where N is the number of quantizer cells. A good quantizer has both a small rate and high *fidelity*, or quality of the decoded message. High fidelity is equivalent to low *distortion*. Vector quantization is a powerful source coding technique for both univariate and multivariate sources. In the following section, we describe the most common measure of distortion for VQ design. Then, in the remainder of the chapter, we formulate the theory for distortion measures that can be used to design VQ's that are optimal for hypothesis testing.

Estimation-Optimal Vector Quantization

Much of the research done on vector quantization has focused on the rth-power distortion [46] given by

$$\begin{aligned} D &= \frac{1}{k} E\left[\|x - Q(x)\|^r\right] \\ &= \frac{1}{k} \int \|x - Q(x)\|^r q(x) dx \end{aligned}$$

where $q(x)$ is the probability density of the source. If we set $r = 2$ and use $D \cdot k$ as the distortion measure, we get the commonly-used mean square reconstruction error (MSE):

$$\text{MSE } = E\left[\|x - Q(x)\|^2\right].$$

Since reconstruction MSE is a widely-used distortion measure for estimators [55], we refer to minimum MSE quantizers as *estimation-optimal* quantizers.

The asymptotic (in N) rth-power distortion of a many-point VQ has been determined in [46]. To determine this value, a "sequence approach" was taken, in which a sequence of quantizers $\{Q_N\}$ was considered. Each quantizer in the sequence was characterized by its specific point density and specific inertial profile.[1] Then, assuming that the sequences of specific point densities and specific inertial profiles converge to functions $\zeta(x)$ and $m(x)$, called the *point density* and *inertial profile*, respectively, and assuming the sequence of diameter functions converges to zero, it is shown that

$$\lim_{N\to+\infty} N^{2/k} E\left[\|x - Q_N(x)\|^2\right] = \int \frac{q(x)m(x)}{\zeta(x)^{2/k}} dx. \tag{3.18}$$

Equation (3.18), known as *Bennet's integral*, gives an approximation to the MSE of a many-point VQ. For a given inertial profile, the point density that minimizes Bennet's integral, obtained by Hölder's inequality or calculus of variations, is given by [43, 47]

$$\zeta(x) = \frac{[q(x)m(x)]^{\frac{k}{k+2}}}{\int [q(y)m(y)]^{\frac{k}{k+2}} dy}.$$

Using this in (3.18) gives

$$\text{MSE} \approx \frac{1}{N^{2/k}} \left(\int [q(x)m(x)]^{\frac{k}{k+2}} dx \right)^{\frac{k+2}{k}}.$$

A conjecture by Gersho [26], believed by many to be true, states that many-point VQ's that are optimal with respect to rth-power distortion have cells that are approximately congruent. Furthermore, the moment of inertia of the congruent cells is the minimum moment of inertia $m^*_{s,k}$ of all cells that tessellate in $\mathbb{R}^k$. A quantizer with congruent cells

[1] Note that the specific inertial profile in [46] differs slightly from our definition.

has a constant inertial profile and, as a result, the estimation-optimal point density is given by [26, 46]

$$\zeta^e(x) = \frac{q(x)^{\frac{k}{k+2}}}{\int q(y)^{\frac{k}{k+2}} dy}. \tag{3.19}$$

The MSE of an N-point, k-dimensional VQ can then be upper bounded by the so-called Zador-Gersho formula [26, 75]:

$$\text{MSE} \lessapprox \frac{m^*_{s,k}}{N^{2/k}} \left(\int q(x)^{\frac{k}{k+2}} dx \right)^{\frac{k+2}{k}}. \tag{3.20}$$

The covariation profile (the limit of the sequence of specific covariation profiles) of the estimation-optimal quantizer is also constant, as the cells are congruent, and is equal to a multiple of the identity matrix. This follows from a theorem by Zamir and Feder [77] that states that the components of the error vector of an optimal lattice quantizer (a quantizer for which each cell is a translation of a basic cell) are uncorrelated.

Bennet's integral is well suited to the analysis of the estimation performance of structured quantizers. In [46] it is shown that the (estimation) performance loss of a suboptimal quantizer, defined as the ratio of the quantizer's distortion (MSE) to that of the optimal quantizer with the same rate, can be factored into the product of two individual losses called the *point density loss* and the *cell shape loss*. As their names suggest, these losses are attributable to the suboptimality of the quantizer's point density and inertial profile (cell shapes), respectively.

3.3 Lossless Quantizers for Distributed Hypothesis Testing

In the remainder of this chapter, we develop procedures for designing vector quantizers that are optimal for the distributed hypothesis testing problem illustrated in Figure 3.1. We focus on the case of i.i.d. observations $x^{(1)}, \ldots, x^{(n)}$, but point out that the analysis can be extended to independent, non-identical observations as well. Our goal is to design the quantizers $Q_1, \ldots, Q_n$ in Figure 3.1 so that the performance of the hypothesis test utilizing the n quantized observations is maximized. Unless otherwise stated, we assume that the

hypothesis test is a Neyman-Pearson test. Each channel $C_1, \ldots, C_n$ is assumed to have the same capacity and therefore each quantizer will have the same rate and the same number of cells N. Again, the analysis can be extended to cover the case of different quantizer rates. Clearly, the optimal quantizers are strongly dependent on the source densities q_0 and q_1. In fact, for certain densities there exist quantizers that have no effect on hypothesis testing performance. We refer to these quantizers as *lossless quantizers for distributed hypothesis testing* or simply *lossless quantizers*. When a quantizer degrades hypothesis testing performance, it is said to be *lossy*. Recall that for estimation objectives, all quantizers are lossy. In this section, we determine conditions on the source densities under which lossless quantizers exist. We show that, for these cases, which are restrictive and uncommon, the lossless quantizers are easy to derive. In the forthcoming sections, we derive optimal lossy quantizers for the more common cases in which quantization necessarily degrades hypothesis testing performance.

We begin by defining the notation. The i.i.d. observations, source densities, joint source densities, and hypotheses are given by equations (3.1), (3.2), and (3.3). Recall that each element of $\mathbf{x}$ is k-dimensional and thus $\mathbf{x} \in \mathbb{R}^{kn}$. Let the ith quantizer Q_i in Figure 3.1 have the N cells $\{S_{i,1}, \ldots, S_{i,N}\}$ and codebook points $\{x_{i,1}, \ldots, x_{i,N}\}$.

Next, we define

$$Q^{(n)}(\mathbf{x}) = \left[Q_1\big(x^{(1)}\big), \ldots, Q_n\big(x^{(n)}\big)\right].$$

Although $Q^{(n)}$ is not an explicit quantizer, it is equivalent to a kn-dimensional *product quantizer* [46] with N^n cells which we will denote $\{R_1, \ldots, R_{N^n}\}$ and codebook points $\{\mathbf{x}_1, \ldots, \mathbf{x}_{N^n}\}$. The cells and codebook points of $Q^{(n)}$ are n-fold Cartesian products of the cells and codebook points of the constituent quantizers $Q_1, \ldots Q_n$. Thus

$$\begin{aligned} \{R_1, \ldots, R_{N^n}\} &= \{S_{1,j_1} \times \cdots \times S_{n,j_n} : j_1, \ldots j_n \in \{1, \ldots N\}\}, \\ \{\mathbf{x}_1, \ldots, \mathbf{x}_{N^n}\} &= \{[x_{1,j_1}, \ldots, x_{n,j_n}] : j_1, \ldots j_n \in \{1, \ldots N\}\} \end{aligned}$$

and the product quantizer codebook point $[x_{1,j_1}, \ldots, x_{n,j_n}]$ lies in the cell given by $S_{1,j_1} \times \cdots \times S_{n,j_n}$.

The probability mass functions of the observation $x^{(i)}$ after quantization by the quantizer Q_i, under hypotheses H_0 and H_1, are given by

$$\begin{aligned} \bar{q}_{0,Q_i,j} &= \int_{S_{i,j}} q_0(x)dx \\ \bar{q}_{1,Q_i,j} &= \int_{S_{i,j}} q_1(x)dx \end{aligned} \tag{3.21}$$

for $j \in \{1, \ldots, N\}$. Note that since $x^{(1)}, \ldots, x^{(n)}$ are i.i.d., the quantized observations are also independent, though not necessarily identically distributed. The joint probability mass functions of the quantized observation $Q^{(n)}(\mathbf{x})$ are

$$\begin{aligned} \bar{q}_0^{(n)}\left([x_{1,j_1}, \ldots, x_{n,j_n}]\right) &= \prod_{i=1}^{n} \bar{q}_{0,Q_i,j_i} \\ \bar{q}_1^{(n)}\left([x_{1,j_1}, \ldots, x_{n,j_n}]\right) &= \prod_{i=1}^{n} \bar{q}_{1,Q_i,j_i}. \end{aligned} \tag{3.22}$$

The probabilities of type I and II errors of a Neyman-Pearson test using the quantized observation $Q^{(n)}(\mathbf{x})$ are partial sums of the pmf's in (3.22). From (3.21), it is clear that these error probabilities are independent of the codebooks of quantizers $Q_1, \ldots, Q_n$. From (3.10), the discriminations between the marginal and joint pmf's of the quantized sources, under the two hypotheses, are also independent of the quantizers' codebooks. Thus, the quantizers may be characterized solely by their partitions, or cells.

The following theorem shows that for certain hypothesis testing objectives with i.i.d. observations, the optimal quantizers $Q_1, \ldots, Q_n$ must have identical cells. This allows us to restrict our attention to the design of a single quantizer. We then derive conditions under which this single quantizer is lossless.

Theorem 3.1 *Let the observations $x^{(1)}, \ldots, x^{(n)}$ be i.i.d. under each hypothesis. If the quantizers $Q_1, \ldots, Q_n$ maximize the discrimination $L\left(\bar{q}_0^{(n)} \| \bar{q}_1^{(n)}\right)$, then their cells are equivalent.*

Proof: Since the observations are independent, the discrimination between the joint pmf's is the sum of the individual discriminations [10]:

$$L\big(\bar{q}_0^{(n)} \| \bar{q}_1^{(n)}\big) = \sum_{i=1}^{n} L(\bar{q}_{0,Q_i} \| \bar{q}_{1,Q_i})$$

where $\bar{q}_{0,Q_i}$ and $\bar{q}_{1,Q_i}$ are the pmf's of $Q_i(x^{(i)})$, whose probabilities are given by (3.21). Now suppose the quantizers $Q_1, \ldots, Q_n$ maximize $L\big(\bar{q}_0^{(n)} \| \bar{q}_1^{(n)}\big)$. Then, for each $i \in \{1, \ldots, n\}$, the discrimination $L(\bar{q}_{0,Q_i} \| \bar{q}_{1,Q_i})$ must be maximized by the ith quantizer Q_i. This discrimination is a function of the marginal densities q_0 and q_1 as well as the ith quantizer's cells. Therefore, each quantizer must have the same cells. □

Identical-partition quantizers are also optimal for asymptotic probability of false alarm or miss. To prove this, we must first extend the relations (3.15) to the case of independent, non-identical observations. Doing so gives the following asymptotic equations for $\hat{\alpha}$ and $\hat{\beta}$, the probabilities of type I and II errors with n quantized observations $Q_1\big(x^{(1)}\big), \ldots, Q_n\big(x^{(n)}\big)$:

$$\begin{aligned} \log \hat{\alpha} &\approx -\sum_{i=1}^{n} L(\hat{q}_{\lambda,Q_i} \| \bar{q}_{0,Q_i}) \\ \log \hat{\beta} &\approx -\sum_{i=1}^{n} L(\hat{q}_{\lambda,Q_i} \| \bar{q}_{1,Q_i}) \end{aligned}$$

where $\hat{q}_{\lambda,Q_i}$ is the tilted mass function associated with $\bar{q}_{0,Q_i}$ and $\bar{q}_{1,Q_i}$ (see Section 3.5.1 and equation (3.32)). By arguments similar to those in the proof of Theorem 3.1, we conclude that quantizers that are optimal with respect to $\hat{\alpha}$ or $\hat{\beta}$ have identical cells.

We now focus on quantizers $Q_1, \ldots, Q_n$ with identical cells. Further, since a quantizer's codebook does not affect hypothesis testing performance, we assume the quantizers have identical codebooks as well. Thus, the quantizers are identical and will be denoted Q. The product quantizer $Q^{(n)}$ then consists of n identical constituent quantizers Q and

$$Q^{(n)}(\mathbf{x}) = \Big[Q\big(x^{(1)}\big), \ldots, Q\big(x^{(n)}\big)\Big].$$

Next we define $\bar{q}_0$ and $\bar{q}_1$ to be the pmf's of the two sources after quantization with Q:

$$\begin{aligned} \bar{q}_{0,i} &= \int_{S_i} q_0(x)dx \\ \bar{q}_{1,i} &= \int_{S_i} q_1(x)dx \end{aligned}$$

for $i \in \{1, \ldots, N\}$. Similarly, $\bar{q}_0^{(n)}$ and $\bar{q}_1^{(n)}$ are the pmf's of $Q^{(n)}(\mathbf{x})$ under hypotheses H_0 and H_1, respectively. Finally, we define $\mathcal{U}_0^{(n)}$ and $\mathcal{U}_1^{(n)}$ to be the decision regions of the Neyman-Pearson test with n unquantized observations and threshold T, given by

$$\frac{q_0^{(n)}(\mathbf{x})}{q_1^{(n)}(\mathbf{x})} \underset{<}{\overset{H_0}{>}} T. \tag{3.23}$$

A quantizer is considered lossless if it does not degrade hypothesis testing performance. This can be true for a particular Neyman-Pearson threshold or for all thresholds, in which case $Q^{(n)}(\mathbf{x})$ is a sufficient statistic. Therefore, we define two types of lossless quantizers. A quantizer Q is a *sufficient quantizer* for distributed hypothesis testing if $Q^{(n)}(\mathbf{x})$ is a sufficient statistic for deciding between H_0 and H_1 based on observation of $\mathbf{x}$. A quantizer Q is a *Neyman-Pearson quantizer* for threshold T if there is a decision rule using $Q^{(n)}(\mathbf{x})$ with type I and II error probabilities $\hat{\alpha}$ and $\hat{\beta}$ given by $\hat{\alpha} = \alpha$ and $\hat{\beta} = \beta$ where α and β are the type I and II error probabilities of the decision rule (3.23). Note that a quantizer may be a Neyman-Pearson quantizer for more than one threshold. Note also that a sufficient quantizer is a Neyman-Pearson quantizer for any threshold.

3.3.1 Sufficient Quantizers

The following theorem indicates when a quantizer is sufficient.

Theorem 3.2 *Q is sufficient if and only if*

$$\frac{q_0(x)}{q_1(x)} = \frac{\bar{q}_{0,i}}{\bar{q}_{1,i}}$$

almost everywhere on S_i for all $i = 1, \ldots, N$.

Proof: The statistic $Q^{(n)}(\mathbf{x})$ is sufficient for deciding between H_0 and H_1 based on observation of $\mathbf{x}$ if and only if [10]

$$L(q_0^{(n)}||q_1^{(n)}) = L(\bar{q}_0^{(n)}||\bar{q}_1^{(n)}).$$

Since the observations $x^{(1)}, \ldots, x^{(n)}$ are i.i.d., this is equivalent to

$$L(q_0||q_1) = L(\bar{q}_0||\bar{q}_1).$$

The discrimination between the sources q_0 and q_1 is

$$\begin{aligned} L(q_0||q_1) &= \int q_0(x) \log \frac{q_0(x)}{q_1(x)} dx \\ &= \sum_{i=1}^{N} \int_{S_i} q_0(x) \log \frac{q_0(x)}{q_1(x)} dx. \end{aligned}$$

The discrimination between the quantized sources $\bar{q}_0$ and $\bar{q}_1$ is

$$L(\bar{q}_0||\bar{q}_1) = \sum_{i=1}^{N} \bar{q}_{0,i} \log \frac{\bar{q}_{0,i}}{\bar{q}_{1,i}}.$$

The loss in discrimination due to quantization is

$$\begin{aligned} L(q_0||q_1) - L(\bar{q}_0||\bar{q}_1) &= \sum_{i=1}^{N} \int_{S_i} q_0(x) \log \frac{q_0(x)\bar{q}_{1,i}}{q_1(x)\bar{q}_{0,i}} dx \\ &\geq \sum_{i=1}^{N} \int_{S_i} q_0(x) \left[1 - \frac{q_1(x)\bar{q}_{0,i}}{q_0(x)\bar{q}_{1,i}}\right] dx \\ &= 1 - \sum_{i=1}^{N} \frac{\bar{q}_{0,i}}{\bar{q}_{1,i}} \int_{S_i} q_1(x) dx \\ &= 0. \end{aligned} \tag{3.24}$$

The inequality in the second step of (3.24) derives from $\log(1/a) \geq 1-a$. Thus the inequality $L(q_0||q_1) \geq L(\bar{q}_0||\bar{q}_1)$ holds with equality if and only if

$$\frac{q_1(x)\bar{q}_{0,i}}{q_0(x)\bar{q}_{1,i}} = 1$$

almost everywhere on S_i for all $i = 1, \ldots, N$, which proves the theorem. □

Theorem 3.2 gives a condition on the quantizer cells that ensures sufficiency. The next theorem gives a condition on the source densities that ensures the existence of a sufficient quantizer.

Theorem 3.3 *A sufficient quantizer exists if and only if the likelihood ratio $q_0(x)/q_1(x)$ is piecewise-constant almost everywhere on $\mathbb{R}^k$.*

Proof: Suppose the likelihood ratio is piecewise constant almost everywhere. Then we can write

$$\frac{q_0(x)}{q_1(x)} = \sum_{i=1}^{K} a_i I_{A_i}(x) + z(x)$$

where $\{A_1, \ldots, A_K\}$ is a set of disjoint, exhaustive, and connected regions in $\mathbb{R}^k$, $a_i \in \mathbb{R}$ for all $i = 1, \ldots, K$, $I_U(x)$ is the indicator function of the set U, and $z(x)$ is zero almost everywhere. [2]

Let the quantizer Q have $N = K$ cells $\{S_1, \ldots, S_N\}$ where $S_i = A_i$ for all i. Now, $q_0(x)/q_1(x) = a_i$ almost everywhere on S_i. Therefore

$$\frac{\bar{q}_{0,i}}{\bar{q}_{1,i}} = \frac{\int_{S_i} q_0(y)dy}{\int_{S_i} q_1(y)dy} = \frac{\int_{S_i} a_i q_1(y)dy}{\int_{S_i} q_1(y)dy} = a_i.$$

Thus

$$\frac{\bar{q}_{0,i}}{\bar{q}_{1,i}} = \frac{q_0(x)}{q_1(x)}$$

almost everywhere on S_i. By Theorem 3.2, Q is sufficient.

Next, let $l(x)$ be the likelihood ratio and suppose $l(x)$ is not piecewise-constant almost everywhere. Then for an N-point quantizer with $N < +\infty$, there is a cell S_i such that $l(x)$ is not constant almost everywhere on S_i. From Theorem 3.2, the quantizer is not sufficient.

□

Note that Theorem 3.3 is equivalent to stating that a sufficient quantizer exists if and only if the joint likelihood ratio can be written in the form

$$\frac{q_0^{(n)}(\mathbf{x})}{q_1^{(n)}(\mathbf{x})} = \prod_{j=1}^{n} \sum_{i=1}^{N} a_i I_{A_i}\left(x^{(j)}\right) + z(\mathbf{x})$$

where $z(\mathbf{x})$ is zero almost everywhere on $\mathbb{R}^{kn}$.

[2] We take this to be the definition of piecewise-constant almost everywhere.

3.3.2 Neyman-Pearson Quantizers

The previous section gave a condition on the quantizer's cells that guaranteed sufficiency and a condition on the sources that ensured existence of a sufficient quantizer. In this section, we prove similar theorems for Neyman-Pearson quantizers. The first theorem states that each cell of the product quantizer $Q^{(n)}$ associated with a Neyman-Pearson quantizer Q is contained in one of the Neyman-Pearson regions $\mathcal{U}_0^{(n)}$ or $\mathcal{U}_1^{(n)}$.

Theorem 3.4 *A quantizer Q is a Neyman-Pearson quantizer for threshold T if and only if for every cell R_i of $Q^{(n)}$ with $P(R_i|H_0) > 0$ and $P(R_i|H_1) > 0$, $V\big(R_i \cap \mathcal{U}_0^{(n)}\big) = 0$ or $V\big(R_i \cap \mathcal{U}_1^{(n)}\big) = 0$ where $V(U)$ is the volume of the set U.*

Proof: First assume the conditions of the theorem hold. We note that $T > 0$, $\mathcal{U}_0^{(n)} = \{\mathbf{x} : q_0^{(n)}(\mathbf{x}) > Tq_1^{(n)}(\mathbf{x})\}$, and $\mathcal{U}_1^{(n)} = \{\mathbf{x} : q_0^{(n)}(\mathbf{x}) < Tq_1^{(n)}(\mathbf{x})\}$. If $P(R_i|H_0) = 0$, then $q_0^{(n)}(\mathbf{x}) = 0$ a.e. on R_i and, therefore, $V\big(R_i \cap \mathcal{U}_0^{(n)}\big) = 0$. Similarly, if $P(R_i|H_1) = 0$, then $V\big(R_i \cap \mathcal{U}_1^{(n)}\big) = 0$. Thus, for each cell R_i with non-zero probability under at least one hypothesis, one of the volumes is zero.

Now define the following decision rule based on observation of $Q^{(n)}(\mathbf{x}) = \mathbf{x}_i$:

$$V\big(R_i \cap \mathcal{U}_0^{(n)}\big) \underset{H_1}{\overset{H_0}{\gtrless}} V\big(R_i \cap \mathcal{U}_1^{(n)}\big). \tag{3.25}$$

This rule can be implemented in a more straightforward manner if the codebook is selected properly. Later, we will see that this is actually a Neyman-Pearson test on the quantized data. Recall that for any i, one of the volumes in (3.25) must be zero.

The probability of false alarm with this decision rule is

$$\hat{\alpha} = \sum_{\{i:V(R_i \cap \mathcal{U}_0^{(n)})=0\}} \bar{q}_0^{(n)}(\mathbf{x}_i) = \sum_{\{i:V(R_i \cap \mathcal{U}_0^{(n)})=0\}} \int_{R_i} q_0^{(n)}(\mathbf{x}) d\mathbf{x} = \int_{\mathcal{U}_1^{(n)}} q_0^{(n)}(\mathbf{x}) d\mathbf{x} = \alpha$$

where α is the probability of false alarm of the Neyman-Pearson decision rule with unquantized data (3.23). It can be similarly shown that the probability of miss is unchanged as well. Thus, Q is a Neyman-Pearson quantizer for threshold T, and (3.25) is a Neyman-Pearson test.

Next assume there is a cell R_i such that $P(R_i|H_0) > 0$, $P(R_i|H_1) > 0$, $V\big(R_i \cap \mathcal{U}_0^{(n)}\big) > 0$, and $V\big(R_i \cap \mathcal{U}_1^{(n)}\big) > 0$. Let $\hat{D}$ be a decision rule based on observation of $Q^{(n)}(\mathbf{x})$ and define the decision rule $D(\mathbf{x}) = \hat{D}\big(Q^{(n)}(\mathbf{x})\big)$. Clearly D and $\hat{D}$ have the same performance. Now, D must make the same decision for all $\mathbf{x} \in R_i$. However, R_i intersects both Neyman-Pearson regions with non-zero volume and has non-zero probability under both hypotheses. Therefore, the decision regions of D are not the Neyman-Pearson regions and, by the Neyman-Pearson theorem, the type I and II error probabilities of D and $\hat{D}$ are *not* equal to α and β. □

Theorem 3.5 *A Neyman-Pearson quantizer exists if and only if $\mathcal{U}_0^{(n)}$ and $\mathcal{U}_1^{(n)}$ are unions of n-fold Cartesian products of a finite number of connected regions in $\mathbb{R}^k$, i.e. there is a set $A = \{A_1, \ldots, A_K\}$ of connected regions in $\mathbb{R}^k$ such that $\mathcal{U}_0^{(n)}$ and $\mathcal{U}_1^{(n)}$ are unions of sets in B where*

$$B = \left\{ \text{all sets of the form } \bigtimes_{j=1}^{n} A_{i_j}, \text{ where } i_j \in \{1, \ldots, K\} \right\}.$$

Proof: Suppose there is a set A as defined above such that $\mathcal{U}_0^{(n)}$ and $\mathcal{U}_1^{(n)}$ are both finite unions of sets in B. Let the quantizer Q have $N = K$ cells $\{S_1, \ldots, S_N\}$ where $S_i = A_i$ for all i. Then the cells of the product quantizer $Q^{(n)}$ are the sets in B. Thus $\mathcal{U}_0^{(n)}$ and $\mathcal{U}_1^{(n)}$ are unions of cells of $Q^{(n)}$. Since the Neyman-Pearson regions are disjoint, it follows that no cell intersects both regions. From Theorem 3.4, the quantizer is a Neyman-Pearson quantizer.

Next, assume that there is no set A such that $\mathcal{U}_0^{(n)}$ and $\mathcal{U}_1^{(n)}$ are unions of Cartesian products of sets in A. Let Q be an N-cell, k-dimensional quantizer with cells $\{S_1, \ldots, S_N\}$. Then the cells of $Q^{(n)}$ are

$$R = \left\{ \text{all sets of the form } \bigtimes_{j=1}^{n} S_{i_j} \text{ where } i_j \in \{1, \ldots, N\} \right\}.$$

Since $\mathcal{U}_0^{(n)}$ and $\mathcal{U}_1^{(n)}$ are not unions of cells of $Q^{(n)}$, it follows that at least one cell of $Q^{(n)}$ must intersect both Neyman-Pearson regions with positive volume. □

3.3.3 Examples of Lossless Quantizers

We now give some examples of sources for which lossless quantizers exist. First we show that for $n = 1$, a Neyman-Pearson quantizer always exists. For $n > 1$, however, lossless quantizers usually do not exist.

Non-Distributed Hypothesis Testing

When the observation $\mathbf{x}$ comes from a single source ($n = 1$), a Neyman-Pearson quantizer always exists. Thus, for the non-distributed hypothesis testing scenario, it is possible to design a quantizer for which no performance loss is incurred. Furthermore, in some cases this quantizer can have a rate of only one bit.

To prove this, we first note that for $n = 1$, the Neyman-Pearson regions $\mathcal{U}_0^{(n)}$ and $\mathcal{U}_1^{(n)}$ for threshold T are subsets of $\mathbb{R}^k$. Assume each of these regions is connected. Then clearly the conditions of Theorem 3.5 are satisfied with $A = \{\mathcal{U}_0^{(n)}, \mathcal{U}_1^{(n)}\}$ and $n = 1$. Next, let the quantizer $Q = Q^{(n)}$ have cells $\mathcal{U}_0^{(n)}$ and $\mathcal{U}_1^{(n)}$. By Theorem 3.4, Q is a Neyman-Pearson quantizer for threshold T. Since Q has only two cells, its rate is one bit. This formulation can easily be extended to the case of non-connected Neyman-Pearson regions. In such cases, a Neyman-Pearson quantizer must have more than two cells.

Next, suppose a quantizer is a Neyman-Pearson quantizer for K different thresholds for $n = 1$. Then there are up to K pairs of type I and II error probabilities that are unaffected by the quantizer. Therefore, the ROC curve of the Neyman-Pearson test with quantized data will intersect that of the unquantized curve in up to K places as shown in Figure 3.2. Suppose now that we have $n = 2$ observations and two quantizers. In most cases, when $n > 1$ Neyman-Pearson quantizers do not exist. Thus the quantized ROC curve will not intersect the unquantized curve. This is also illustrated in Figure 3.2.

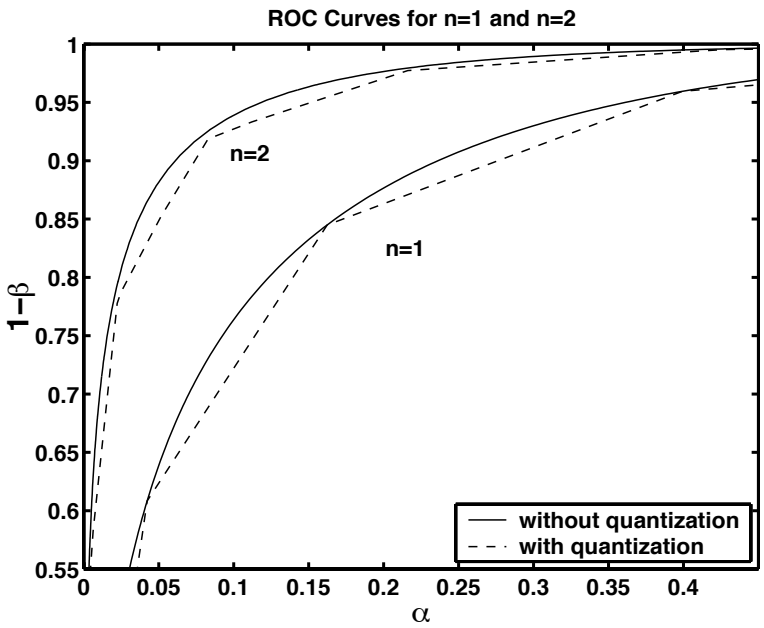

Figure 3.2: ROC curves of Neyman-Pearson tests with quantized and unquantized data for $n = 1$ and $n = 2$ observations. For $n = 1$, the quantizer is a Neyman-Pearson quantizer for several thresholds.

Piecewise-Constant Sources

When both sources q_0 and q_1 are piecewise-constant, it follows that the likelihood ratio q_0/q_1 is also piecewise-constant and thus by Theorem 3.3, a sufficient quantizer exists. Let $\{A_1, \ldots, A_{N_0}\}$ be the set of connected regions in $\mathbb{R}^k$ on which q_0 is constant. Similarly, let $\{B_1, \ldots, B_{N_1}\}$ be the connected regions where q_1 is constant. Let the quantizer Q have a set of cells $S = \{S_1, \ldots, S_N\}$ consisting of all non-empty intersections of regions in A with regions in B. Then it can easily be seen from Theorem 3.2 that Q is sufficient. The number of cells N is no more than $N_0 N_1$ and the rate of the quantizer is upper bounded by $\log_2 N_0 + \log_2 N_1$.

Gaussian Sources

When both sources are Gaussian, the likelihood ratio is an exponential function and from Theorem 3.3, no sufficient quantizer exists. Let $q_0 \sim \mathcal{N}(\underline{\mu}_0, K_0)$ and $q_1 \sim \mathcal{N}(\underline{\mu}_1, K_1)$ where $\underline{\mu}_0$ and $\underline{\mu}_1$ are k-dimensional mean vectors and K_0 and K_1 are $k \times k$ covariance

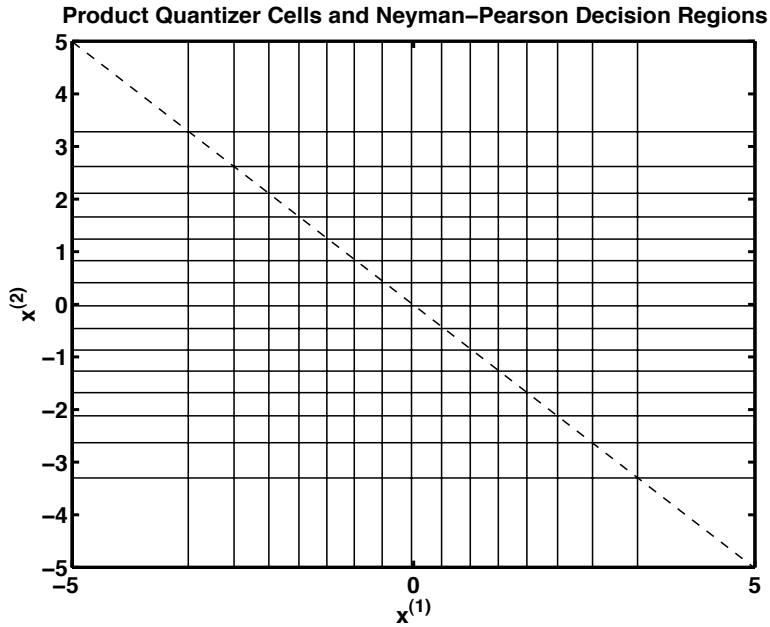

Figure 3.3: Border (dashed line) between Neyman-Pearson regions for Gaussian sources and cells of a two-dimensional product quantizer (solid lines).

matrices. The likelihood ratio is

$$\frac{q_0(x)}{q_1(x)} = \sqrt{\frac{|K_1|}{|K_0|}} \exp\left[-\frac{1}{2}(x-\underline{\mu}_0)^T K_0^{-1}(x-\underline{\mu}_0) - \frac{1}{2}(x-\underline{\mu}_1)^T K_1^{-1}(x-\underline{\mu}_1)\right].$$

Since this function is not piecewise-constant on $\mathbb{R}^k$, no sufficient quantizer exists. Furthermore, for any threshold T with $n > 1$, it can be shown that no Neyman-Pearson quantizer exists.

For example, let $q_0 \sim \mathcal{N}(-\mu, 1)$, $q_1 \sim \mathcal{N}(\mu, 1)$, and $n = 2$. Since the sources are one-dimensional ($k = 1$), the cells of $Q^{(n)}$ are Cartesian products of intervals, or rectangles. However, the Neyman-Pearson regions are

$$\begin{aligned} \mathcal{U}_0^{(n)} &= \left\{\mathbf{x} = [x^{(1)}, x^{(2)}] : x^{(1)} + x^{(2)} > T'\right\} \\ \mathcal{U}_1^{(n)} &= \left\{\mathbf{x} = [x^{(1)}, x^{(2)}] : x^{(1)} + x^{(2)} < T'\right\} \end{aligned}$$

where $T' = -\log T/2\mu$. Thus the Neyman-Pearson regions are separated by a line of slope -1 in the $x^{(1)}, x^{(2)}$ plane. Figure 3.3 shows the border between the Neyman-Pearson regions for $\mu = 2$ and $T = 1$ along with the cells of a product quantizer $Q^{(n)}$. Note that

since the product quantizer must have rectangular cells, some cells must intersect both Neyman-Pearson regions.

3.3.4 Estimation Performance of Lossless Quantizers

The derivations in this section have neglected any consideration of quantizer estimation performance. Indeed, it is quite possible that a lossless quantizer can have very poor estimation performance. For example, a two-cell Neyman-Pearson quantizer can be expected to yield a large reconstruction MSE. To improve estimation performance, one could apply the techniques of Gray *et al.* [29, 50, 53] who consider a Bayes risk weighted rth-power distortion measure. The weighting factor is used to trade detection performance for estimation performance. The weighted distortion is minimized by an iterative descent algorithm.

Another method of improving estimation performance is to refine the lossless quantizer. Let Q and Q' be vector quantizers. The quantizer Q' is a refinement of Q if Q' has more cells than Q and every cell of Q' is a subset of some cell of Q. Now, suppose Q is a lossless quantizer with minimum rate and N cells. The estimation performance of Q can be improved by refining Q. That is, a refinement of Q with $N' > N$ cells may be optimized for estimation performance, perhaps by means of an iterative algorithm. The detection performance will of course remain unchanged. Note that in optimizing the refined quantizer for estimation performance, both the codebook and partition must be considered.

3.4 Lossy Quantizers for Distributed Hypothesis Testing

Theorems 3.3 and 3.5 indicate that lossless quantizers exist only under rather restrictive circumstances. Most sources, including Gaussian sources, do not satisfy the conditions of the theorems. For these sources, quantization always results in a degradation in hypothesis testing performance. In the proceeding sections, we will be concerned with the design of quantizers for these cases. Although there will always be a loss in performance, this loss can be minimized by proper design of the quantizer.

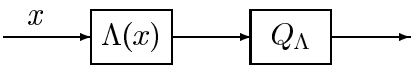

Figure 3.4: A log-likelihood ratio quantizer.

3.4.1 Sequences of Quantizers

Asymptotic, or high-rate quantization analysis is commonly used to obtain interesting insights into the behavior of many-point quantizers. Bennet's integral [26, 46] is the product of such analysis. The most commonly used technique of asymptotic analysis is the sequence approach. This technique has been introduced in Section 3.2.2 and is described in detail in Appendix E where it is used to derive asymptotic losses in discrimination due to quantization. The idea behind the sequence approach is to consider a sequence of quantizers $\{Q_N\}$. Each quantizer in the sequence has associated with it a specific point density, specific inertial profile, specific covariation profile, and diameter function. Assuming the first three sequences of functions converge to functions $\zeta(x)$, $m(x)$, $M(x)$, and that the sequence of diameter functions converges to zero, the limiting behavior of the quantizer sequence can be determined. The resulting formulas can then be used to approximate the behavior of a many-point quantizer by assuming it is part of such a sequence.

3.4.2 Log-Likelihood Ratio Quantizers

The performance of Neyman-Pearson hypothesis tests is unaffected by processing of the observations as long as the processing produces a sufficient statistic. For example, quantization with a sufficient quantizer (if one exists) has no effect on hypothesis testing performance. If a sufficient quantizer does not exist, then it is reasonable to quantize a sufficient statistic, such as the log-likelihood ratio, rather than the raw data, as the required rate may be less than for a quantizer applied to the raw data. It has been determined that this approach is optimal for various detection-related objectives [54, 68]. A *log-likelihood*

ratio quantizer or *LLR quantizer* Q is a function given by

$$Q(x) = Q_\Lambda(\Lambda(x)) \tag{3.26}$$

where $\Lambda(x)$ is the log-likelihood ratio. The scalar quantizer Q_Λ is called the *constituent quantizer* of Q. Figure 3.4 depicts schematically a log-likelihood ratio quantizer. Note that a k-dimensional vector quantizer Q is equivalent to some LLR quantizer if and only if $\Lambda(x) : \mathbb{R}^k \to \mathbb{R}$ is a one-to-one correspondence. Note also that the boundaries of an LLR quantizer correspond to subsets of $\mathbb{R}^k$ on which the log-likelihood ratio is constant (level sets of $\Lambda(x)$). Finally, recall from Section 3.3 that the codebook of the quantizer is inconsequential for hypothesis testing.

Let $q_{\Lambda,0}(l)$ and $q_{\Lambda,1}(l)$ be the probability densities of $\Lambda(x)$ under hypotheses H_0 and H_1, respectively. Consider a sequence of LLR quantizers $\{Q_N\}$ where $Q_N(x) = Q_{N,\Lambda}(\Lambda(x))$ and assume that the sequence of diameter functions associated with the constituent quantizers $Q_{N,\Lambda}$ converges to zero. Then by Bennet's integral, the mean square reconstruction error of the quantized log-likelihood ratio under either hypothesis converges to zero:

$$\begin{aligned} E\left[(l - Q_{N,\Lambda}(l))^2 | H_0\right] &= \int q_{\Lambda,0}(l)(l - Q_{N,\Lambda}(l))^2 dl \to 0 \\ E\left[(l - Q_{N,\Lambda}(l))^2 | H_1\right] &= \int q_{\Lambda,1}(l)(l - Q_{N,\Lambda}(l))^2 dl \to 0. \end{aligned} \tag{3.27}$$

From (3.27), it is evident that the sequence of quantized observations $\{Q_N(x)\}$ converges to the sufficient statistic $\Lambda(x)$ and therefore, the degradation in hypothesis testing performance vanishes as $N \to +\infty$. Therefore, an LLR quantizer whose constituent quantizer has small cells should provide good hypothesis testing performance. In Section 3.6.4, we discuss optimization of the constituent quantizer. On the other hand, it is certainly possible that the reconstruction MSE of the sequence $\{Q_N\}$ does not converge to zero, especially when $\Lambda(x)$ is many-to-one. For example, if $k = 2$ and the sources are $q_0 \sim \mathcal{N}([\mu_0, \mu_0], I)$ and $q_1 \sim \mathcal{N}([\mu_1, \mu_1], I)$, then the constituent quantizer will attempt to preserve the sum of the components of the observation vector. Thus, the cells of the LLR quantizer will be "strips"

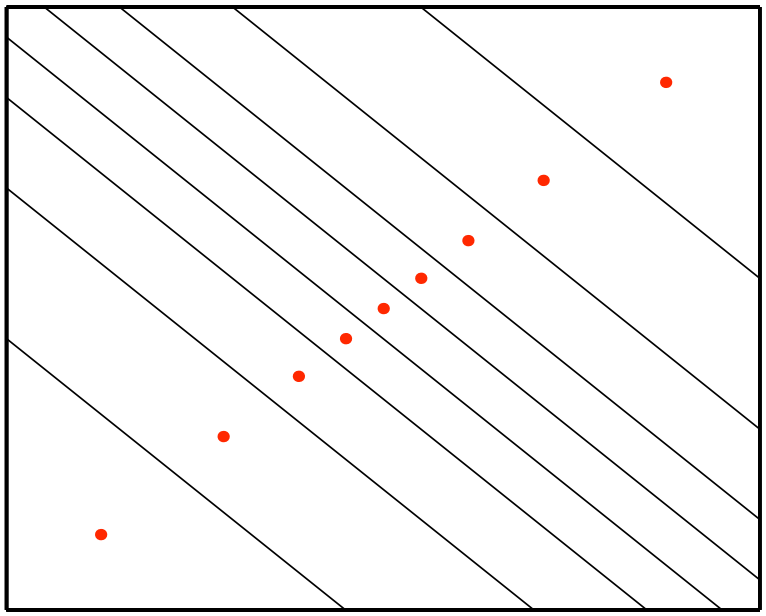

Figure 3.5: Log-likelihood ratio quantizer for two-dimensional Gaussian sources with identity covariance matrices.

of slope -1 as shown in Figure 3.5 and the sequence of diameter functions associated with Q_N does not converge to zero.[3] Clearly, the MSE will not converge to zero either.

3.4.3 Estimation-Optimal Quantizers

Quantizers that minimize Bennet's integral, or estimation-optimal quantizers, have been introduced in Section 3.2.2. Here we simply point out that an estimation-optimal quantizer must yield reasonable hypothesis testing performance as finely-quantized raw data certainly yields high resolution of the sufficient statistic. However, for a given rate, an optimal LLR quantizer should be expected to provide better hypothesis testing performance than an estimation-optimal quantizer, as the LLR quantizer discards any information not relevant for hypothesis testing.

[3]Of course, the sequence of diameter functions associated with the constituent quantizers $Q_{N,\Lambda}$ does converge to zero, as we have assumed.

3.4.4 Small-Cell Quantizers

A many-point quantizer for which most cells have small diameter can be considered to be part of a sequence of quantizers whose diameter functions converge to zero. Such a quantizer will be referred to as a *small-cell quantizer*. Bennet's integral (3.18) is obtained using the sequence approach and is therefore only applicable to small-cell quantizers. An estimation-optimal quantizer is a small-cell quantizer, whereas an LLR quantizer may or may not be a small-cell quantizer. Note also that any small-cell quantizer should have reasonable estimation performance, as the reconstruction MSE is proportional to $N^{-2/k}$. In the next section, we introduce quantizers that are optimized for hypothesis testing performance under a small-cell constraint.

3.5 Asymptotic Analysis of Quantization for Hypothesis Testing

In Section 3.2.1, the discriminations $L(q_0||q_1)$, $L(q_1||q_0)$, $L(q_\lambda||q_0)$, and $L(q_\lambda||q_1)$ were shown to play an important role in the determination of hypothesis testing performance. In general, a Neyman-Pearson hypothesis test performs better as these quantities are increased. In this section, we analyze the asymptotic behavior of these discriminations with quantized data. Throughout the analysis we assume that the quantizers are small-cell quantizers. We also assume that the quantizers' codebook points are the centroids of their cells (see Appendix E). Since a quantizer's codebook does not affect its hypothesis testing performance, this assumption is not restrictive.

Note that the small-cell assumption precludes consideration of LLR quantizers for many cases. However, several beneficial optimization procedures emerge from this analysis. First, we are able to obtain the optimal LLR quantizer by optimizing the constituent quantizer. Secondly, mixed detection-estimation objectives can be optimized with the small-cell assumption. These procedures and several more will be discussed in Section 3.6.

3.5.1 Asymptotic Discrimination Losses

Below, we give formulas for the loss in discrimination between the two sources due to quantization and for the loss in discrimination between each source and the tilted source.

Loss in Discrimination between Sources

In Appendix E.1, we determine the asymptotic loss in discrimination between the two sources due to quantization following the sequence approach. Thus, we consider a sequence of quantizers $\{Q_N\}$ whose diameter functions converge to zero and define the discrimination between the sources quantized with the Nth quantizer as $\hat{L}_N = L(\bar{q}_{0,N} \| \bar{q}_{1,N})$ where $\bar{q}_{0,N}$ and $\bar{q}_{1,N}$ are the pmf's of the quantized sources (see Appendix E.1). The loss in discrimination incurred by quantization with the Nth quantizer is thus $\Delta L_N = L - \hat{L}_N$. The asymptotic loss, or distortion, is given by

$$\begin{aligned} \lim_{N \to +\infty} N^{2/k} \Delta L_N &= \frac{1}{2} \int \frac{q_0(x)}{\zeta(x)^{2/k}} \mathrm{tr}(F(x)M(x)) dx \\ &= \frac{1}{2} \int \frac{q_0(x)\mathcal{F}(x)}{\zeta(x)^{2/k}} dx \end{aligned} \tag{3.28}$$

where

$$\mathcal{F}(x) = \nabla\Lambda(x)^T M(x) \nabla\Lambda(x) \tag{3.29}$$

is the *Fisher covariation profile* (See Section 3.5.2). For large N, we can thus write

$$\Delta L_N \approx \frac{1}{2N^{2/k}} \int \frac{q_0(x)\mathcal{F}(x)}{\zeta(x)^{2/k}} dx. \tag{3.30}$$

Note that as the objective function ΔL_N is asymmetric in q_0 and q_1, the asymptotic loss (3.30) is also asymmetric. The formula (3.30) will be used in Section 3.6.1 to derive discrimination-optimal quantizers.

Loss in Discrimination between Tilted Density and Source Densities

To determine the effect of quantization on the asymptotic probabilities of type I and II errors, we use (3.15) to show that for n large

$$\begin{aligned} \hat{\alpha} &\approx e^{-nL(\hat{q}_{\lambda,N}||\bar{q}_{0,N})} \\ \hat{\beta} &\approx e^{-nL(\hat{q}_{\lambda,N}||\bar{q}_{1,N})} \end{aligned} \tag{3.31}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the probabilities of type I and II errors after quantization by an N-point quantizer and $\hat{q}_{\lambda,N}$ is the quantized tilted mass function whose probabilities for $i = 1, \ldots, N$ are given by

$$\hat{q}_{\lambda,N,i} = \frac{\bar{q}_{0,N,i}^{1-\lambda} \cdot \bar{q}_{1,N,i}^{\lambda}}{\sum_{j=1}^{N} \bar{q}_{0,N,j}^{1-\lambda} \cdot \bar{q}_{1,N,j}^{\lambda}}. \tag{3.32}$$

Of course (3.31) assumes that the hypothesis test with quantized data is a Neyman-Pearson test. To obtain the error exponents in (3.31), define the discrimination losses $\Delta L_{0,N} = L(q_\lambda||q_0) - L(\hat{q}_{\lambda,N}||\bar{q}_{0,N})$ and $\Delta L_{1,N} = L(q_\lambda||q_1) - L(\hat{q}_{\lambda,N}||\bar{q}_{1,N})$. Just as Stein's lemma provides motivation to minimize $L(q_0||q_1) - L(\bar{q}_{0,N}||\bar{q}_{1,N})$, equation (3.31) motivates us to minimize $\Delta L_{0,N}$ and $\Delta L_{1,N}$. In Appendix E.2, we again use a sequence approach to obtain $\Delta L_{0,N}$ and $\Delta L_{1,N}$:

$$\lim_{N\to+\infty} N^{2/k} \Delta L_{0,N} = \frac{1}{2} \int \frac{q_\lambda(x)\mathcal{F}(x)}{\zeta(x)^{2/k}} \left[\lambda^2 + \lambda(1-\lambda)(L(q_\lambda||q_0) - \Lambda_0(x))\right] dx \tag{3.33}$$

$$\lim_{N\to+\infty} N^{2/k} \Delta L_{1,N} = \frac{1}{2} \int \frac{q_\lambda(x)\mathcal{F}(x)}{\zeta(x)^{2/k}} \left[(1-\lambda)^2 + \lambda(1-\lambda)(L(q_\lambda||q_1) - \Lambda_1(x))\right] dx \tag{3.34}$$

where

$$\Lambda_0(x) = \log \frac{q_\lambda(x)}{q_0(x)}, \quad \text{and} \quad \Lambda_1(x) = \log \frac{q_\lambda(x)}{q_1(x)}. \tag{3.35}$$

In Section 3.6.3, we use the asymptotic losses given in (3.33) and (3.34) to design optimal vector quantizers. In the remainder of this section, we discuss some important features of the asymptotic loss formulas.

3.5.2 Fisher Covariation Profile

The Fisher covariation profile is so named for its relation to the Fisher information matrices of the source densities q_0 and q_1. The Fisher information of source q_0 is $E[\mathcal{I}_0(x)|H_0]$ and the Fisher information of q_1 is $E[\mathcal{I}_1(x)|H_1]$ [21] where $\mathcal{I}_0(x)$ and $\mathcal{I}_1(x)$ are given by

$$\begin{aligned} \mathcal{I}_0(x) &= \nabla \log q_0(x) \nabla \log q_0(x)^T \\ \mathcal{I}_1(x) &= \nabla \log q_1(x) \nabla \log q_1(x)^T. \end{aligned} \tag{3.36}$$

From (3.29) and (3.36) we have

$$\mathcal{F}(x) = \mathrm{tr}(M(x)(\mathcal{I}_0(x) + \mathcal{I}_1(x)) - 2\sqrt{\mathrm{tr}(\mathcal{I}_0(x)M(x)^2\mathcal{I}_1(x))}.$$

Quantizers with Ellipsoidal Cells

We define an ellipsoidal cell to be a quantizer cell with an ellipsoidal boundary. If a quantizer's cells are ellipsoidal, some interesting properties of the covariation profile and Fisher covariation profile emerge. The following theorem relates the covariation profile on such a cell to a quadratic form associated with the cell's boundary.

Theorem 3.6 *Let S be an ellipsoidal quantizer cell and let R be a symmetric positive definite matrix such that the boundary of S is a level set of the quadratic form x^TRx. Let the cell's codebook point be its centroid and let M be the specific covariation profile of a point in S. Then $M = \gamma R^{-1}$ where $\gamma > 0$.*

Proof: Without loss of generality, assume S is centered at the origin. The matrix R can be orthogonally diagonalized and represented by $R = U\Psi U^T$ where $\Psi = \mathrm{diag}\{\psi_1, \ldots, \psi_k\}$, $\psi_1, \ldots, \psi_k$ are the (positive) eigenvalues of R, and U is a matrix whose columns are the orthonormal eigenvectors of R. The cell S can be written $S = \{x \in \mathbb{R}^k : x^TRx \leq c\}$ where $c > 0$. The covariation profile is

$$M = \frac{1}{V(S)^{1+2/k}} \int_S xx^T dx.$$

Using the change of variables $y = U^T x$, this becomes

$$M = \frac{1}{V(S)^{1+2/k}} U \int_{S'} y y^T dy U^T$$

where $S' = \{y : y^T \Psi y \leq c\}$. Again using a change of variables $z = \sqrt{\Psi} y$, we get

$$M = \frac{1}{V(S)^{1+2/k}} U \sqrt{\Psi}^{-1} \int_{S''} z z^T dz \sqrt{\Psi}^{-1} \left| \sqrt{\Psi}^{-1} \right| U^T$$

where S'' is a k-dimensional sphere of radius $\sqrt{c}$. Since the covariation profile of a sphere is a positive identity, the theorem is proven. □

3.5.3 Discriminability

When $k = 1$ or $M \propto I$, the Fisher covariation profile is proportional to a function that we call the *discriminability*. This function is the square magnitude of the log-likelihood ratio gradient:

$$D(q_0, q_1, x) = \|\nabla \Lambda(x)\|^2 \tag{3.37}$$

where $\Lambda(x) = \log(q_0(x)/q_1(x))$. Essentially, $D(q_0, q_1, x)$ is a measure of the usefulness of x in deciding between the two hypotheses H_0 and H_1. Note that $D(x)$ can be written

$$D(q_0, q_1, x) = \left\| \frac{\nabla q_0(x)}{q_0(x)} - \frac{\nabla q_1(x)}{q_1(x)} \right\|^2. \tag{3.38}$$

Equation (3.38) indicates that if the two source densities q_0 and q_1 are equal and have equal gradients at a point x, then the discriminability at x is zero. It is somewhat intuitive that a function measuring the usefulness of an observation for hypothesis testing should be zero in such a situation. However, this is certainly not the only case in which the discriminability is zero. The following theorem shows that when the discriminability is zero on a region, the discrimination (Kullback-Leibler distance) between q_0 and q_1 is minimized.

Theorem 3.7 *Let A be a connected subset of $\mathbb{R}^k$ and let q_0 and q_1 be probability densities on $\mathbb{R}^k$ such that $D(q_0, q_1, x) = 0$ on A. Let $\mathcal{Q}_1$ be the family of densities that are equal to q_1 on $\mathbb{R}^k \setminus A$. Then*

$$L(q_0||q_1) = \inf_{p_1 \in \mathcal{Q}_1} L(q_0||p_1).$$

Proof: From (3.37), since $D(q_0, q_1, x) = 0$ on A, the log-likelihood ratio must be constant on A. Let this constant be c. Then $q_0(x) = e^c q_1(x)$ on A. Let $p_1 \in \mathcal{Q}_1$. Then

$$\begin{aligned} L(q_0||p_1) - L(q_0||q_1) &= \int_A q_0(x) \log \frac{q_0(x)}{p_1(x)} dx - \int_A q_0(x) \log \frac{q_0(x)}{q_1(x)} dx \\ &= \int_A q_0(x) \log \left[\frac{q_0(x)}{p_1(x)} \cdot \frac{q_1(x)}{q_0(x)} \right] dx \\ &= \int_A q_0(x) \log \left[\frac{q_0(x)}{p_1(x)} e^{-c} \right] dx \\ &= \int_A q_0(x) \log \frac{q_0(x)}{p_1(x)} dx - c P_{A|0} \end{aligned}$$

where $P_{A|0} = \int_A q_0(x) dx$. Similarly, define $P_{A|1} = \int_A q_1(x) dx = \int_A p_1(x) dx$ and note that $P_{A|0} = e^c P_{A|1}$. Next, define the densities

$$\begin{aligned} q_{A,0}(x) &= \frac{q_0(x)}{P_{A|0}} I_A(x) \\ p_{A,1}(x) &= \frac{p_1(x)}{P_{A|1}} I_A(x) \end{aligned}$$

where $I_A(x)$ is the indicator function of set A. Continuing, we have

$$\begin{aligned} L(q_0||p_1) - L(q_0||q_1) &= P_{A|0} \int q_{A,0}(x) \log \left[\frac{q_{0,A}(x)}{p_{A,1}(x)} \cdot \frac{P_{A|0}}{P_{A|1}} \right] dx - c P_{A|0} \\ &= P_{A|0} \left(L(q_{A,0}||p_{A,1}) + \log \frac{P_{A|0}}{P_{A|1}} - c \right) \\ &\geq 0 \end{aligned} \tag{3.39}$$

where we have used the fact that discrimination is never negative [10]. The inequality in (3.39) is an equality if and only if $q_{A,0}(x) = p_{A,1}(x)$ on A. This is equivalent to $q_0(x) = e^c p_1(x)$ or $p_1(x) = q_1(x)$ on A. Thus, the theorem is proven. □

3.5.4 Comparison to Bennet's Integral

Equation (3.28) indicates that the loss in discrimination due to quantization by a sequence of N-point, small-cell VQ's converges to zero at the rate of $N^{2/k}$. This is the same rate of convergence of the reconstruction MSE given by Bennet's integral (3.18). Note also the similarity between the distortion formulas (3.33), (3.34), and (3.18).

3.6 Optimal Small-Cell Quantizers for Hypothesis Testing

In this section, we use the asymptotic discrimination losses derived in Section 3.5 to optimize small-cell quantizers for hypothesis testing performance. The optimal quantizers are characterized by their point densities and covariation profiles. We restrict attention to those small-cell quantizers with congruent cells and those with ellipsoidal cells.

3.6.1 Maximum Discrimination

Motivated by Stein's lemma, we seek to maximize the discrimination between the sources q_0 and q_1 after undergoing quantization. To optimize the VQ with respect to asymptotic discrimination loss, as given by (3.30), it is necessary to jointly optimize two functions, namely the point density $\zeta(x)$ and the covariation profile $M(x)$. First, the discrimination-optimal point density can be obtained in a manner similar to that for the estimation-optimal quantizer:

$$\zeta^d(x) = \frac{[q_0(x)\mathcal{F}(x)]^{\frac{k}{k+2}}}{\int [q_0(y)\mathcal{F}(y)]^{\frac{k}{k+2}} dy}. \tag{3.40}$$

The discrimination loss with the optimal point density is then

$$\Delta L_N \approx \frac{1}{2N^{2/k}} \left(\int [q_0(x)\mathcal{F}(x)]^{\frac{k}{k+2}} dx \right)^{\frac{k+2}{k}}. \tag{3.41}$$

We present two optimization techniques. The first assumes the quantizer has congruent cells with minimum moment of inertia. The second assumes ellipsoidal cells.

Congruent Cells

If the quantizer's cells are congruent, the covariation profile $M(x)$ and Fisher covariation profile $\mathcal{F}(x)$ are constant and the point density given by equation (3.40) completely characterizes the optimal quantizer. Furthermore, assuming the congruent cells have minimum moment of inertia, and thus the same shape as the cell found in the estimation-optimal quantizer, the covariation profile is a positive identity matrix and the optimal point density can be written in terms of the discriminability function:

$$\zeta^d(x) = \frac{[q_0(x)\|\nabla\Lambda(x)\|^2]^{\frac{k}{k+2}}}{\int [q_0(y)\|\nabla\Lambda(y)\|^2]^{\frac{k}{k+2}} dy}. \tag{3.42}$$

Ellipsoidal Cells

Ellipsoidal cells can not cover $\mathbb{R}^k$ without overlap and thus can not partition $\mathbb{R}^k$. However, as $N \to +\infty$ it is possible that a quantizer's cells can be close to ellipsoidal. Studying this type of quantizer yields important insights.

With the ellipsoidal-cell assumption, we have some control over the covariation profile $M(x)$ which we use to minimize the discrimination loss (3.41). Since $M(x)$ is symmetric and positive definite (see Appendix E), it can be spectrally decomposed:

$$M(x) = \sum_{i=1}^{k} \phi_i(x) v_i(x) v_i(x)^T$$

where $\{\phi_1(x), \ldots, \phi_k(x)\}$ are the positive eigenvalues of $M(x)$ corresponding to orthonormal eigenvectors $\{v_1(x), \ldots, v_k(x)\}$. Thus the Fisher covariation profile is

$$\mathcal{F}(x) = \sum_{i=1}^{k} \phi_i(x) \left(\nabla\Lambda(x)^T v_i(x)\right)^2.$$

Now, minimization of ΔL_N involves an optimization of the eigenvalues and eigenvectors of the covariation profile matrix at each point x under constraints imposed by the total number of cells N. We take a simplified approach and minimize $\mathcal{F}(x)$ with no constraints. Accordingly, the eigenvector corresponding to the minimum eigenvalue of the optimal covariation profile matrix should be parallel to $\nabla\Lambda(x)$. All other eigenvectors will be orthogonal

to $\nabla\Lambda(x)$ and the Fisher covariation profile becomes

$$\mathcal{F}(x) = \phi_{\min}(x)||\nabla\Lambda(x)||^2$$

where $\phi_{\min}(x)$ is the minimum eigenvalue of $M(x)$ and the corresponding eigenvector is $v_{\min} = \nabla\Lambda(x)/||\nabla\Lambda(x)||$.

Now, let S denote the ellipsoidal cell that contains x. From Theorem 3.6, there is a matrix R and a scalar $\gamma > 0$ such that the boundary of S is a level set of the quadratic form associated with R and $M(x) = \gamma R^{-1}$. Since $M(x)$ and R are inversely related, the eigenvector of R corresponding to its maximum eigenvalue is parallel to $\nabla\Lambda(x)$ and all other eigenvectors are orthogonal to $\nabla\Lambda(x)$. Therefore, the minor axis of cell S is parallel to the gradient of the log-likelihood ratio. Finally, since the gradient of a function is always orthogonal to the function's level sets, it follows that the minor axis of cell S is orthogonal to the level set of Λ at x. For large N, we see that this implies that the ellipsoidal cells should be aligned with the level sets of the log-likelihood ratio. An example for $k = 2$ is shown in Figure 3.6.

The above arguments suggest that as $N \to +\infty$, the log-likelihood ratio is approximately constant on each cell of the discrimination-optimal quantizer. Therefore, the discrimination-optimal quantizer essentially quantizes the log-likelihood ratio, as does a log-likelihood ratio quantizer. This suggests that the best hypothesis testing performance of a quantizer with a given rate is achieved by an LLR quantizer, since an LLR quantizer preserves the log-likelihood ratio without sacrificing rate due to a small-cell assumption. In Section 3.6.3, we will see that the quantizer that yields the best ROC curve also preserves the log-likelihood ratio, again pointing to the optimality of LLR quantizers for hypothesis testing. In Section 3.6.5, a joint detection-estimation criterion will be considered for which the small-cell assumption must be invoked and the LLR quantizer may not be considered.

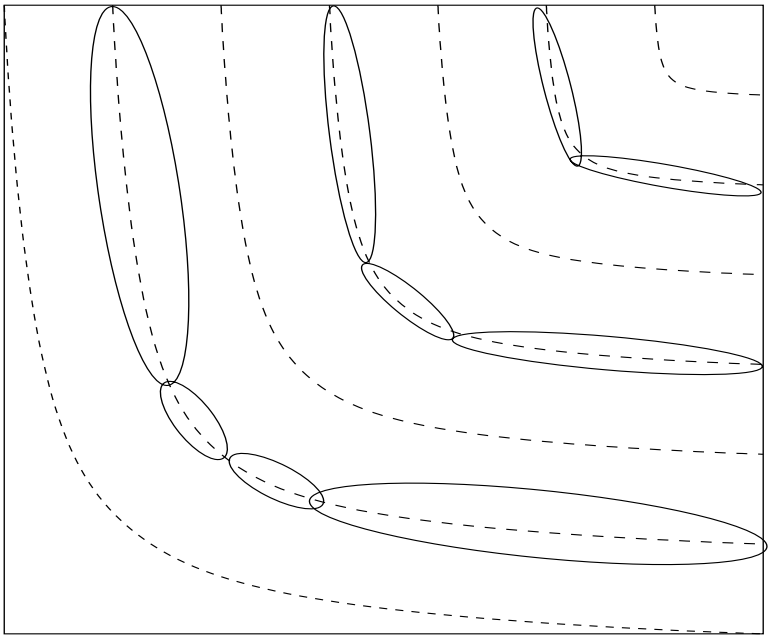

Figure 3.6: Contours of log-likelihood ratio $\Lambda(x)$ (dashed lines) and some cells of an optimal ellipsoidal-cell quantizer.

Shortcomings of Maximum Discrimination Objective

Although Stein's lemma justifies maximization of the error exponent $L(q_0||q_1)$ after quantization, it says nothing of the asymptotic performance of a Neyman-Pearson test with a given threshold, say T. The error exponent in Stein's lemma is applicable only when for each n, the Neyman-Pearson test with the minimum probability of miss meeting the false alarm constraint $\alpha \leq \alpha^*$ is used. Furthermore, the error exponent is independent of the actual constraint as long as $\alpha^* \in (0, 1)$. Therefore, Stein's lemma can not be used to determine the asymptotic behavior of both α and β. Nor can it provide insight into the asymptotic characteristics of the ROC curve.

3.6.2 Maximum Power

The error exponent formulas (3.15) give the asymptotic values of both α and β as functions of the Neyman-Pearson threshold, which determines the parameter λ. In this section, we use these formulas to show how to design quantizers optimal for detection probability with a false alarm constraint. This optimization is difficult and therefore in the

next section a more tractable optimization method is presented for maximizing the area under the ROC curve. Since the probabilities of type I and type II errors are dependent on the discriminations between the tilted source and the actual sources, we first make the following definitions for notational convenience:

$$\begin{aligned} L_0(\lambda) &= L(q_\lambda \| q_0) \\ L_1(\lambda) &= L(q_\lambda \| q_1) \\ \hat{L}_0(\lambda, \zeta) &= L(\hat{q}_{\lambda,N} \| \bar{q}_0) \\ \hat{L}_1(\lambda, \zeta) &= L(\hat{q}_{\lambda,N} \| \bar{q}_1) \\ \Delta L_0(\lambda, \zeta) &= \Delta L_{0,N} \\ \Delta L_1(\lambda, \zeta) &= \Delta L_{1,N}. \end{aligned} \tag{3.43}$$

From equations (3.33) and (3.34), it is evident that one may choose the point density $\zeta(x)$ and covariation profile $M(x)$ of the vector quantizer to minimize either ΔL_0 or ΔL_1. Correspondingly, the increase in false alarm probability or probability of miss is minimized. However, minimizing one of these quantities will undoubtedly result in an unacceptably large value of the other. Instead, we focus on the two optimality criteria discussed in Section 3.2.1. Here we discuss maximization of the probability of detection after quantization $1-\hat{\beta}$ under a constraint on the false alarm probability $\hat{\alpha}$. In Section 3.6.3, we discuss maximization of the area under the ROC curve after quantization.

Accordingly, suppose we are given the maximum allowable value of $\hat{\alpha}$, say $\hat{\alpha}^*$. From (3.31), assuming a large number of observations n, we can express the probabilities of type I and type II errors after quantization as

$$\begin{aligned} \hat{\alpha}(\lambda, \zeta) &\approx e^{-n\hat{L}_0(\lambda,\zeta)} \\ \hat{\beta}(\lambda, \zeta) &\approx e^{-n\hat{L}_1(\lambda,\zeta)}. \end{aligned}$$

The problem, then, is to choose λ^* and ζ^* such that

$$(\lambda^*, \zeta^*) = \underset{(\lambda,\zeta)\in\Phi}{\operatorname{argmin}}\, \hat{\beta}(\lambda, \zeta) \tag{3.44}$$

where

$$\Phi = \{(\lambda, \zeta) : \hat{\alpha}(\lambda, \zeta) \leq \hat{\alpha}^*\}.$$

Note that the problem may be stated equivalently as

$$(\lambda^*, \zeta^*) = \underset{(\lambda,\zeta)\in\Phi}{\operatorname{argmax}} \hat{L}_1(\lambda, \zeta) \tag{3.45}$$

and the set Φ can be expressed as

$$\Phi = \left\{(\lambda, \zeta) : \hat{L}_0(\lambda, \zeta) \geq -\frac{1}{n}\log\hat{\alpha}^*\right\}.$$

To derive the optimal quantizer, (3.44) indicates that we must optimize over λ and ζ. As the latter is a function, this is not a simple task. In the next subsection, a more tractable optimization method is presented.

3.6.3 Maximum Area under ROC Curve

Here we adopt the area under the ROC curve with quantized data as the objective function that we seek to maximize. In so doing, we are effectively seeking a threshold-independent quantizer that yields a family of Neyman-Pearson tests (indexed by λ) acting on quantized data that is optimal in an average sense.

We first note that the optimal hypothesis test with the n observations $\mathbf{x} = [x^{(1)}, \ldots, x^{(n)}]$, each having been quantized by the N-cell quantizer Q, is a Neyman-Pearson test of the form

$$\log \frac{\bar{q}_0^{(n)}\left(Q^{(n)}(\mathbf{x})\right)}{\bar{q}_1^{(n)}\left(Q^{(n)}(\mathbf{x})\right)} \underset{H_1}{\overset{H_0}{\gtrless}} nT \tag{3.46}$$

where

$$\bar{q}_0^{(n)}\big(Q^{(n)}(\mathbf{x})\big) = \prod_{i=1}^{n} \bar{q}_0\big(Q(x^{(i)})\big), \qquad \bar{q}_1^{(n)}\big(Q^{(n)}(\mathbf{x})\big) = \prod_{i=1}^{n} \bar{q}_1\big(Q(x^{(i)})\big)$$

and $\bar{q}_0(\cdot)$, $\bar{q}_1(\cdot)$ are the (k-dimensional) pmf's of a single quantized observation under hypotheses H_0 and H_1, respectively. Clearly, the log-likelihood ratio in (3.46) can take on at most N^n values. Thus, the ROC curve of the Neyman-Pearson test will consist of at

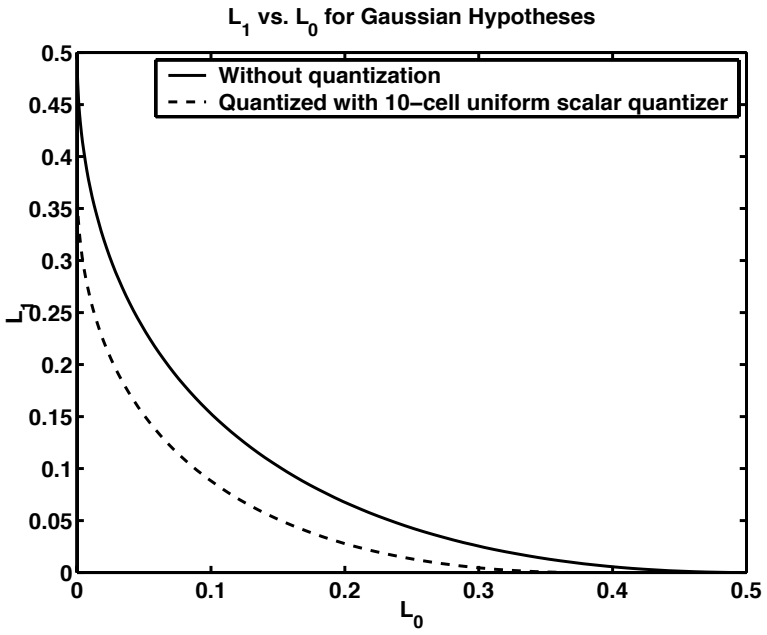

Figure 3.7: Example of $L_1(L_0)$ and $\hat{L}_1(\hat{L}_0)$.

most N^n pairs $(\hat{\alpha}, 1-\hat{\beta})$. However, a continuous ROC curve may be obtained by assuming a randomized Neyman-Pearson test [35]. The $(\hat{\alpha}, 1-\hat{\beta})$ pairs of the non-randomized test will be joined by linear segments in the randomized test. As $n \to +\infty$, the ROC curve of the randomized test becomes a smooth function parameterized by λ. From (3.15), $\hat{\alpha}$ and $\hat{\beta}$ are decreasing functions of $\hat{L}_0$ and $\hat{L}_1$, respectively, for n large. Therefore, we instead maximize the area under the curve $\hat{L}_1(\hat{L}_0)$. The optimal quantizer will be referred to as the ROC-optimal quantizer. The $\hat{L}_1(\hat{L}_0)$ curve can be represented parametrically since $\hat{L}_1$ and $\hat{L}_0$ are both functions of λ. When $\lambda = 0$, $q_\lambda(x) = q_0(x)$. Similarly, when $\lambda = 1$, $q_\lambda(x) = q_1(x)$. Thus we have

$$\begin{aligned}
\hat{L}_0(0) &= 0 \\
\hat{L}_1(0) &= L(\bar{q}_{0,N} \| \bar{q}_{1,N}) \\
\hat{L}_0(1) &= L(\bar{q}_{1,N} \| \bar{q}_{0,N}) \\
\hat{L}_1(1) &= 0.
\end{aligned}$$

Figure 3.7 shows an example of the curve $L_1(L_0)$ for $q_0 \sim \mathcal{N}(0,1)$ and $q_1 \sim \mathcal{N}(1,1)$. Also

shown is the curve $\hat{L}_1(\hat{L}_0)$ when the quantizer is a uniform scalar quantizer with ten cells and support $[-10, 10]$. Note that $L(\bar{q}_{0,N}||\bar{q}_{1,N})$ and $L(\bar{q}_{1,N}||\bar{q}_{0,N})$ may be obtained by using (3.28) for large N. Let $\hat{A}$ be the area under the curve $\hat{L}_1(\hat{L}_0)$. Then

$$\begin{aligned}\hat{A} &= \int_0^{L(\bar{q}_{1,N}||\bar{q}_{0,N})} \hat{L}_1(\hat{L}_0) d\hat{L}_0 \\ &= \int_0^1 \hat{L}_1(\lambda) \frac{d}{d\lambda} \hat{L}_0(\lambda) d\lambda.\end{aligned}$$

Thus we seek the functions

$$\{\zeta^o, M^o\} = \text{argmax } \hat{A}.$$

Derivation of $\hat{A}$

To derive $\hat{A}$, we first define

$$\begin{aligned}f_0(x,\lambda) &= q_\lambda(x) \left[\lambda^2 + \lambda(1-\lambda)(L_0(\lambda) - \Lambda_0(x,\lambda))\right] \\ f_1(x,\lambda) &= q_\lambda(x) \left[(1-\lambda)^2 + \lambda(1-\lambda)(L_1(\lambda) - \Lambda_1(x,\lambda))\right]. \end{aligned} \tag{3.47}$$

Then we can write

$$\begin{aligned}\hat{L}_0(\lambda) &= L_0(\lambda) - \frac{1}{2N^{2/k}} \int \frac{\mathcal{F}(x)}{\zeta(x)^{2/k}} f_0(x,\lambda) dx \\ \hat{L}_1(\lambda) &= L_1(\lambda) - \frac{1}{2N^{2/k}} \int \frac{\mathcal{F}(x)}{\zeta(x)^{2/k}} f_1(x,\lambda) dx\end{aligned}$$

and

$$\frac{d}{d\lambda}\hat{L}_0(\lambda) = \frac{d}{d\lambda} L_0(\lambda) - \frac{1}{2N^{2/k}} \int \frac{\mathcal{F}(x)}{\zeta(x)^{2/k}} \cdot \frac{\partial}{\partial\lambda} f_0(x,\lambda) dx.$$

Thus

$$\begin{aligned}\hat{L}_1(\lambda)\frac{d}{d\lambda}\hat{L}_0(\lambda) &= L_1(\lambda)\frac{d}{d\lambda}L_0(\lambda) - \\ &\quad \frac{1}{2N^{2/k}} \int \frac{\mathcal{F}(x)}{\zeta(x)^{2/k}} \left[L_1(\lambda)\frac{\partial}{\partial\lambda} f_0(x,\lambda) + f_1(x,\lambda)\frac{d}{d\lambda}L_0(\lambda)\right] dx + \\ &\quad o\left(\frac{1}{N^{2/k}}\right).\end{aligned}$$

The area $\hat{A}$ is thus

$$\hat{A} = A - \frac{1}{2N^{2/k}} \int \frac{\mathcal{F}(x)\eta(x)}{\zeta(x)^{2/k}} dx + o\left(\frac{1}{N^{2/k}}\right)$$

where

$$A = \int_0^1 L_1(\lambda) \frac{d}{d\lambda} L_0(\lambda) d\lambda$$

is the area under the curve $L_1(L_0)$ and

$$\eta(x) = \int_0^1 \left[L_1(\lambda) \frac{\partial}{\partial \lambda} f_0(x, \lambda) + f_1(x, \lambda) \frac{d}{d\lambda} L_0(\lambda) \right] d\lambda. \tag{3.48}$$

Finally, we can write

$$\lim_{N \to +\infty} N^{2/k} (A - \hat{A}) = \frac{1}{2} \int \frac{\mathcal{F}(x) \eta(x)}{\zeta(x)^{2/k}} dx. \tag{3.49}$$

Note the resemblance of (3.49) to (3.28). Essentially, the source density $q_0(x)$ in (3.28) has been replaced by $\eta(x)$ in (3.49). Note from (3.48) that $\eta(x)$ is independent of the quantizer. Although $\eta(x)$ is difficult to calculate, we determine this function numerically for some example sources in Section 3.7. For these examples, $\eta(x)$ is always positive. Thus $\eta(x)$ can be thought of as a density that has been averaged (over λ).

Now the quantizer optimization procedures described in Section 3.6.1 can be applied here with $q_0(x)$ replaced by $\eta(x)$. The ROC-optimal point density is

$$\zeta^o(x) = \frac{[\mathcal{F}(x)\eta(x)]^{\frac{k}{k+2}}}{\int [\mathcal{F}(y)\eta(y)]^{\frac{k}{k+2}} dy} \tag{3.50}$$

and the resulting loss in area under the $L_1(L_0)$ curve, with the optimal point density is

$$\Delta A_N \approx \frac{1}{2N^{2/k}} \left(\int [\mathcal{F}(x)\eta(x)]^{\frac{k}{k+2}} dx \right)^{\frac{k+2}{k}}. \tag{3.51}$$

Next, we focus on congruent-cell and ellipsoidal-cell quantizers as in Section 3.6.1. The congruent-cell quantizer is completely characterized by the optimal point density (3.50) which, in the case of minimum-moment-of-inertia cells is given by

$$\zeta^o(x) = \frac{[\eta(x) \|\nabla \Lambda(x)\|^2]^{\frac{k}{k+2}}}{\int [\eta(y) \|\nabla \Lambda(y)\|^2]^{\frac{k}{k+2}} dy}. \tag{3.52}$$

Based on the resemblance of (3.51) to (3.28) we see that, in the case of ellipsoidal cells, the arguments of Section 3.6.1 are once again applicable. The ROC-optimal, ellipsoidal-cell quantizer will preserve the log-likelihood ratio. This conclusion again suggests optimality of LLR quantizers for hypothesis testing.

3.6.4 Optimal Log-Likelihood Ratio Quantizers

Next we show how to use the asymptotic small-cell theory to optimize the constituent quantizer of an LLR quantizer. The resulting LLR quantizer will provide the best hypothesis testing performance of all the quantizers considered, but because of potentially large cells in the LLR quantizer, its estimation performance will in general be poor.

As in Section 3.4.2, let $q_{\Lambda,0}(l)$ and $q_{\Lambda,1}(l)$ be the probability densities of $\Lambda(x)$ under hypotheses H_0 and H_1, respectively. Let Q be a log-likelihood ratio quantizer with constituent quantizer Q_Λ. The optimization procedure of Section 3.6.3 can be used to derive an optimal point density $\zeta_\Lambda(l)$ of Q_Λ. This procedure will be demonstrated in Section 3.7.2 for Gaussian sources.

3.6.5 Mixed Objective Function

Here we consider a mixed detection-estimation objective. When detection performance and reconstruction MSE are both optimization criteria, the techniques of Gray *et al.* [29, 50, 53] can be used to iteratively optimize a quantizer as described in Section 3.3.4. We can also use the asymptotic optimization methods of this section to design quantizers with mixed detection-estimation objectives. This permits an understanding of the features of the optimal quantizer through its point density function. The minimization criterion is a weighted combination of the reconstruction MSE and the area loss given by (3.49). Note that the small-cell assumption is necessary if the reconstruction MSE is to be made small. Thus we assume small cells. We also focus on congruent-cell quantizers and optimize only the point density.

To derive the mixed objective function, we begin with the Bayes risk weighted distortion used in [50]:

$$J_1 = c(D + rB)$$

where $c, r \in [0, +\infty)$, D is the reconstruction MSE, and B is Bayes risk. In [50], $c = 1$ was

assumed without loss of generality. The objective J_1 is equivalent to the following convex combination:

$$J_2 = (1-\rho)B + \rho D, \quad \rho \in [0,1]$$

with $r = (1-\rho)/\rho$ and $c = \rho$. In [50], an iterative algorithm was developed to minimize J_2. It was found that (for $n = 1$ observation) as $\rho \to 0$, the iteratively-optimized quantizer converges to a Neyman-Pearson quantizer for threshold T (where T is the threshold in the Bayes test corresponding to the chosen Bayes risk). As $\rho \to 1$, the quantizer converges to the small-cell estimation-optimal quantizer that results from the Lloyd optimization, described in Appendix F.

As we are considering distributed hypothesis testing environments with many observations, we use an asymptotic objective function analogous to J_2. Let the Bayes risk be equivalent to the total probability of error P_e. Then, for n large

$$B = P_e \approx e^{-nC}$$

where C is the Chernoff information given by (3.16). Thus, it is reasonable to replace the Bayes risk term in J_2 with the loss ΔC in Chernoff information due to quantization. This yields the following objective:

$$J_3 = (1-\rho)\Delta C + \rho D.$$

Without knowledge of the priors, or equivalently the Neyman-Pearson threshold, we require a threshold-independent detection objective. The ROC objective of Section 3.6.3 is an excellent choice. We therefore replace ΔC with ΔA, the loss in area under the $L_1(L_0)$ curve, to get

$$J_4 = (1-\rho)\Delta A + \rho D.$$

Using the asymptotic formulas (3.18) and (3.49) we get the following objective for many-point congruent-cell quantizers with minimum moment of inertia:

$$J = \int \frac{\rho q(x) + (1-\rho)p(x)}{\zeta(x)^{2/k}} dx \tag{3.53}$$

where the "density" $p(x)$ is given by

$$p(x) = \frac{\eta(x)\|\nabla\Lambda(x)\|^2}{\int \eta(y)\|\nabla\Lambda(y)\|^2 dy}.$$

The mixed objective J is similar to J_4, with the detection and estimation terms normalized. The optimal point density for the mixed objective is

$$\zeta^J(x) = \frac{[\rho q(x) + (1-\rho)p(x)]^{\frac{k}{k+2}}}{\int [\rho q(y) + (1-\rho)p(y)]^{\frac{k}{k+2}} dy}.$$

It is easy to see that for $\rho = 0$, ζ^J is the ROC-optimal, congruent-cell point density while for $\rho = 1$, ζ^J is the estimation-optimal point density.

Mixed Detection-Estimation Objective without Small-Cell Assumption

Finally, we note that the refinement procedure described in Section 3.3.4 can be used to improve the estimation performance of LLR quantizers. In this case, however, the LLR quantizer is not lossless. Therefore, care must be taken to first ensure adequate detection performance and then improve estimation performance by refinement.

3.7 Numerical Examples

In this section, we demonstrate the concepts and procedures described in Section 3.6 through some numerical examples. We focus on one and two-dimensional congruent-cell quantizers for a variety of sources.

3.7.1 Scalar Gaussian Sources

As a first example, consider scalar, unit-variance Gaussian sources with different means:

$$\begin{aligned} q_0 &\sim \mathcal{N}(\mu_0, 1) \\ q_1 &\sim \mathcal{N}(\mu_1, 1). \end{aligned} \tag{3.54}$$

ROC-Optimal, Discrimination-Optimal, and Estimation-Optimal Quantizers

Assume the priors P_0 and P_1 are equal. The estimation-optimal point density can then be derived based on the mixture density $q = (q_0 + q_1)/2$. The log-likelihood ratio $\Lambda(x)$ is

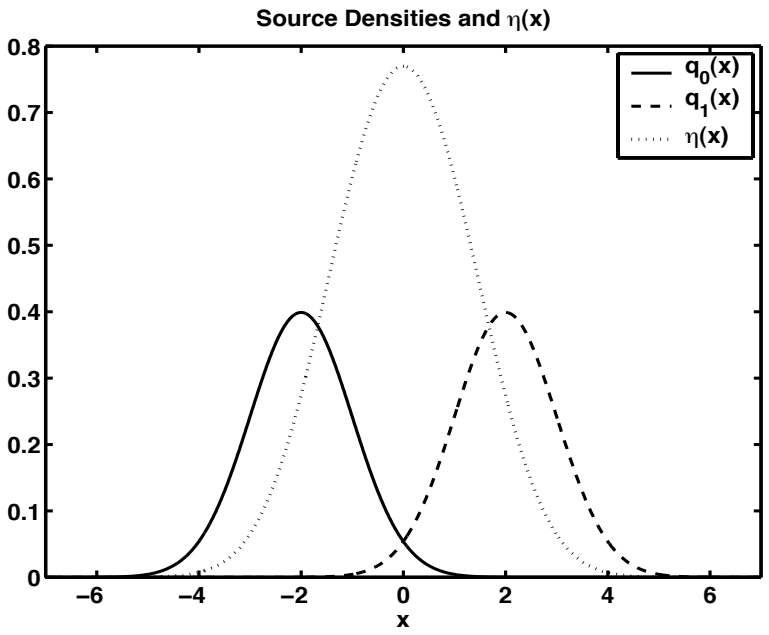

Figure 3.8: Source densities and $\eta(x)$ for one-dimensional Gaussian example.

given by

$$\Lambda(x) = -\frac{1}{2}(\mu_0^2 - \mu_1^2) + (\mu_0 - \mu_1)x.$$

Therefore, the discriminability function is constant. For scalar quantizers, the covariation profile is always constant as the cells are all intervals. Thus the discrimination-optimal and ROC-optimal point densities are given by equations (3.42) and (3.52), respectively. From these equations, we see that the discrimination-optimal quantizer should concentrate its points underneath density q_0 while the ROC-optimal quantizer concentrates its points underneath the function $\eta(x)$.

Figure 3.8 shows the sources q_0 and q_1 with $\mu_0 = -2$ and $\mu_1 = 2$ along with the function $\eta(x)$. Note that $\eta(x)$ takes a maximum at $x = 0$ where the two source densities cross. In Figure 3.9, the ROC-optimal, discrimination-optimal, and estimation-optimal point densities are plotted. As the priors are equal, the estimation-optimal point density has peaks at the maxima of the source densities. With the constant discriminability function, the ROC-optimal and discrimination-optimal point densities are maximized at points where $\eta(x)$ and $q_0(x)$ are maximized, respectively.

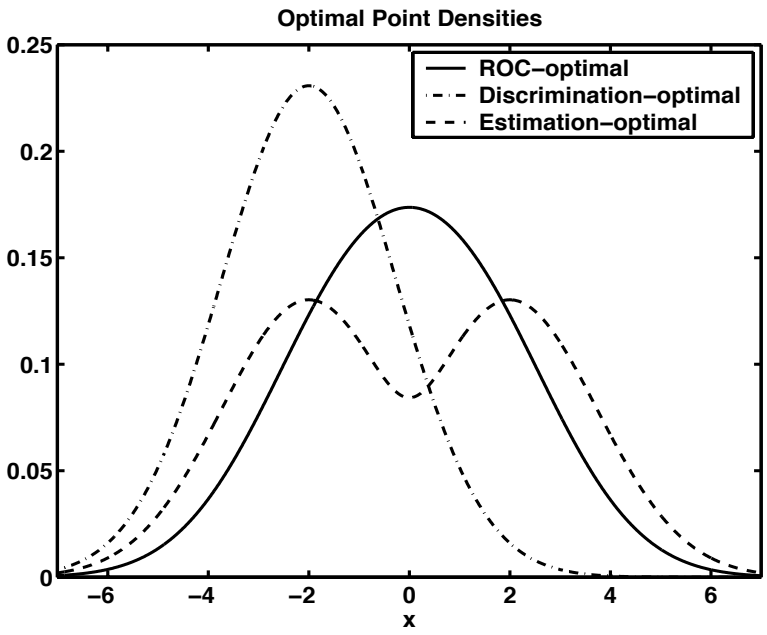

Figure 3.9: ROC-optimal, discrimination-optimal, and estimation-optimal point densities for one-dimensional Gaussian example.

In Figures 3.10, 3.11, and 3.12, the performance of scalar quantizers with the various optimal point densities is compared. The quantizers were obtained using the Lloyd algorithm (see Appendix F). Figure 3.10 shows the $L_1(L_0)$ curves with no quantization and with quantization by the ROC-optimal, discrimination-optimal, and estimation-optimal quantizers with $N = 8$ cells. As expected, the ROC-optimal quantizer performs the best as the area underneath its curve is clearly the largest. It is interesting to note that the $L_1(L_0)$ performance of the discrimination-optimal quantizer is quite poor. Recall that this quantizer is optimized only for $L(\bar{q}_0||\bar{q}_1)$. Since $L_1 = L(q_0||q_1)$ and $L_0 = 0$ for $\lambda = 0$, the value of $L(\bar{q}_0||\bar{q}_1)$ for each quantizer is the ordinate of the $L_1(L_0)$ curve at $L_0 = 0$. Observe that the discrimination-optimal curve is the largest at $L_0 = 0$. Thus, the discrimination-optimal quantizer does indeed maximize the discrimination $L(\bar{q}_0||\bar{q}_1)$, but its performance averaged over all λ is poor. Figure 3.11 shows the ROC curves of Neyman-Pearson tests with $n = 2$ i.i.d. observations with no quantization and with quantization by the various optimal quantizers with $N = 16$ cells. Note that the formulas (3.15) and (3.31) are accurate only as the

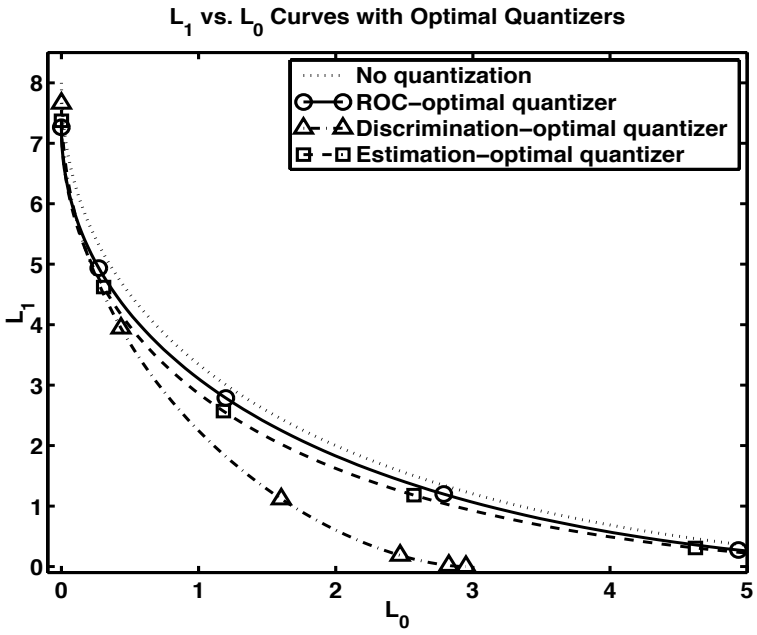

Figure 3.10: $L_1(L_0)$ curves without quantization and with quantization by ROC-optimal, discrimination-optimal, and estimation-optimal quantizers with $N = 8$ cells for one-dimensional Gaussian example. ROC-optimal quantizer has best performance, on average, while detection-optimal quantizer yields largest value of $L(\bar{q}_0 \| \bar{q}_1)$.

number of observations n becomes large and therefore the ROC-optimal quantizer may or may not actually yield an optimum ROC curve. However, for this example we see that the ROC-optimal quantizer does indeed have the best performance. Finally, in Figure 3.12 the estimation performance of the three quantizers with $N = 16$ cells is compared. The reconstruction MSE of each quantizer is plotted versus the prior probability $P_0 = P(H_0)$. The estimation-optimal quantizer is assumed to have knowledge of the priors. As expected, the estimation-optimal quantizer yields the minimum reconstruction MSE of the three considered quantizers. Note the extremely poor performance of the discrimination-optimal quantizer for $P_0 < 1$. Recall that the discrimination-optimal quantizer concentrates its points mostly underneath density q_0. For $P_0 = 1$, the discrimination-optimal and estimation-optimal quantizers are the same. For $P_0 < 1$, however, the discrimination-optimal quantizer differs quite greatly from the estimation-optimal quantizer. See for example Figure 3.9, which shows the two point densities for $P_0 = 1/2$.

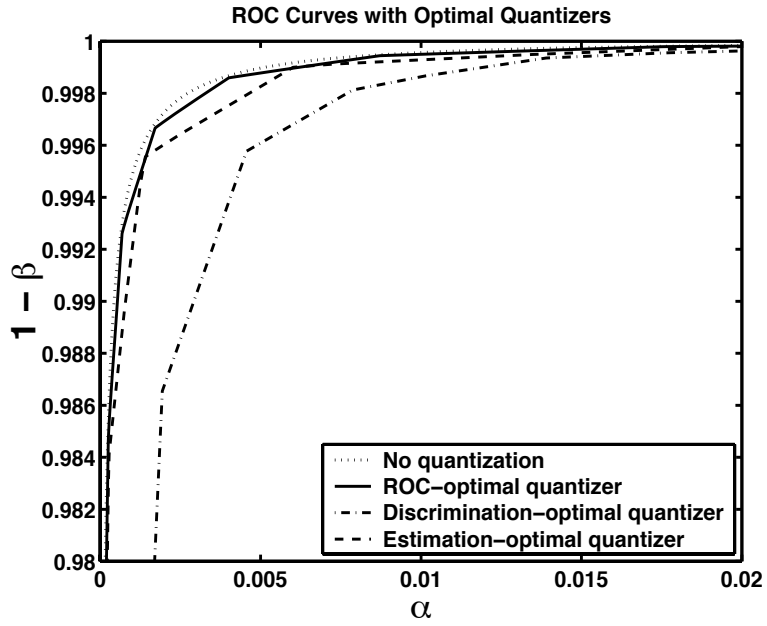

Figure 3.11: ROC curves with $n = 2$ observations and data quantized by ROC-optimal, discrimination-optimal, and estimation-optimal quantizers with $N = 16$ cells for one-dimensional Gaussian example.

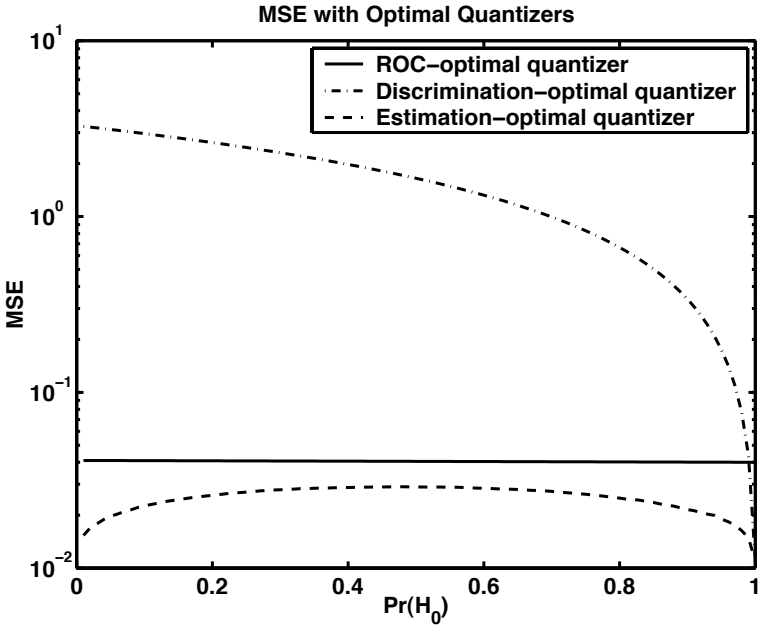

Figure 3.12: Reconstruction MSE with ROC-optimal, discrimination-optimal, and estimation-optimal quantizers with $N = 16$ cells for one-dimensional Gaussian example.

Chernoff-Information-Optimal Quantizer

Equal-variance Gaussian sources permit an additional optimization procedure: maximization of Chernoff information. Recall that the Chernoff information C, given in equation (3.16), is the greatest possible exponent in the probability of error of a Bayes test. Note from equation (3.16) that maximization of Chernoff information is equivalent to maximizing $L(\hat{q}_{\lambda^*}\|\bar{q}_0) = L(\hat{q}_{\lambda^*}\|\bar{q}_1)$ where λ^* is chosen such that the two discriminations are equal. Thus, this optimization criterion differs from the ROC optimality criterion in that the parameter λ is fixed.

Again, assume the sources given by (3.54). It can easily be shown that the tilted density is Gaussian with unit variance:

$$q_\lambda \sim \mathcal{N}(\mu_\lambda, 1)$$

where $\mu_\lambda = (1-\lambda)\mu_0 + \lambda\mu_1$. The log-likelihood ratios $\Lambda_0(x)$ and $\Lambda_1(x)$ given by (3.35) are

$$\begin{aligned} \Lambda_0(x) &= -\frac{1}{2}(\mu_\lambda - \mu_0)^2 + x(\mu_\lambda - \mu_0) \\ \Lambda_1(x) &= -\frac{1}{2}(\mu_\lambda - \mu_1)^2 + x(\mu_\lambda - \mu_1). \end{aligned}$$

Now, the discrimination loss $\Delta L_{0,N}$ given in equation (3.33) can be written

$$\Delta L_{0,N} \approx \frac{\mathcal{F}}{2N^{2/k}} \int \frac{q_\lambda(x)}{\zeta(x)^{2/k}} dx \left(\lambda^2 + \lambda(1-\lambda)\left(E_{q_\lambda}[\Lambda_0(x)] - E_p[\Lambda_0(x)]\right)\right)$$

where

$$E_{q_\lambda}[\Lambda_0(x)] = \int q_\lambda(x)\Lambda_0(x)dx, \qquad E_p[\Lambda_0(x)] = \int p(x)\Lambda_0(x)dx,$$

and

$$p(x) = \frac{q_\lambda(x)/\zeta(x)^{2/k}}{\int q_\lambda(y)/\zeta(y)^{2/k}dy}.$$

Note that the Fisher covariation profile $\mathcal{F}$ is constant. If the point density $\zeta(x)$ is symmetric about μ_λ, then the "density" $p(x)$ is also symmetric about μ_λ. Since $\Lambda_0(x)$ is linear in x, it

is easily seen that $E_{q_\lambda}[\Lambda_0(x)] = E_p[\Lambda_0(x)]$ and thus

$$\Delta L_{0,N} \approx \frac{\lambda^2 \mathcal{F}}{2N^{2/k}} \int \frac{q_\lambda(x)}{\zeta(x)^{2/k}} dx.$$

By similar arguments it can be shown that

$$\Delta L_{1,N} \approx \frac{(1-\lambda)^2 \mathcal{F}}{2N^{2/k}} \int \frac{q_\lambda(x)}{\zeta(x)^{2/k}} dx.$$

To maximize Chernoff information, the discriminations $\hat{L}_0$ and $\hat{L}_1$ must be equal. It is easy to see that using $\lambda = 1/2$ gives this result. The Chernoff-information-optimal point density for the Gaussian sources given by (3.54) is thus

$$\zeta^{\mathrm{Ch}}(x) = \frac{q_{1/2}(x)^{1/3}}{\int q_{1/2}(y)^{1/3} dy}.$$

For $\mu_0 = 0$ and $\mu_1 = 8$, Figure 3.13 shows the optimal point density for Chernoff information ζ^{Ch}, along with the ROC-optimal point density ζ^o. Both point densities are maximized at $x = 4$, where the two source densities cross. The Chernoff-information-optimal quantizer places more emphasis at this point, however. In Figure 3.14, the $L_1(L_0)$ curve is plotted along with the quantized curves for both quantizers with $N = 8$ cells. Note that the intersection of each of these curves with the unit-slope line gives the corresponding Chernoff information. The Chernoff-optimal curve lies above the ROC-optimal curve in a region close to the intersection with the unit-slope line, thus yielding greater Chernoff information. On the other hand, the area under the ROC-optimal curve is greater, as expected. Note that the Chernoff-optimal quantizer is optimized specifically for $\lambda = 1/2$, and not for any other value of λ.

Finally, we note that this analysis can be extended to obtain higher-dimensional Chernoff-information-optimal VQ's for Gaussian sources with identity covariance matrices. For these cases, we must restrict attention to quantizers with point densities and covariation profiles that are symmetric about $\underline{\mu}_\lambda$, the mean of the tilted density. For example, restricted polar quantizers [45] and some shape-gain quantizers [27] satisfy this constraint.

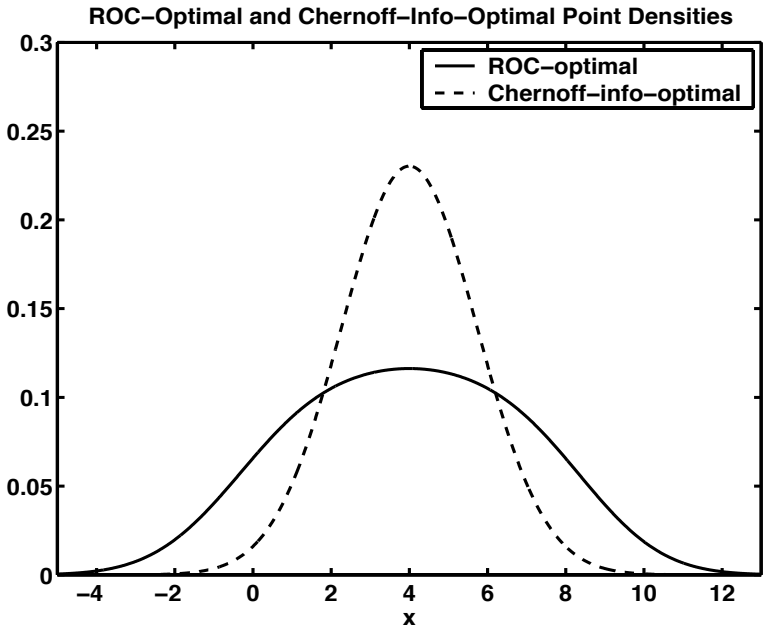

Figure 3.13: Optimal point densities for ROC area and Chernoff information for one-dimensional Gaussian sources with $\mu_0 = 0$ and $\mu_1 = 8$.

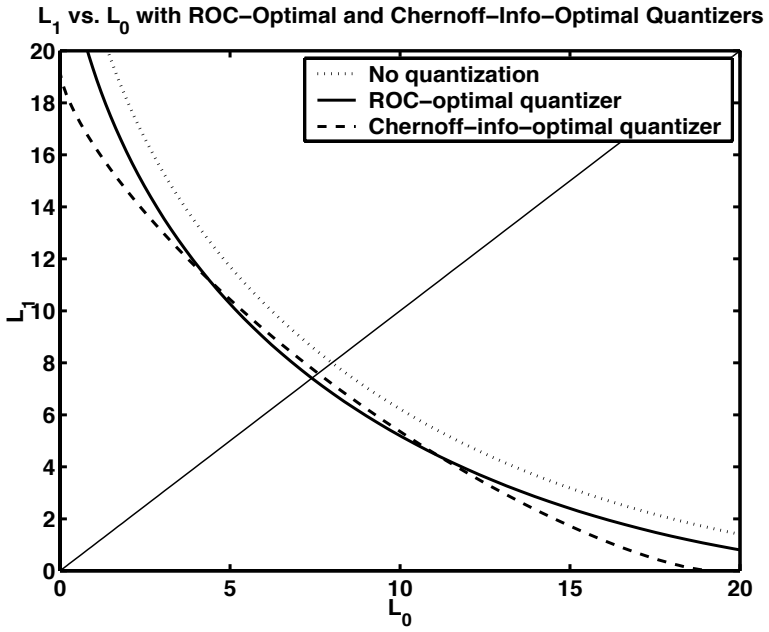

Figure 3.14: $L_1(L_0)$ curves without quantization and with quantization by ROC-optimal and Chernoff-information-optimal quantizers for one-dimensional Gaussian sources with $N = 8$, $\mu_0 = 0$, and $\mu_1 = 8$.

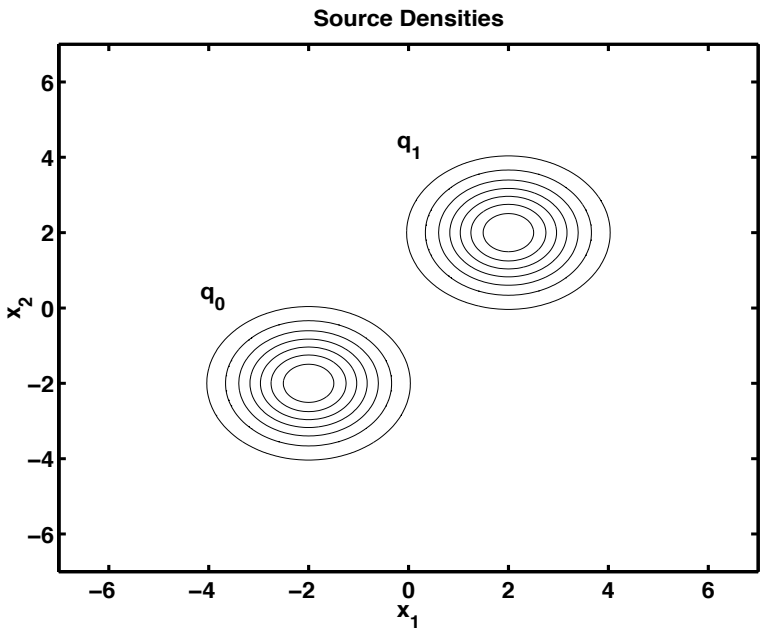

Figure 3.15: Source densities for two-dimensional uncorrelated Gaussian example.

3.7.2 Two-Dimensional Uncorrelated Gaussian Sources

Next, consider two-dimensional Gaussian sources with identity covariance matrices:

$$\begin{aligned} q_0 &\sim \mathcal{N}\left(\underline{\mu}_0, I\right) \\ q_1 &\sim \mathcal{N}\left(\underline{\mu}_1, I\right) \end{aligned}$$

where $\underline{\mu}_0 = [\mu_0, \mu_0]$ and $\underline{\mu}_1 = [\mu_1, \mu_1]$ are the mean vectors.

ROC-Optimal, Discrimination-Optimal, and Estimation-Optimal Quantizers

As in the scalar Gaussian example, the discriminability function is constant for two-dimensional Gaussian sources with identity covariance matrices. For congruent-cell quantizers, the discrimination-optimal and ROC-optimal point densities are again given by equations (3.42) and (3.52), respectively.

Figure 3.15 shows contours of the two source densities for $\mu_0 = -2$ and $\mu_1 = 2$. In Figure 3.16, several functions are shown. Observe that the function $\eta(x)$ takes a maximum in a region between the peaks of the source densities. The log-likelihood ratio and the discriminability function are shown in Figures 3.16b and 3.16c. The constant discriminability

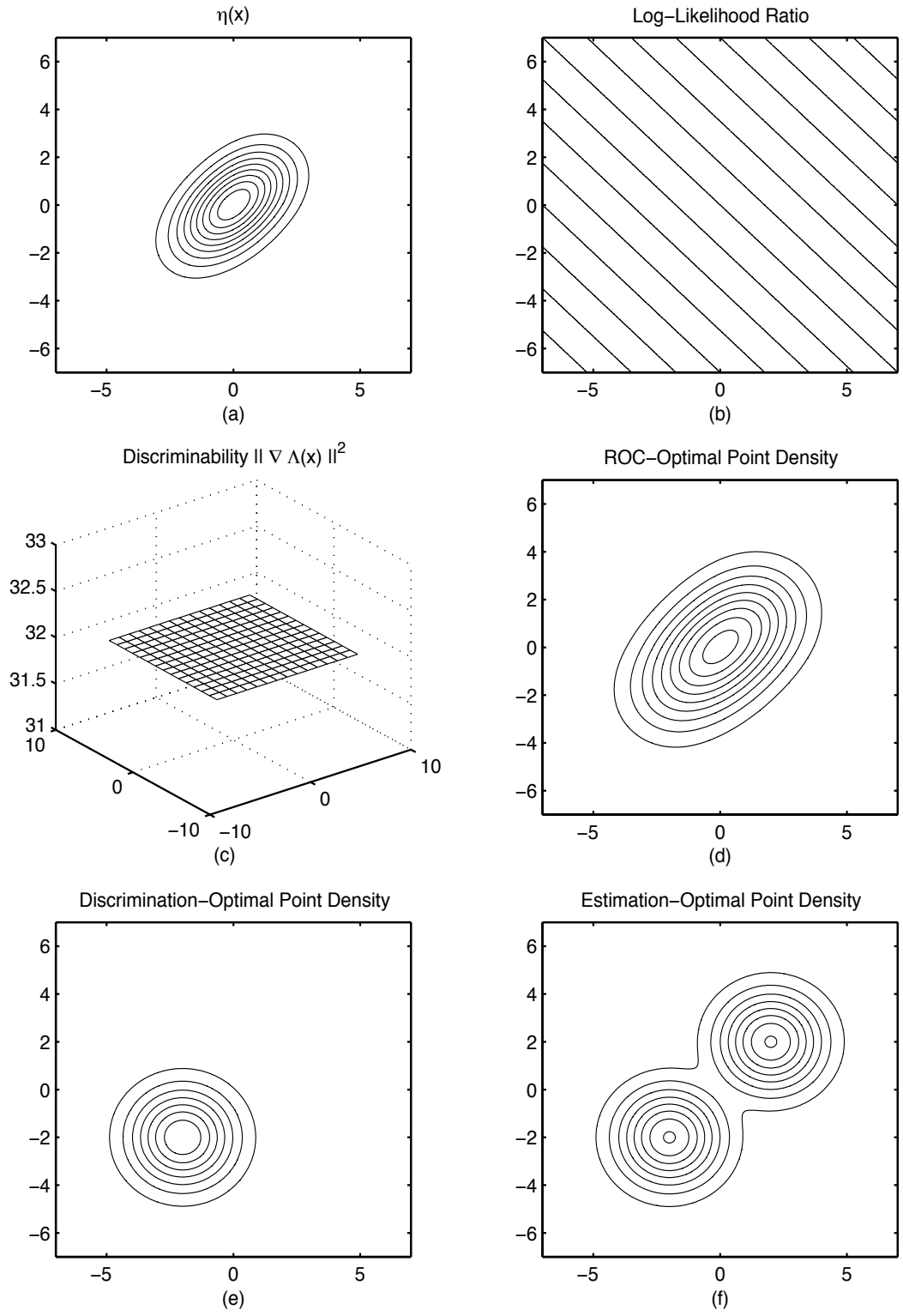

Figure 3.16: Two-dimensional uncorrelated Gaussian example: (a) $\eta(x)$, (b) log-likelihood ratio $\Lambda(x)$, (c) discriminability $\|\nabla\Lambda(x)\|^2$, (d) ROC-optimal point density, (e) discrimination-optimal point density, (f) estimation-optimal point density.

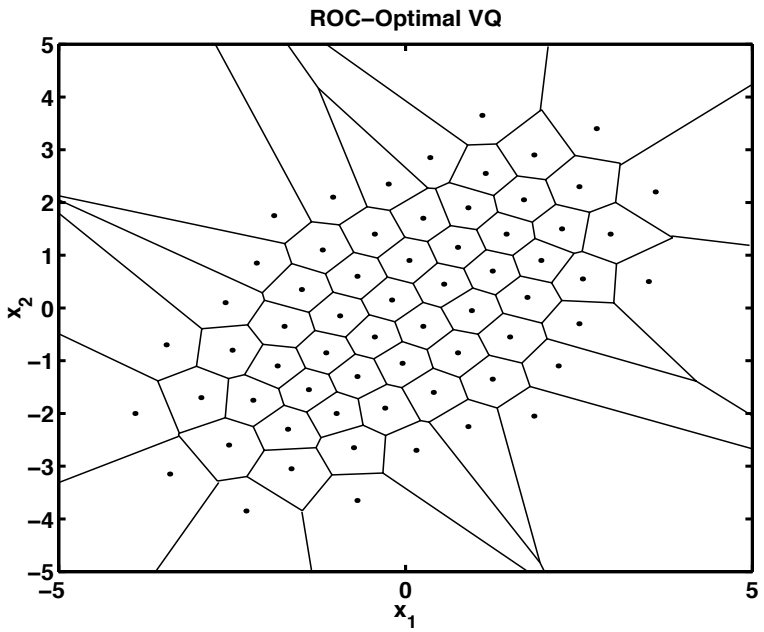

Figure 3.17: ROC-optimal 64-cell vector quantizer for two-dimensional uncorrelated Gaussian example.

function results in the ROC-optimal point density having contours aligned with those of $\eta(x)$ and the discrimination-optimal point density having contours aligned with those of q_0. Finally, the estimation-optimal point density for equal priors is shown in Figure 3.16f.

The two-dimensional congruent-cell quantizers with $N = 64$ cells with the ROC-optimal, discrimination-optimal, and estimation-optimal point densities are shown in Figures 3.17, 3.18, and 3.19, respectively. These quantizers were again obtained using the generalized Lloyd, or LBG algorithm. This algorithm and its utility for obtaining optimal congruent-cell quantizers is described in Appendix F.

The hypothesis testing performance of the 64-cell quantizers in Figures 3.17, 3.18, and 3.19 is compared in Figure 3.20. Similar to the scalar Gaussian example, the ROC-optimal quantizer performs the best, while the discrimination-optimal quantizer yields the largest discrimination between quantized sources, but performs poorly on average.

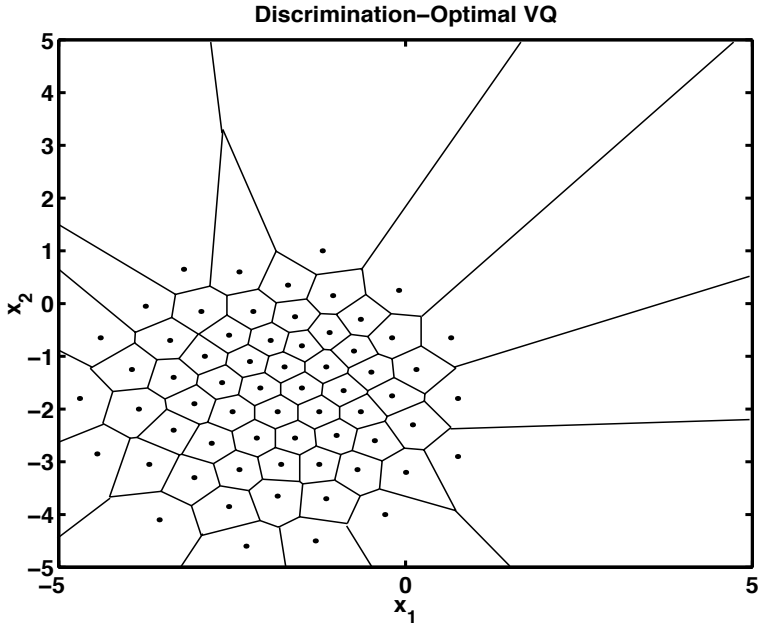

Figure 3.18: Discrimination-optimal 64-cell vector quantizer for two-dimensional uncorrelated Gaussian example.

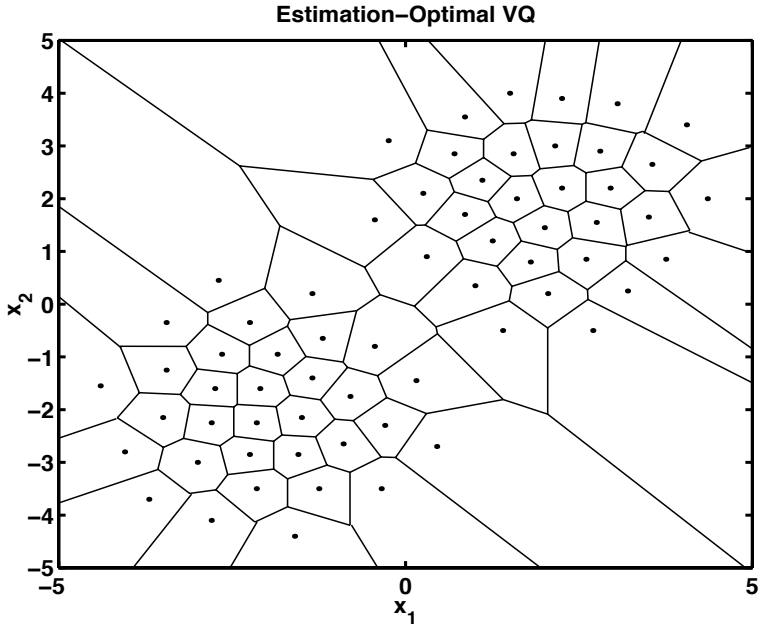

Figure 3.19: Estimation-optimal 64-cell vector quantizer for two-dimensional uncorrelated Gaussian example.

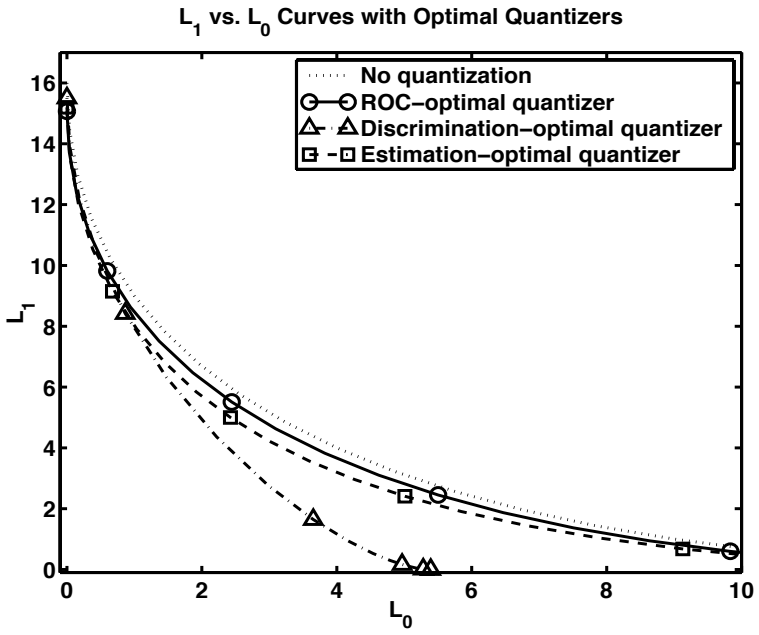

Figure 3.20: $L_1(L_0)$ curves without quantization and with quantization by ROC-optimal, discrimination-optimal, and estimation-optimal quantizers with $N = 64$ cells for two-dimensional uncorrelated Gaussian example. ROC-optimal quantizer has best performance, on average, while detection-optimal quantizer yields largest value of $L(\bar{q}_0 \| \bar{q}_1)$.

Optimal Log-Likelihood Ratio Quantizer

Next, the optimal log-likelihood ratio quantizer for this example is obtained by first noting that the log-likelihood ratio $\Lambda(x)$ is

$$\Lambda(x) = \frac{1}{2}\left(\|\underline{\mu}_1\|^2 - \|\underline{\mu}_0\|^2\right) + (\mu_0 - \mu_1)(x_1 + x_2)$$

where x_1 and x_2 are the two components of the source vector x. Thus, the densities $q_{\Lambda,0}(l)$ and $q_{\Lambda,1}(l)$ of the log-likelihood ratio under the two hypotheses are Gaussian and the ROC-optimal constituent quantizer Q_Λ, which quantizes $\Lambda(x)$, is easily obtained. The cells and codebook points of the resultant ROC-optimal LLR quantizer with $N = 64$ are shown in Figure 3.21. Note that the codebook points are arbitrarily chosen to lie in the centroids of their cells. Each cell boundary is a level set (contour) of $\Lambda(x)$ (see Figure 3.16b).

The hypothesis testing performance of the 64-cell ROC-optimal LLR quantizer is compared to that of the 64-cell ROC-optimal congruent-cell quantizer in Figure 3.22. Observe that the LLR quantizer yields better hypothesis testing performance. The reconstruction

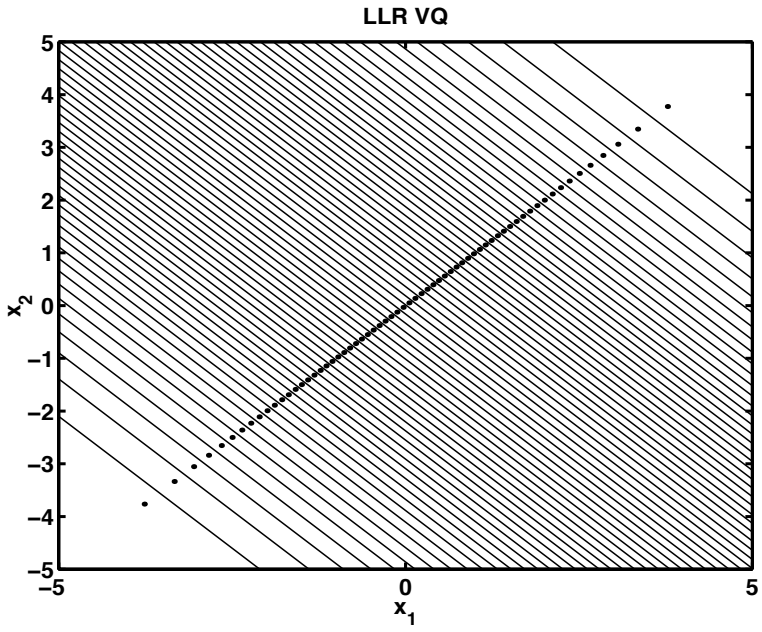

Figure 3.21: Optimal 64-cell log-likelihood ratio quantizer for two-dimensional uncorrelated Gaussian example.

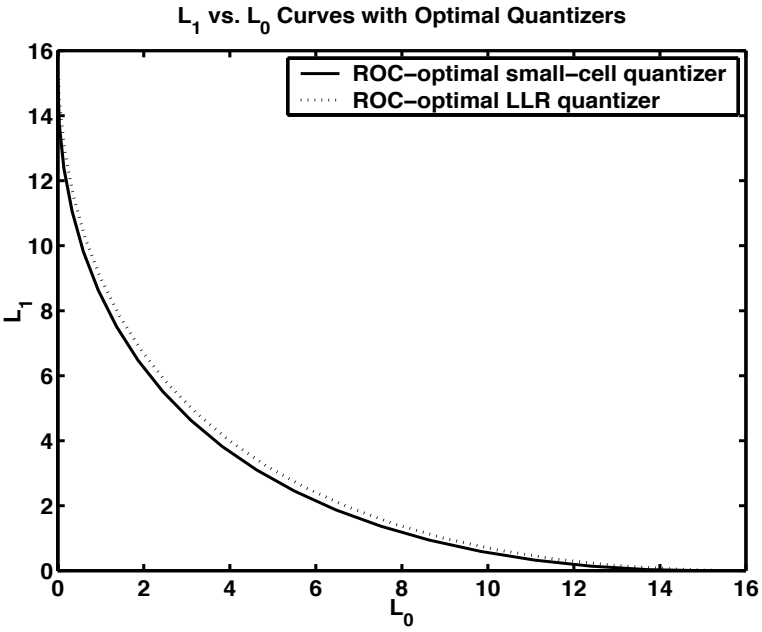

Figure 3.22: $L_1(L_0)$ curves with 64-cell ROC-optimal congruent-cell quantizer and 64-cell ROC-optimal LLR quantizer for two-dimensional uncorrelated Gaussian example.

Type of quantizer	MSE
Estimation-opt	0.1181
ROC-opt, congruent-cell	0.1709
ROC-opt LLR quantizer	1.0051

Table 3.1: Reconstruction MSE of 64-cell estimation-optimal quantizer, ROC-optimal congruent cell-quantizer, and ROC-optimal LLR quantizer for two-dimensional uncorrelated Gaussian example.

MSE of the two quantizers along with that of the 64-cell estimation-optimal quantizer assuming equal priors is shown in Table 3.1. The congruent-cell quantizer has a much lower reconstruction MSE than that of the LLR quantizer. The long "strip" cells of the LLR quantizer renders its estimation performance quite poor. Note that if we let N become large, the reconstruction MSE of the optimal LLR quantizer will not vanish as the diameter function does not converge to zero.

Optimal Mixed-Objective Quantizer

An optimal 64-cell mixed-objective quantizer for this example is shown in Figure 3.23 for $\rho = 1/2$. Recall that the mixed objective function J given by (3.53) incorporates both area loss ΔA and reconstruction MSE. Thus this quantizer concentrates its points between the source density peaks as does the ROC-optimal quantizer in Figure 3.17, as well as underneath the peaks as does the estimation-optimal quantizer in Figure 3.19.

The detection and estimation performance of the optimal mixed-objective quantizer as a function of the parameter ρ is shown in Figure 3.24. As expected, the detection performance degrades as ρ is varied from 0 to 1, while the estimation performance improves.

3.7.3 Two-Dimensional Correlated Gaussian Sources

The next example assumes the following sources:

$$q_0 \sim \mathcal{N}\left(\underline{\mu}_0, K_0\right)$$
$$q_1 \sim \mathcal{N}\left(\underline{\mu}_1, K_1\right)$$

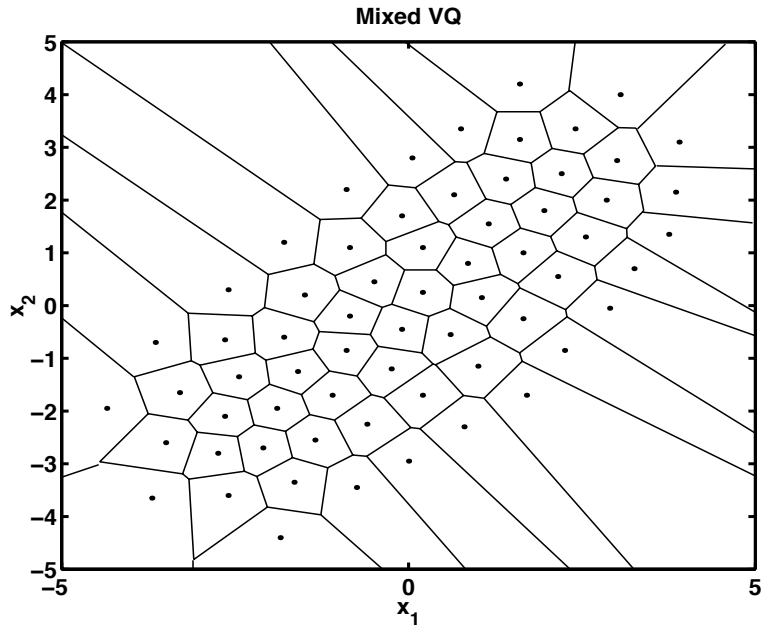

Figure 3.23: Optimal 64-cell vector quantizer with mixed objective function with $\rho = 1/2$ for two-dimensional uncorrelated Gaussian example.

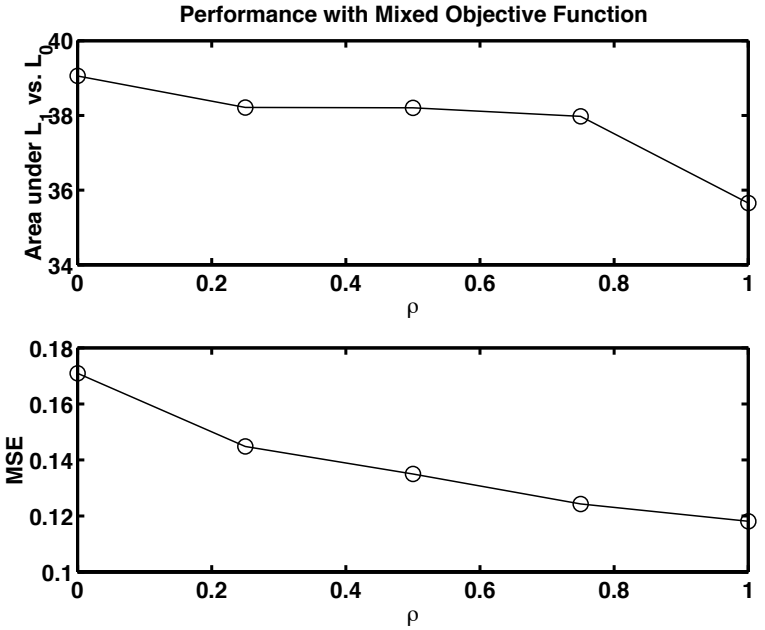

Figure 3.24: Detection and estimation performance of optimal 64-cell vector quantizer with mixed objective function for two-dimensional uncorrelated Gaussian example.

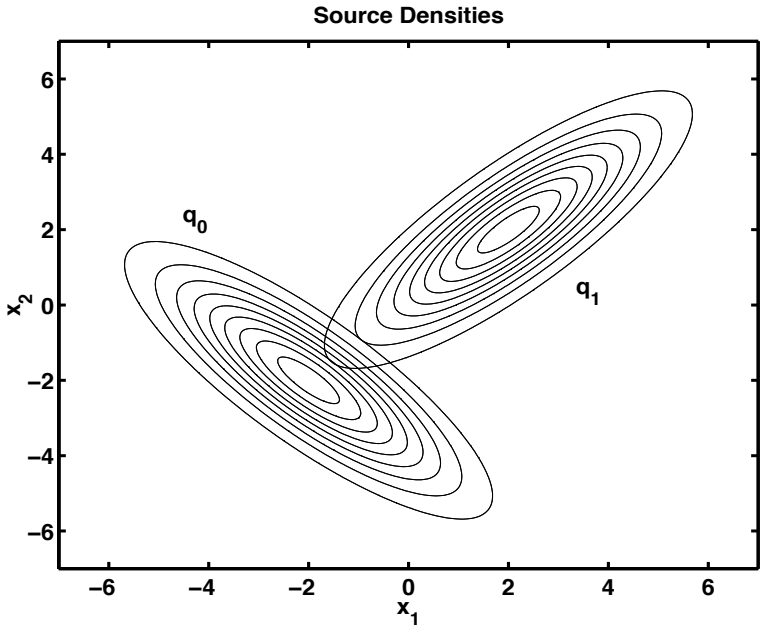

Figure 3.25: Source densities for two-dimensional correlated Gaussian example.

where $\underline{\mu}_0 = [-2, -2]$, $\underline{\mu}_1 = [2, 2]$, and

$$K_0 = \begin{bmatrix} 3 & -2.5 \\ -2.5 & 3 \end{bmatrix}, \quad K_1 = \begin{bmatrix} 3 & 2.5 \\ 2.5 & 3 \end{bmatrix}.$$

The source densities are shown in Figure 3.25.

Figure 3.26 shows several functions associated with this example. Unlike the uncorrelated Gaussian example, the discriminability function is not constant. Recall that equations (3.42) and (3.52) give the discrimination-optimal and ROC-optimal point densities assuming congruent cells. The contours of the discrimination-optimal and ROC-optimal point densities are therefore no longer aligned with those of $q_0(x)$ and $\eta(x)$ as was the case when the discriminability function was constant.

The 64-point optimal congruent-cell quantizers are shown in Figures 3.27, 3.28, and 3.29. The hypothesis testing performance of these quantizers is compared in Figure 3.30, which shows their $L_1(L_0)$ curves. As in the two previous examples, the ROC-optimal quantizer has the best performance (maximum area) of all three quantizers and the discrimination-optimal quantizer yields the largest value of $L(\bar{q}_0 \| \bar{q}_1)$.

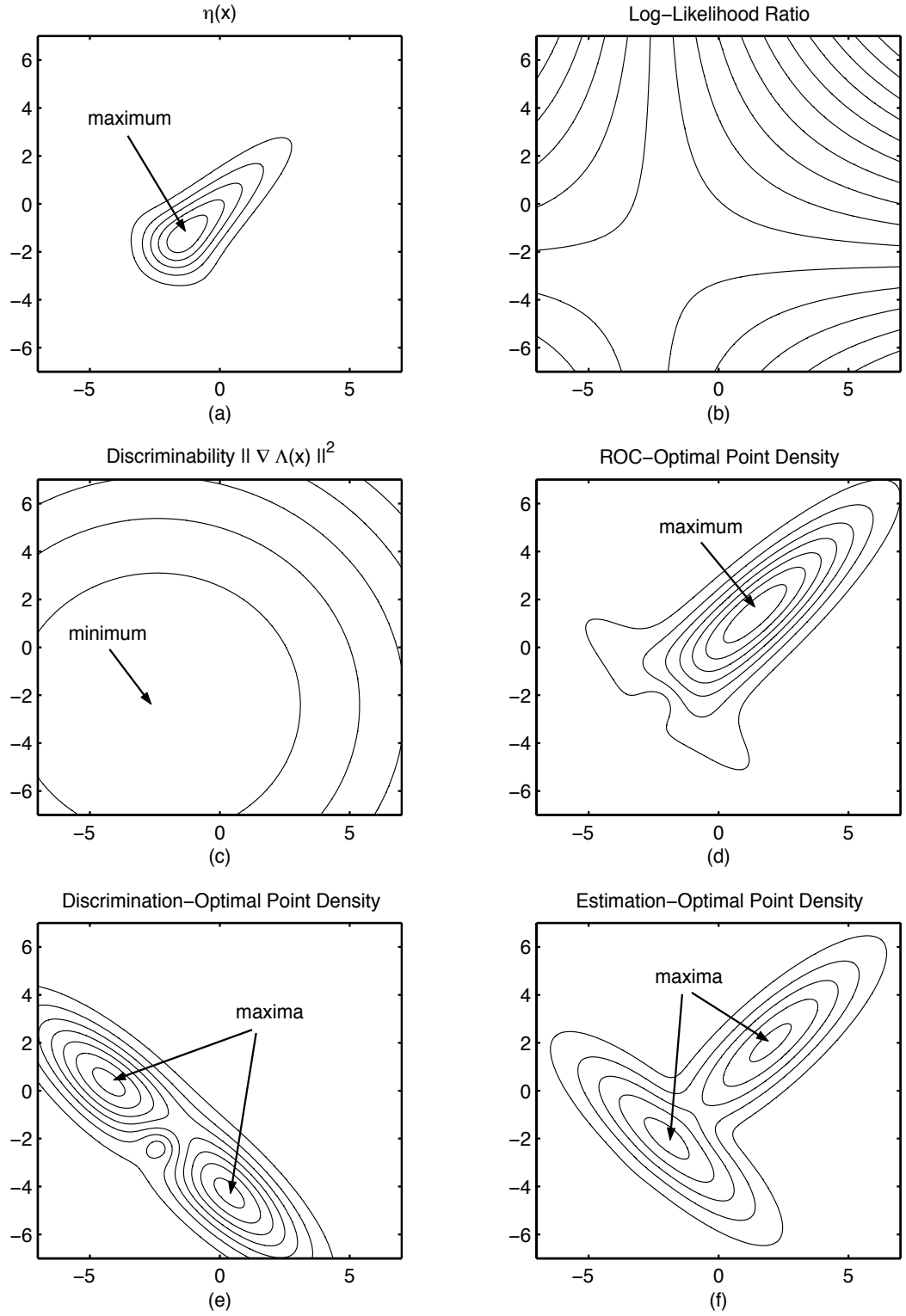

Figure 3.26: Two-dimensional correlated Gaussian example: (a) $\eta(x)$, (b) log-likelihood ratio $\Lambda(x)$, (c) discriminability $\|\nabla\Lambda(x)\|^2$, (d) ROC-optimal point density, (e) discrimination-optimal point density, (f) estimation-optimal point density.

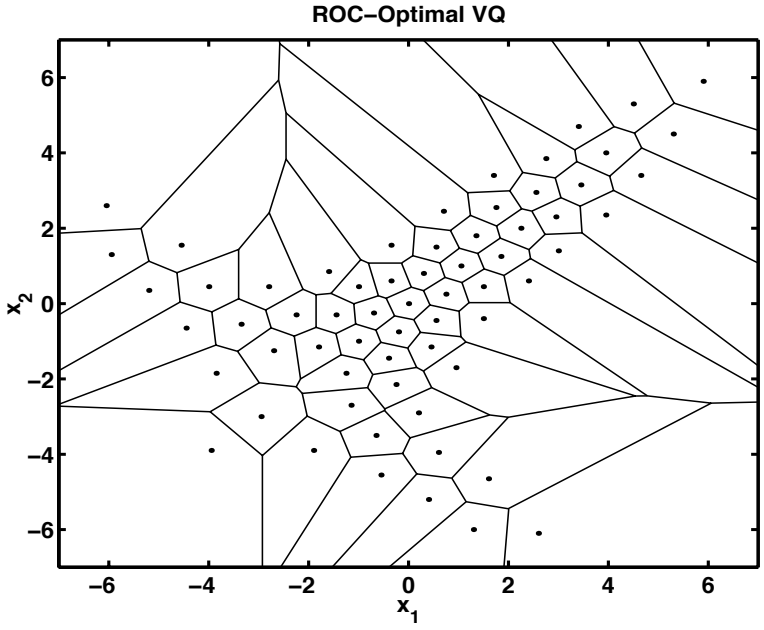

Figure 3.27: ROC-optimal 64-cell vector quantizer for two-dimensional correlated Gaussian example.

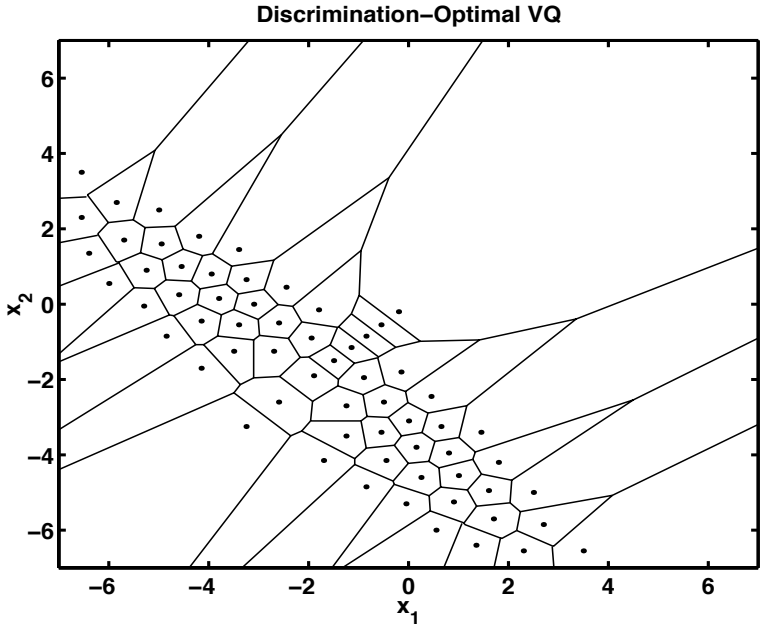

Figure 3.28: Discrimination-optimal 64-cell vector quantizer for two-dimensional correlated Gaussian example.

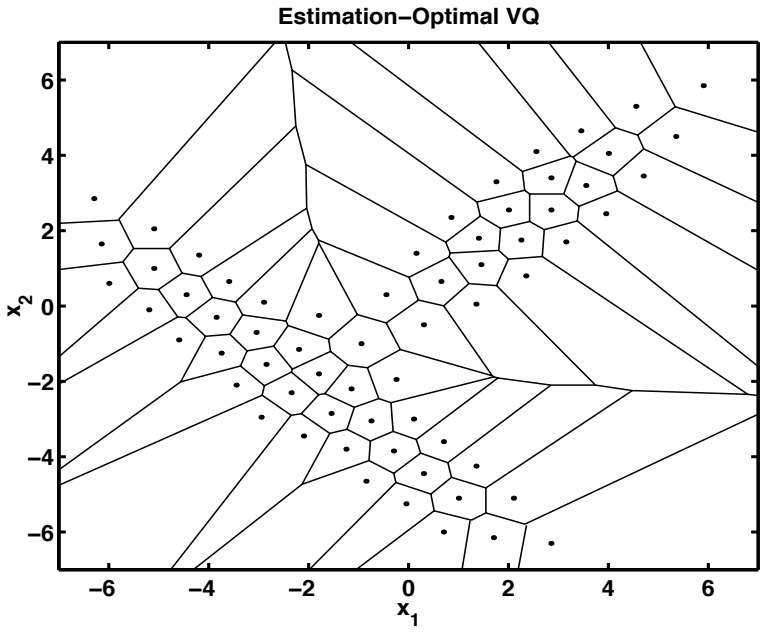

Figure 3.29: Estimation-optimal 64-cell vector quantizer for two-dimensional correlated Gaussian example.

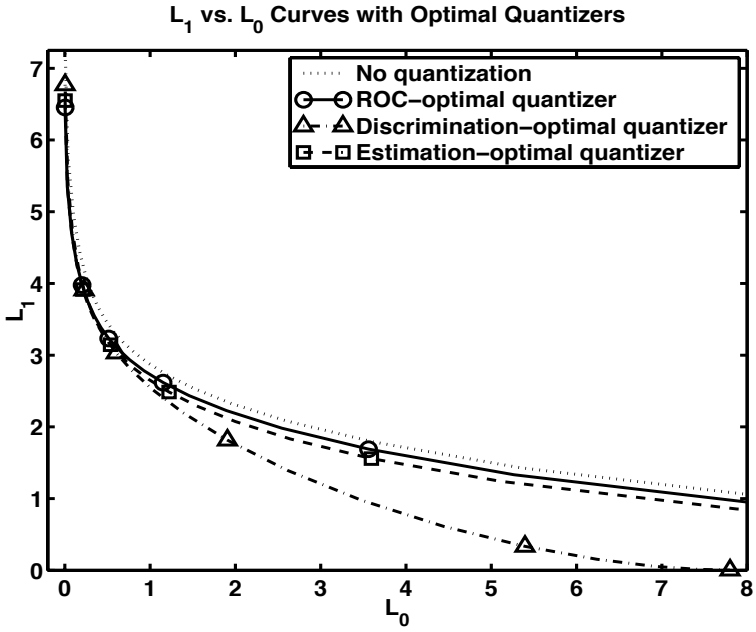

Figure 3.30: $L_1(L_0)$ curves without quantization and with quantization by ROC-optimal, discrimination-optimal, and estimation-optimal quantizers with $N = 64$ cells for two-dimensional correlated Gaussian example. ROC-optimal quantizer has best performance, on average, while detection-optimal quantizer yields largest value of $L(\bar{q}_0||\bar{q}_1)$.

3.7.4 Triangular Sources

We now consider hypothesis testing with scalar, finite-support, linear densities. Let the densities be given by

$$\begin{aligned} q_0(x) &= 2x+1, \ |x| \leq 1/2 \\ q_1(x) &= -2x+1, \ |x| \leq 1/2 \end{aligned}$$

with $q_0(x) = q_1(x) = 0$ for $|x| > 1/2$.

Figure 3.31 shows the source densities as well as $\eta(x)$, the log-likelihood ratio gradient, and the ROC-optimal point density. Similar to the one-dimensional Gaussian example, we see that $\eta(x)$ takes a maximum at the point $x = 0$ where the two densities cross. The gradient of the log-likelihood ratio increases in magnitude as $|x|$ increases. Thus, the discriminability is largest near the endpoints of the interval $B = [-1/2, 1/2]$. The ROC-optimal point density also increases as $|x|$ increases. This implies that more points should be placed near the endpoints of B, rather than near the center. Recall from equation (3.52) that for $k = 1$ the ROC-optimal point density is

$$\zeta^o(x) = \frac{[\|\nabla\Lambda(x)\|^2\eta(x)]^{1/3}}{\int[\|\nabla\Lambda(y)\|^2\eta(y)]^{1/3}dy}. \tag{3.55}$$

Thus, the shape of the optimal point density is determined by the product of the discriminability function and $\eta(x)$. For this example, the discriminability function is convex $\cup$ while $\eta(x)$ is convex $\cap$. The point density is convex $\cup$ since the curvature of the discriminability function is greater than that of $\eta(x)$.

It may seem counterintuitive that the ROC-optimal quantizer should cluster more points near the endpoints of B as opposed to the center where the densities cross, as in the Gaussian case (Section 3.7.1). Recall, however, that in the Gaussian case, the discriminability function is constant and thus does not affect the optimal point density. In this example, the discriminability has a significant impact on the optimal point density. The importance of the discriminability can be seen by comparing the $L_1(L_0)$ curves for various quantizers.

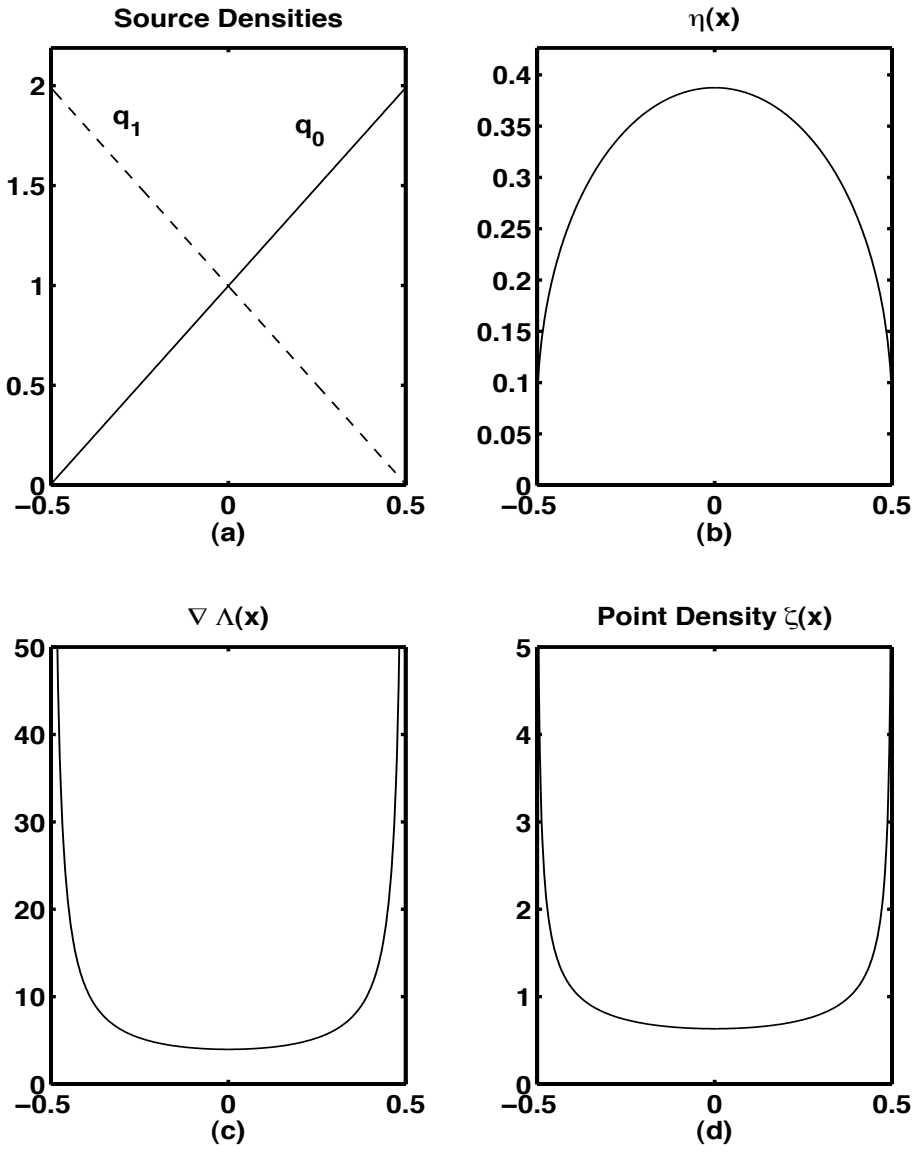

Figure 3.31: Triangular source example: (a) source densities, (b) $\eta(x)$, (c) gradient (derivative) of log-likelihood ratio, (d) ROC-optimal point density.

N	A_L	A_L^o	A_L^η
4	0.1543	0.1120	0.1066
8	0.1543	0.1408	0.1344
16	0.1543	0.1503	0.1469
32	0.1543	0.1527	0.1508

Table 3.2: Areas under $L_1(L_0)$ curves without quantization and with quantization for one-dimensional triangular example. A_L = area with no quantization, A_L^o = area after quantization with ROC-optimal quantizer, A_L^η = area after quantization with quantizer using point density ζ^η.

Define the suboptimal point density ζ^η as

$$\zeta^\eta(x) = \frac{\eta(x)^{1/3}}{\int \eta(y)^{1/3} dy}.$$

This can be viewed as the optimal point density when the discriminability function is neglected. Figure 3.32 shows the $L_1(L_0)$ curve along with the corresponding curves after quantization with eight-cell scalar quantizers using point densities ζ^o and ζ^η. The quantizers were obtained using the Lloyd algorithm (see Appendix F). The quantizer with point density ζ^o outperforms that with point density ζ^η as its curve is uniformly greater. Similar results were obtained using quantizers with $3 \leq N \leq 32$. Let the areas under the $L_1(L_0)$ curves with no quantization, quantization with point density ζ^o, and quantization with point density ζ^η, be defined as A_L, A_L^o, and A_L^η, respectively. Table 3.2 shows these areas for $N = 4,\ 8,\ 16,$ and 32. For all cases, the ROC-optimal quantizer outperforms the quantizer with point density ζ^η.

Further evidence of the optimality of ζ^o for this example can be seen by comparing the areas under the ROC curves obtained with each of the quantizers (using ζ^o and ζ^η) for reasonably large values of n. From equation (3.31) and Figure 3.32, the ROC curve area should be larger for a quantizer using point density ζ^o than for a quantizer using ζ^η for large n. Let A_{ROC}^o be the area under the ROC curve when the data is quantized with point density ζ^o. Similarly, let A_{ROC}^η be the area when point density ζ^η is used. Note that we are assuming a randomized Neyman-Pearson test [35] to get a continuous ROC curve. Next,

ΔA_{ROC}	$n = 3$	$n = 4$	$n = 5$
$N = 4$	-6.79×10^{-4}	1.53×10^{-5}	2.46×10^{-4}
$N = 5$	-1.61×10^{-4}	3.62×10^{-4}	4.65×10^{-4}
$N = 6$	3.88×10^{-4}	6.45×10^{-4}	6.07×10^{-4}
$N = 7$	5.60×10^{-4}	6.76×10^{-4}	5.87×10^{-4}
$N = 8$	4.68×10^{-4}	5.71×10^{-4}	4.97×10^{-4}

Table 3.3: Difference in ROC curve areas for one-dimensional triangular example. $\Delta A_{\text{ROC}} = A^o_{\text{ROC}} - A^\eta_{\text{ROC}}$ where A^o_{ROC} = area under ROC curve after quantization with ROC-optimal quantizer and A^η_{ROC} = area under ROC curve after quantization with quantizer using point density ζ^η.

define $\Delta A_{\text{ROC}} = A^o_{\text{ROC}} - A^\eta_{\text{ROC}}$. Table 3.3 shows the difference in areas ΔA_{ROC} for various values of N and n. As n becomes large, the difference in area is positive. Thus, although seemingly counterintuitive, clustering points near the endpoints of the interval B is optimal for detection when the number of observations is reasonably large.

It is important to note that for $n = 1$ observation, the asymptotic formulas (3.15) do not apply and the optimal point density ζ^o will not necessarily produce an ROC curve with maximum area. In fact, if the likelihood ratio is monotonic in x, as it is in this example, the quantized ROC curve will intersect the unquantized ROC curve in up to $N+1$ points. The borders of the one-dimensional quantizer cells correspond to the Neyman-Pearson thresholds at which the intersections occur. Thus, for $n = 1$, the optimal quantizer should concentrate points in areas corresponding to regions of large curvature [66] on the ROC curve. For this example, this region is near the origin. Thus, the optimal quantizer for $n = 1$ differs greatly from the optimal quantizer for $n \gg 1$.

In Figure 3.33, the log-likelihood ratio $\Lambda(x)$ is plotted. Note that as $|x| \to 1/2$, $\Lambda(x)$ becomes more and more steep. Thus, a quantizer that concentrates its points near $x = \pm 1/2$ will preserve more values of the sufficient statistic $\Lambda(x)$.

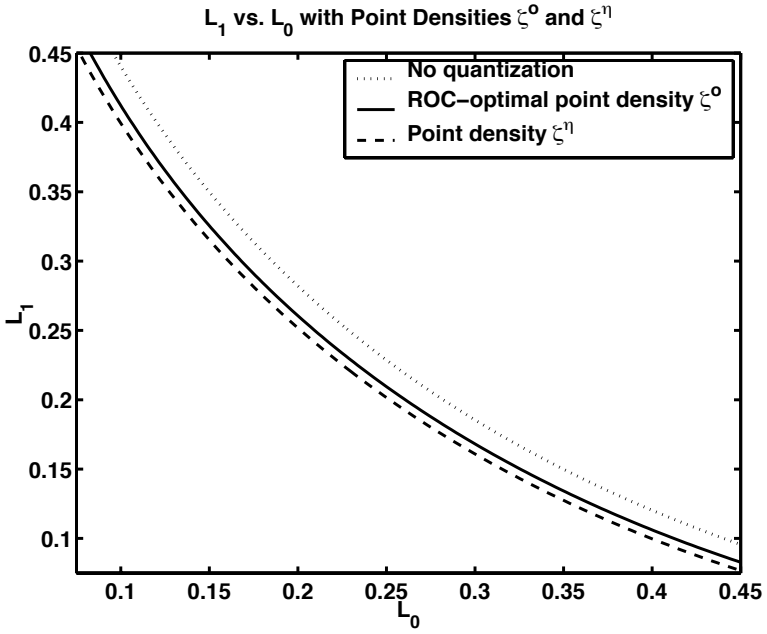

Figure 3.32: $L_1(L_0)$ curves for one-dimensional triangular example without quantization and with quantization using point densities ζ^o and ζ^η, with $N = 8$ cells.

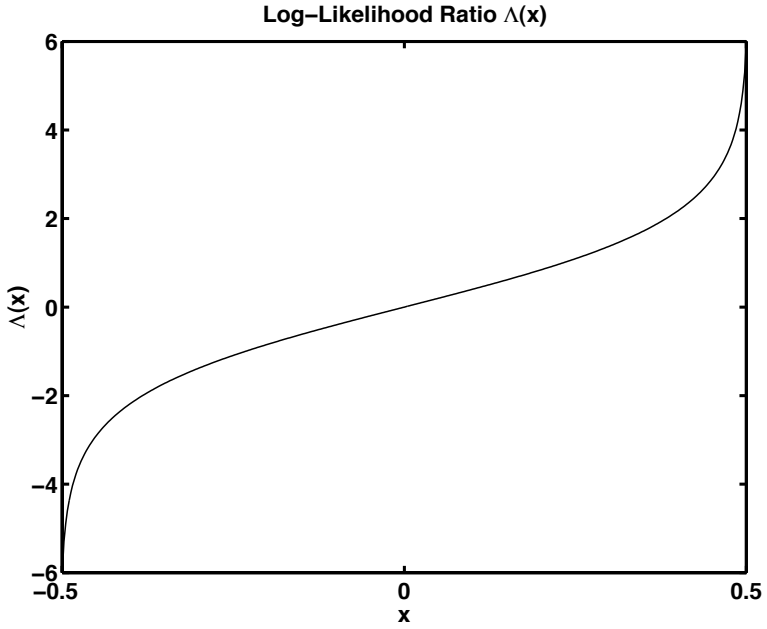

Figure 3.33: Log-likelihood ratio for one-dimensional triangular example.

3.7.5 Piecewise-Constant Sources

Although Section 3.3.3 showed that sufficient quantizers always exist when the sources are piecewise-constant, derivation of the ROC-optimal point density for piecewise-constant sources illustrates some important concepts. As a simple example, let $q_0(x)$ be uniform on the interval $[-1, 1]$ and let $q_1(x)$ be

$$q_1(x) = \begin{cases} 1, & |x| \leq 1/3 \\ 1/4, & 1/3 < |x| \leq 1 \\ 0, & |x| > 1 \end{cases} .$$

It is clear that since q_0 and q_1 are constant on the intervals $[-1, -1/3), [-1/3, 1/3]$, and $(1/3, 1]$, the likelihood ratio will also be constant on these intervals. It follows that the tilted density along with the functions f_0, f_1, and η are also piecewise-constant on the intervals. It can also be shown that $\eta(x) > 0$ for $|x| < 1$. Next, the derivative of the log-likelihood ratio is given by

$$\nabla\Lambda(x) = -(\log 4)\delta(x + 1/3) + (\log 4)\delta(x - 1/3)$$

where $\delta(\cdot)$ is the Dirac delta function. Thus, it immediately follows that the optimal point density is zero everywhere, but at the points $-1/3$ and $1/3$. Figure 3.34 shows the source densities, $\eta(x)$, $\nabla\Lambda(x)$, and the ROC-optimal point density.

A sufficient statistic for this example is

$$S(x) = I_{[-1/3,1/3]}(x).$$

Thus it is sufficient to know which of the intervals $[-1, -1/3), [-1/3, 1/3]$, or $(1/3, 1]$ contains x. Therefore, any scalar quantizer that contains cells with borders (thresholds) at $-1/3$ and $1/3$ is a sufficient quantizer. The asymptotic theory asserts that a many-point quantizer must concentrate its points close to $-1/3$ and $1/3$. This guarantees the existence of thresholds at $-1/3$ and $1/3$ and thus, sufficiency.

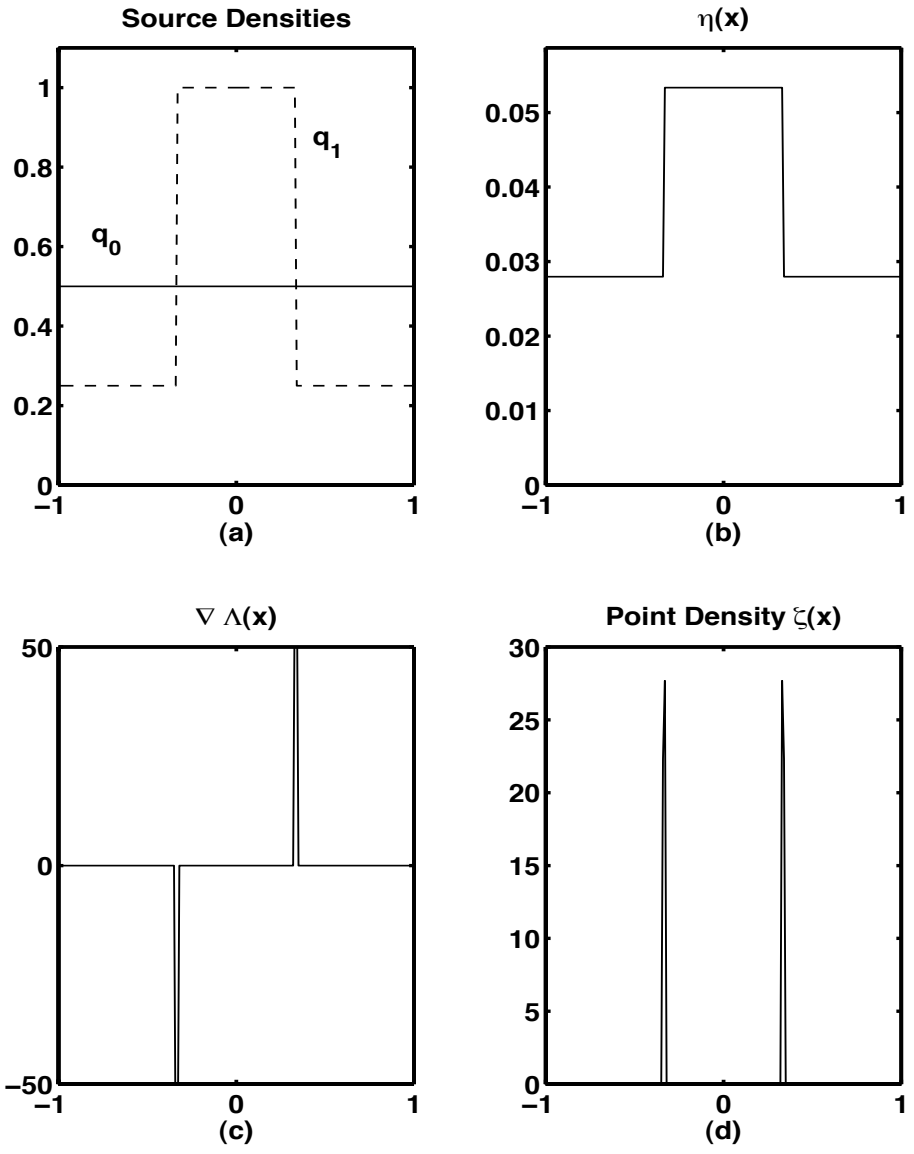

Figure 3.34: Piecewise-constant source example: (a) source densities, (b) $\eta(x)$, (c) gradient (derivative) of log-likelihood ratio, (d) ROC-optimal point density.

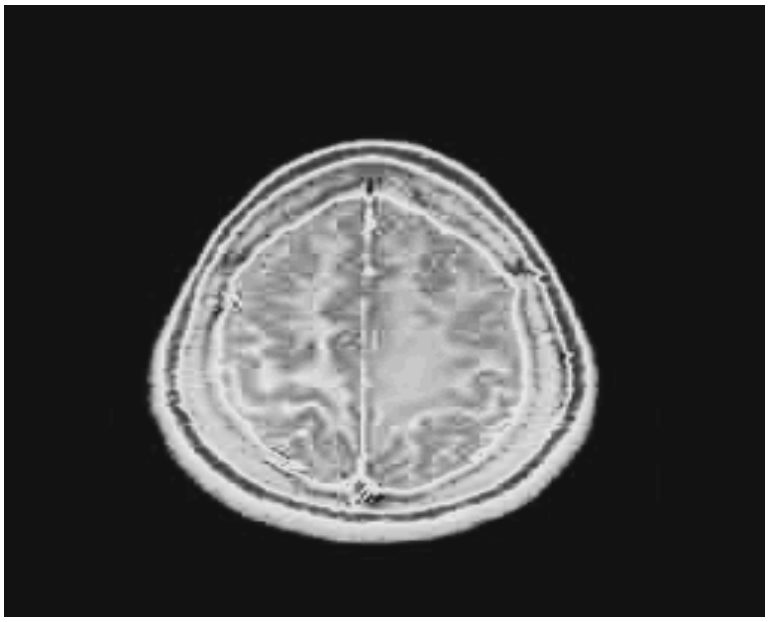

Figure 3.35: Null density for two-dimensional image example.

3.7.6 Two-Dimensional Image Sources

Lastly, we consider sources whose probability densities are given by two-dimensional images. This is the case when the data comes from a photon counter that observes an environment. Most photons will come from the brightest areas in the environment. Therefore, the two-dimensional image that is formed from the received photons resembles a probability density of the location of the photons.

Figures 3.35 and 3.36 are two images that can be used to represent the null and alternate source densities. These images are cross-sections of a human brain. The alternate density contains a tumor in the lower-right quadrant while the null density is tumor free. In the areas where the tumor is not located, the images are nearly identical. The "background" regions of the images are both zero and the likelihood ratio is defined to be one here. On the "brain-minus-tumor" region, the likelihood ratio is nearly constant and thus has very low discriminability. The discriminability is largest at the edges of the brain and at the tumor location. Figure 3.37 shows the ROC-optimal point density for these two source densities. Most points should be concentrated near the edge of the brain and at the tumor location.

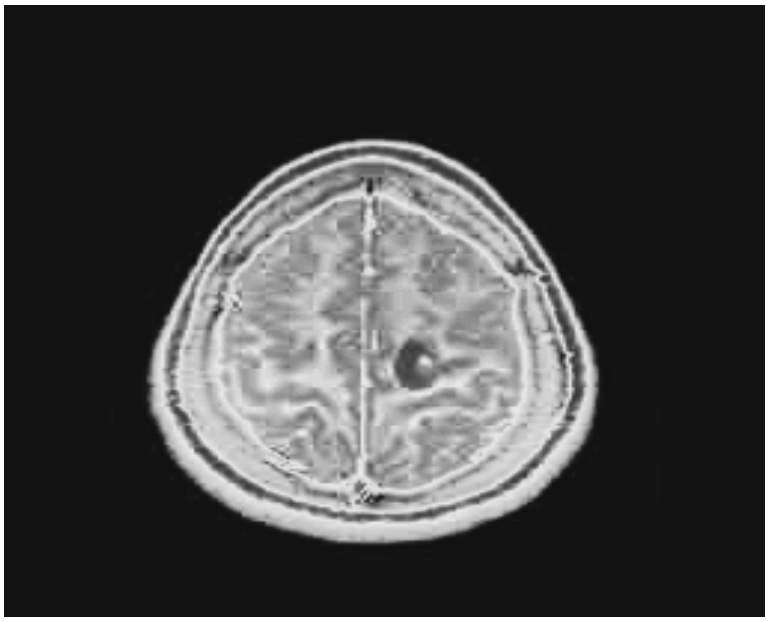

Figure 3.36: Alternate density for two-dimensional image example.

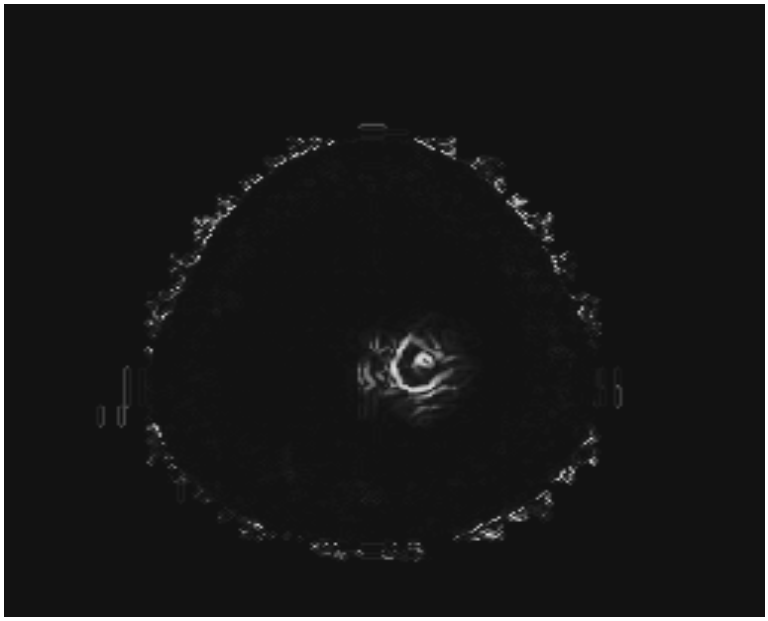

Figure 3.37: ROC-optimal point density for two-dimensional image example.

3.8 Conclusion

Optimal quantization procedures for distributed hypothesis testing environments have been developed in this chapter. Conditions on the source densities under which quantizers exist that result in no loss in hypothesis testing performance have been presented. These conditions, given in Theorems 3.3 and 3.5, are rather restrictive and are not commonplace in practice. Given these conditions and the asymptotic hypothesis testing performance theory of Section 3.2.1 – including Stein's lemma and the exponential decay rates of the probabilities of type I and II errors – the need for discrimination-based objective functions is evident. Formulas for discrimination losses due to quantization by a many-point, small-cell vector quantizer have been determined and shown to resemble Bennet's integral formula for the reconstruction MSE. These formulas suggest that fine quantization in regions where the discriminability function is large is optimal for detection. Optimal small-cell quantizers for maximization of the discrimination between two sources and for maximization of ROC curve area have been derived under a congruent-cell assumption. Additionally, the optimal ellipsoidal-cell quantizers for these two objectives have been discussed. In the limit, as the number of quantizer cells becomes large, these quantizers preserve the log-likelihood ratio. This conclusion points to the optimality of the log-likelihood ratio quantizer for hypothesis testing performance. Accordingly, the optimal log-likelihood ratio quantizer has been derived by optimizing the point density of the scalar constituent quantizer. The estimation performance, as measured by reconstruction MSE, of the log-likelihood ratio quantizer has been shown to be poor, due to violation of the small-cell condition, which is required of estimation-optimal quantizers. By assuming the small cell condition, mixed objective functions have been optimized, which can be used to trade detection performance for estimation performance. Numerical examples of the various optimal quantizers have been presented for several types of scalar and two-dimensional sources. These examples have demonstrated the important concepts introduced in the chapter regarding detection

and estimation performance of vector quantizers. In particular, the discriminability function and, more generally, the Fisher covariation profile have been shown to have significant influence in the placement of codebook points in quantizers optimal for hypothesis testing.

APPENDICES

APPENDIX A

Derivation of Bound on Register Power Consumption

Consider a zero-mean wide-sense stationary Gaussian random sequence x_k, with autocorrelation sequence $R(\tau) = E[x_k x_{k+\tau}]$, uniformly quantized to B bits and loaded into a B-bit register, whose maximum and minimum values are $+1$ and -1, respectively. The quantizer cells are shown in Figure A.1.

Let n_k denote the number of bits that flip when loading the value x_{k+1} into the register. Then

$$E[n_k] = \sum_{i=1}^{B} i\ P(n_k = i) = \sum_{i=0}^{B} P(n_k > i).$$

Next we note that

$$\sum_{i=1}^{B} P(n_k = i) \leq \sum_{i=0}^{B} P(n_k > i) \leq \sum_{i=1}^{B} B \cdot P(n_k = i).$$

Therefore,

$$P(n_k > 0) \leq E[n_k] \leq B \cdot P(n_k > 0). \tag{A.1}$$

Now $P(n_k > 0)$ is equal to the probability that at least one bit flips when loading x_{k+1} into

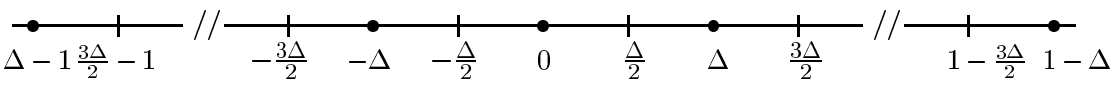

Figure A.1: Uniform scalar quantizer.

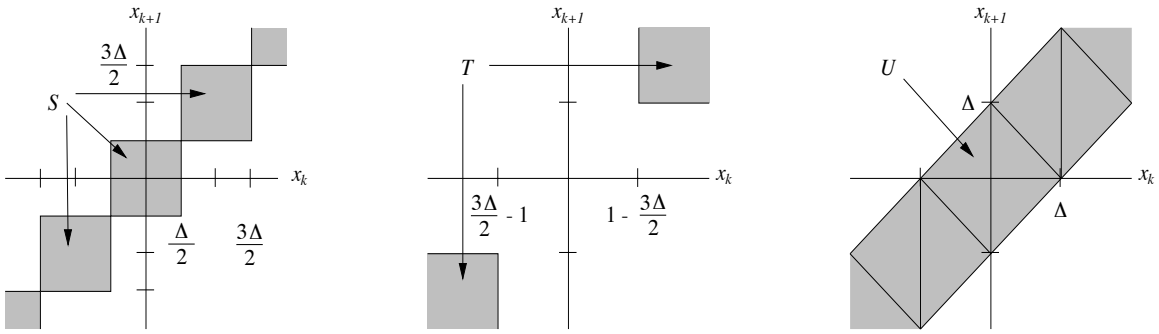

Figure A.2: Regions defined in equation (A.3).

the register. This is equal to one minus the probability that no bits flip. Therefore

$$P(n_k > 0) = 1 - P(n_k = 0). \tag{A.2}$$

No bits in the register will flip if x_k and x_{k+1} lie in the same quantizer cell. The granularity of the quantizer is $\Delta = 2^{-B+1}$. Define

$$\begin{aligned} S &= \bigcup_{m=-\infty}^{\infty} \{(x_k, x_{k+1}) \mid x_k, x_{k+1} \in (\Delta(m-1/2), \Delta(m+1/2))\} \\ T &= \{(x_k, x_{k+1}) \mid x_k, x_{k+1} \in (-\infty, 3\Delta/2 - 1) \cup (1 - 3\Delta/2, +\infty)\} \\ U &= \left\{(x_k, x_{k+1}) \mid |x_k - x_{k+1}| \leq \Delta/\sqrt{2}\right\}. \end{aligned} \tag{A.3}$$

These regions are shown in Figure A.2. Then $P(n_k = 0) = P(S \cup T)$ and

$$P(S) \leq P(n_k = 0) \leq P(S) + P(T). \tag{A.4}$$

For $P(S)$ we have,

$$P(S) = \sum_{m=-\infty}^{\infty} \int_{\Delta(m-1/2)}^{\Delta(m+1/2)} \int_{\Delta(m-1/2)}^{\Delta(m+1/2)} f_{x_k, x_{k+1}}(x, y)\, dxdy$$

where $f_{x_k, x_{k+1}}$ is the joint probability density function of x_k and x_{k+1}. For each integer m let $x_m = y_m = \Delta \cdot m$. By the Mean Value Theorem, for each m there exist $\bar{x}_m, \bar{y}_m$ such

that

$$\begin{aligned} P(S) &= \sum_{m=-\infty}^{\infty} f_{x_k,x_{k+1}}(\bar{x}_m,\bar{y}_m)\Delta^2 \\ &= \sum_{m=-\infty}^{\infty} f_{x_k,x_{k+1}}(x_m,y_m)\Delta^2 + \\ &\quad \sum_{m=-\infty}^{\infty} \left[f_{x_k,x_{k+1}}(\bar{x}_m,\bar{y}_m) - f_{x_k,x_{k+1}}(x_m,y_m)\right]\Delta^2 \end{aligned} \tag{A.5}$$

Next we note that

$$\begin{aligned} P(U) = \sum_{m=-\infty}^{\infty} \quad & \left[\int_{x=(m-1)\Delta}^{m\Delta}\int_{y=(2m-1)\Delta-x}^{x+\Delta} f_{x_k,x_{k+1}}(x,y)\,dxdy + \right. \\ & \left.\int_{x=m\Delta}^{(m+1)\Delta}\int_{y=x-\Delta}^{-x+(2m+1)\Delta} f_{x_k,x_{k+1}}(x,y)\,dxdy\right]. \end{aligned}$$

Again by the Mean Value Theorem, there exist x'_m, y'_m such that

$$\begin{aligned} P(U) &= \sum_{m=-\infty}^{\infty} f_{x_k,x_{k+1}}(x'_m,y'_m)(\Delta\sqrt{2})^2 \\ &= \sum_{m=-\infty}^{\infty} f_{x_k,x_{k+1}}(x_m,y_m)2\Delta^2 + \\ &\quad \sum_{m=-\infty}^{\infty} \left[f_{x_k,x_{k+1}}(x'_m,y'_m) - f_{x_k,x_{k+1}}(x_m,y_m)\right]2\Delta^2. \end{aligned} \tag{A.6}$$

From (A.5) and (A.6),

$$P(S) = \frac{P(U)}{2} + \delta \tag{A.7}$$

where $\delta = \Delta^2 \sum_{m=-\infty}^{\infty} \left[f_{x_k,x_{k+1}}(\bar{x}_m,\bar{y}_m) - f_{x_k,x_{k+1}}(x'_m,y'_m)\right]$. Now $P(T)$ can be bounded using the Chebyshev inequality:

$$\begin{aligned} P(T) &= P\left(|x_k| \geq 1-3\Delta/2, |x_{k+1}| \geq 1-3\Delta/2\right) \\ &\leq \gamma R(0) \end{aligned} \tag{A.8}$$

where $\gamma = (1-3\Delta/2)^{-2}$.

From (A.2), (A.4), (A.7), and (A.8)

$$1 - \frac{P(U)}{2} - (\delta + \gamma R(0)) \leq P(n_k > 0) \leq 1 - \frac{P(U)}{2} - \delta.$$

As $x_k - x_{k+1}$ is Gaussian with mean zero and variance $2R(0) - 2R(1)$, we have

$$P(U) \;=\; \mathrm{erf}\left(\left[2^B\sqrt{2R(0)-2R(1)}\right]^{-1}\right).$$

Next with η the power consumption per bit flip, we have $P_B = \eta E[n_k]$ where P_B is the average power consumed during register loading. Using this and (A.1), we obtain the following two bounds on P_B:

$$\begin{aligned} P_B &\leq B\eta\cdot\left[1-\frac{1}{2}\mathrm{erf}\left(\left[2^B\sqrt{2R(0)-2R(1)}\right]^{-1}\right)-\delta\right] \\ P_B &\geq \eta\cdot\left[1-\frac{1}{2}\mathrm{erf}\left(\left[2^B\sqrt{2R(0)-2R(1)}\right]^{-1}\right)-(\delta+\gamma R(0))\right]. \end{aligned}$$

Finally, since δ is bounded by the maximum variation in f and since $\delta \to 0$ and $\gamma \to 1$ as $\Delta \to 0$, the approximate bound (1.1) holds.

APPENDIX B

Iteration Power of LMS Algorithm

We consider a hardware implementation of the finite-precision LMS algorithm in which all data are stored in $B_d + 1$ bits and all coefficients in $B_c + 1$ bits. We assume that all right-shifted quantities are rounded correctly and that right shifting consumes negligible energy compared to additions and multiplications.

Two multipliers are used: one with B_c-bit multiplier and B_d-bit multiplicand, and one with B_d-bit multiplier and multiplicand. Only magnitudes are multiplied and determination of input magnitude and output sign is assumed to consume negligible energy. Each multiplier is assumed to be a direct table lookup multiplier, implemented by use of a ROM indexed by the magnitudes of the multiplier and multiplicand, that stores the product (magnitude) quantized to $B_1 + B_2$ bits where B_1 and B_2 are the number of bits used for the multiplier and multiplicand [11]. Although the power consumed by such a multiplier is proportional to $2^{B_1+B_2}$, we use a simplified formula and assume the power consumed during a (real) multiplication is $P_{B_1 \times B_2,\text{real}} = \eta_t(B_1 + B_2)$ where η_t is the power per bit of a table lookup operation.

A complex multiplication of $a + bj$ stored in B_1 bits plus sign by $c + dj$ stored in B_2 bits plus sign requires four real B_1-bit by B_2-bit multiplications and two additions of $B_1 + B_2 + 1$ bits. Therefore, the power consumed by such a complex multiplication is $P_{B_1 \times B_2,\text{complex}} = 4\eta_t(B_1 + B_2) + 2\eta_a(B_1 + B_2 + 1)$ where η_a is the power per bit of a real adder.

In the LMS update formula given by (2.2) and (2.3), the calculation of the inner product $\underline{w}_k'^H \underline{x}_k'$ requires p complex multiplications of numbers stored in B_c bits plus sign by numbers stored in B_d bits plus sign and $p-1$ complex (B_d+1)-bit additions. Subtracting this inner product from y_k' requires an additional complex addition of B_d+1 bits. Next, multiplying the conjugate of this quantity (e_k') by $\underline{x}_k'$ we have p complex multiplications in which both multiplier and multiplicand are stored in B_d bits plus sign. Since multiplication by μ requires only a shift, we are left with the addition by $\underline{w}_k'$. This operation requires p complex additions of B_c+1 bits. Therefore, the total power consumed during one iteration of the complex finite-precision LMS algorithm is

$$P_T = 4p\eta_t(3B_d+B_c)+2p\eta_a(3B_d+B_c+2)+2p\eta_a(B_d+B_c+2)$$

which is equivalent to (2.4).

We can derive a similar formula for multiplication using partial product accumulation. In such a multiplier, each real B_1-bit by B_2-bit multiplication requires approximately B_1 additions of B_2 bits [38] and therefore has a power consumption of $B_1B_2\eta_a$. The total real operations are the same as with the table lookup multiplier. This gives the following relation for LMS iteration power using partial product accumulation multiplication:

$$P_T = 4p\eta_a(B_d^2+B_dB_c+2B_d+B_c+2). \tag{B.1}$$

Using equations (2.4) and (B.1) along with (2.19), we can plot P_T as a function of B_T for both power relations. These plots are shown in Figure B.1 for $\rho = 1/2$, $\eta_a \approx 1.4$ mW, $\eta_t \approx 6.8$ mW, and $p = 2$. Note that the formula derived using partial product accumulation is quadratic in B_T while the formula derived using table lookup multiplication is linear in B_T.

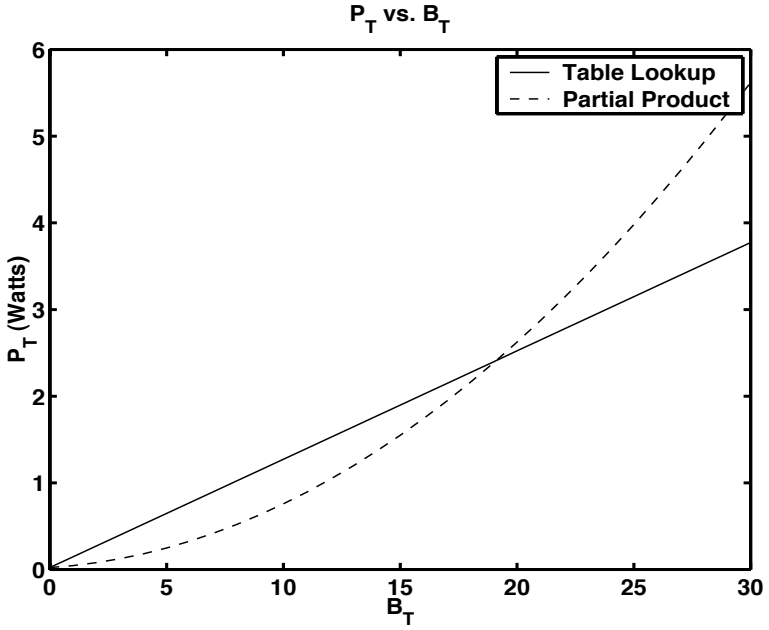

Figure B.1: P_T versus B_T for different power relations with $\eta_a \approx 1.4$ mW and $\eta_t \approx 6.8$ mW.

APPENDIX C

Derivation of Mean Convergence Rate, Weight-Error Covariance, and Excess Mean Square Error

From (2.2) and (2.3), we see that the finite precision LMS update is of the form

$$\underline{w}'_{k+1} = \underline{w}'_k + \mu \underline{x}'_k (y'^*_k - \underline{x}'^H_k \underline{w}'_k + \eta_k) + \underline{\sigma}_k$$

where η^*_k and $\underline{\sigma}_k$ are approximated as zero-mean, white, and uncorrelated with all other signals. The variance of η_k is $2p\sigma_d^2$ and the variance of each component of $\underline{\sigma}_k$ is $2\sigma_c^2$.

C.1 Mean Convergence Rate

Using the method of [65] we obtain (2.7). First define

$$\delta \underline{w}'_{k+1} = \underline{w}'_{k+1} - \underline{w}'_k = \mu \underline{x}'_k e'^*_k + \underline{\sigma}_k$$

and

$$\epsilon'_k = y'_k - \underline{w}'^{oH} \underline{x}'_k, \qquad \underline{\tilde{w}}'_k = \underline{w}'_k - \underline{w}'^o.$$

Then we have

$$e'_k \;\; = \;\; \epsilon'_k - \underline{\tilde{w}}'^H_k \underline{x}'_k + \eta^*_k.$$

Next, use $\underline{\tilde{w}}'_{k+1} - \underline{\tilde{w}}'_k = \delta \underline{w}'_{k+1}$ to obtain

$$\underline{\tilde{w}}'_{k+1} = (I - \mu \underline{x}'_k \underline{x}'^H_k) \underline{\tilde{w}}'_k + \mu \underline{x}'_k \epsilon'^*_k + \mu \underline{x}'_k \eta_k + \underline{\sigma}_k. \tag{C.1}$$

Similar to [65], we assume that $\underline{x}'_k$ and ϵ'_k are independent. Using this and the fact that η_k and $\underline{\sigma}_k$ have zero means, we see that

$$E[\underline{\tilde{w}}'_{k+1}|\underline{x}'_1,\ldots,\underline{x}'_{k-1},\epsilon'_1,\ldots,\epsilon'_{k-1}] = \underline{\tilde{w}}'_k - \mu R_{x'}\underline{\tilde{w}}'_k$$

and taking expectations,

$$E[\underline{\tilde{w}}'_{k+1}] = (I - \mu R_{x'})E[\underline{\tilde{w}}'_k]$$

which is completely analogous to the infinite-precision LMS algorithm. Using the same analysis as in [65, Sec. 5.2], we conclude that $E[\underline{w}'_k] \to \underline{w}'^o$ as long as condition (2.6) is met.

C.2 Steady-State Weight-Error Covariance and Excess Mean Square Error

We begin our derivation of (2.8) and (2.10) in the manner of [15]. We first write the weight update formula for the infinite-precision LMS algorithm

$$\underline{w}_{k+1} = \underline{w}_k + \mu\ \underline{x}_k \left(y_k^* - \underline{x}_k^H \underline{w}_k\right). \tag{C.2}$$

We define

$$\sigma_d^2 = \frac{1}{12}2^{-2B_d}, \qquad \sigma_c^2 = \frac{1}{12}2^{-2B_c}. \tag{C.3}$$

These are the variances of the quantization noises associated with quantizers having B_d and B_c bits (plus sign), respectively. We use primed symbols to represent quantized values and define the following quantities

$$\begin{aligned} \underline{x}'_k &= \underline{x}_k + \underline{\alpha}_k \\ y'_k &= y_k + \beta_k \\ \underline{w}'_k &= \underline{w}_k + \underline{\rho}_k. \end{aligned} \tag{C.4}$$

The components of the vector $\underline{\alpha}_k$ and β_k are complex numbers whose real and imaginary parts are assumed uncorrelated and have variances σ_d^2. Therefore, the variance of β_k and of each component of $\underline{\alpha}_k$ is $2\sigma_d^2$.

Thus we have the representation for the quantized value of the filter output

$$\hat{y}'_k = Q_d(\underline{w}'^H_k \underline{x}'_k) = \underline{w}^H_k \underline{x}_k + \underline{\rho}^H_k \underline{x}_k + \underline{w}^H_k \underline{\alpha}_k + \eta_k \tag{C.5}$$

where η_k is defined above. As in [15], we have ignored the noise product term $\underline{\rho}^H_k \underline{\alpha}_k$ in (C.5), as its power is of order $\sigma_d^2 \sigma_c^2$.

The total error is now

$$s_k - \hat{s}'_k = \frac{1}{a}(y_k - \underline{w}^H_k \underline{x}_k) - \frac{1}{a}(\underline{\rho}^H_k \underline{x}_k + \underline{w}^H_k \underline{\alpha}_k + \eta_k). \tag{C.6}$$

The first term on the right-hand side of (C.6) is the error of the infinite-precision algorithm. The second term is the error due to quantization and will be denoted e_q. Under the hypothesis (2.5), these terms are uncorrelated and the MSE for the finite-precision LMS algorithm is

$$\xi = \frac{1}{a^2} E[|y_k - \underline{w}^H_k \underline{x}_k|^2] + \frac{1}{a^2} E[|\underline{\rho}^H_k \underline{x}_k + \underline{w}^H_k \underline{\alpha}_k + \eta_k|^2]. \tag{C.7}$$

The first term on the right-hand side of (C.7) is the MSE of the infinite-precision LMS algorithm and is equal to $\xi_{\min} + \xi_{\text{excess}}$ [65, 74]. The second term is the excess MSE due to quantization

$$\xi_q = E[|e_q|^2].$$

Under the assumptions (2.5), $\underline{\alpha}_k$, β_k, $\underline{\rho}_k$, and η_k are all uncorrelated and

$$\xi_q = \frac{1}{a^2}\left(E[|\underline{\rho}^H_k \underline{x}_k|^2] + E[|\underline{w}^H_k \underline{\alpha}_k|^2] + E[|\eta_k|^2]\right). \tag{C.8}$$

The term $E[|\eta_k|^2]$ is equal to $2p\ \sigma_d^2$. For the term $E[|\underline{w}^H_k \underline{\alpha}_k|^2]$ we obtain

$$E[|\underline{w}^H_k \underline{\alpha}_k|^2] = 2\sigma_d^2\ E[\|\underline{w}_k\|^2].$$

Note that this differs from the real case studied in [15] by a factor of two. In the steady state, this becomes

$$E[|\underline{w}^H_k \underline{\alpha}_k|^2] = 2\sigma_d^2 \left(\|\underline{w}^o\|^2 + \frac{1}{2} p\mu\xi_{\min}\right). \tag{C.9}$$

Finally, for the first term in (C.8)

$$E[|\underline{\rho}_k^H \underline{x}_k|^2] = \mathrm{tr}(R_x P_k) \tag{C.10}$$

where $P_k = E[\underline{\rho}_k \underline{\rho}_k^H]$.

To derive an expression for P_k, we assume μ is small and use the averaged system techniques of [65, Sec. 9.2]. Define $\underline{\bar{w}}'_k$ as the weight vector at time k of the averaged finite-precision system. Similarly, let $\underline{\bar{w}}_k$ be the averaged infinite-precision system weight vector. Then the finite-precision primary system obeys the recursion (C.1)

$$\underline{\tilde{w}}'_{k+1} = (I - \mu \underline{x}'_k \underline{x}'^H_k)\underline{\tilde{w}}'_k + \mu \underline{x}'_k \epsilon'^*_k + \mu \underline{x}'_k \eta_k + \underline{\sigma}_k.$$

while the finite-precision averaged system recursion is

$$\underline{\bar{w}}'_{k+1} = (I - \mu R_{x'})\underline{\bar{w}}'_k. \tag{C.11}$$

Next define $\underline{u}_k = (\underline{w}_k - \underline{\bar{w}}_k)/\sqrt{\mu}$ and $\underline{u}'_k = (\underline{w}'_k - \underline{\bar{w}}'_k)/\sqrt{\mu}$. Finally, define

$$\begin{aligned} \Gamma &= \lim_{k\to\infty} E[\underline{u}_k \underline{u}_k^H] \\ \Gamma' &= \lim_{k\to\infty} E[\underline{u}'_k \underline{u}'^H_k]. \end{aligned}$$

Now write

$$\underline{w}'_k - \underline{\bar{w}}'_k = (\underline{w}'_k - \underline{w}_k) + (\underline{w}_k - \underline{\bar{w}}_k) + (\underline{\bar{w}}_k - \underline{\bar{w}}'_k). \tag{C.12}$$

Assuming that the first two terms on the right-hand side of (C.12) are uncorrelated and that the last term is small we have

$$P = \lim_{k\to\infty} P_k = \mu(\Gamma' - \Gamma). \tag{C.13}$$

To apply the averaged system techniques of [65] to the finite-precision algorithm, it is necessary that the following conditions hold

$$\begin{aligned} \lim_{\mu\to 0} \Delta_d &= 0 \\ \lim_{\mu\to 0} \Delta_c &= 0 \end{aligned} \tag{C.14}$$

where $\Delta_c = 2^{-B_c}$ and $\Delta_d = 2^{-B_d}$. In the fixed-point, power-of-two algorithm, the magnitude of $\underline{x}'_k e'^*_k$ is stored in $2B_d$ bits before shifting to the right by $q = -\log_2 \mu$ bits. Therefore, we require $q \leq 2B_d$. To avoid slowdown we require $B_c > B_d + \nu$. These requirements ensure that (C.14) holds.

Assuming ϵ_k is white and independent of $\underline{x}_k$, we have

$$\Gamma = \frac{\xi_{\min}}{2} I. \tag{C.15}$$

Proceeding as in [65] for the finite-precision system, assuming ϵ' is white, using (C.1) and (C.11) we find that Γ' satisfies the following equation

$$R_{x'}\Gamma' + \Gamma' R_{x'} = (\sigma^2_{\epsilon'} + 2p\sigma^2_d) R_{x'} + \frac{2}{\mu^2}\sigma^2_c I$$

where $\sigma^2_{\epsilon'} = E[|\epsilon'|^2]$. It is easy to show that $R_{x'}$ and Γ' have the same eigenvectors and

$$\Gamma' = \frac{1}{2}(\sigma^2_{\epsilon'} + 2p\sigma^2_d) I + \frac{2}{\mu^2}\sigma^2_c R^{-1}_{x'}.$$

Next, using $\sigma^2_{\epsilon'} = R^H_{x'y'} R^{-1}_{x'} R_{x'y'} \approx \xi_{\min} + 2\sigma^2_d$ and (C.13) gives

$$P = \mu(p+1)\sigma^2_d I + \frac{1}{\mu}\sigma^2_c R^{-1}_{x'}. \tag{C.16}$$

Finally, using $R_{x'} \approx R_x$ in (C.16) yields

$$\mathrm{tr}(R_x P) = \mu(p+1)\sigma^2_d \mathrm{tr}(R_x) + \frac{p\sigma^2_c}{\mu}. \tag{C.17}$$

As the first term on the right hand side of (C.17) is of order $\mu\sigma^2_d$, it can be ignored.

Now, using (C.8), (C.9), and (C.17) we get

$$\xi_q = \frac{1}{a^2}\left[\frac{p\ \sigma^2_c}{\mu} + 2\sigma^2_d\ (\|\underline{w}^o\|^2 + p)\right]$$

from which, by using (C.3), we obtain (2.10).

APPENDIX D

Derivation of Optimal Bit Allocation Factors

D.1 Total Bit Budget

Using equations (2.10) and (2.19) we can write the excess MSE due to quantization as

$$\xi_q = \alpha_c 2^{-2(1-\rho)B_T} + \alpha_d 2^{-2\rho B_T}. \tag{D.1}$$

Differentiating this equation with respect to ρ we get

$$\frac{d\xi_q}{d\rho} = 2\ln 2\ B_T\ \left(\alpha_c 2^{-2(1-\rho)B_T} - \alpha_d 2^{-2\rho B_T}\right). \tag{D.2}$$

Differentiating again we get

$$\frac{d^2\xi_q}{d\rho^2} = (2\ln 2\ B_T)^2\ \xi_q > 0. \tag{D.3}$$

Equation (D.3) shows that for $\rho \in [0,1]$, $\frac{d^2\xi_q}{d\rho^2} > 0$ and therefore ξ_q is a convex function of ρ. Now, setting $\frac{d\xi_q}{d\rho}$ in (D.2) equal to zero at $\rho = \rho^*$, we have

$$\frac{\alpha_d}{\alpha_c} = 2^{4\rho^* B_T - 2B_T}. \tag{D.4}$$

Equation (D.4) leads directly to (2.20).

D.2 Total Power Budget

To derive (2.23) we first define the following constants

$$A = P_T - 8p\eta_a, \qquad B = 8p(\eta_a + \eta_t), \qquad C = 2p\eta_a. \tag{D.5}$$

Then from (2.22) we have

$$B_T = \frac{A}{B\rho + C}. \tag{D.6}$$

Differentiating (D.1) with respect to ρ gives

$$\begin{aligned} \frac{d\xi_q}{d\rho} &= \ln 2 \; \alpha_c \; 2^{-2(1-\rho)B_T} \cdot \frac{d}{d\rho}[-2(1-\rho)B_T] \quad + \\ & \qquad \ln 2 \; \alpha_d \; 2^{-2\rho \; B_T} \cdot \frac{d}{d\rho}[-2\rho \; B_T]. \end{aligned} \tag{D.7}$$

Differentiating again and using (D.6), we have

$$\begin{aligned} \frac{d^2\xi_q}{d\rho^2} &= \ln 2 \; \alpha_c \; 2^{-2(1-\rho)B_T} \left[-\frac{4AB(B+C)}{(B\rho+C)^3} + \ln 2 \cdot \frac{4A^2(B+C)^2}{(B\rho+C)^4} \right] + \\ & \qquad \ln 2 \; \alpha_d \; 2^{-2\rho B_T} \left[\frac{4ABC}{(B\rho+C)^3} + \ln 2 \cdot \frac{4A^2C^2}{(B\rho+C)^4} \right]. \end{aligned} \tag{D.8}$$

Now, ξ_q is convex if $\frac{d^2\xi_q}{d\rho^2}$ is positive. From (D.8), it is clear that $\frac{d^2\xi_q}{d\rho^2}$ will be positive if the following condition is met:

$$\ln 2 \cdot \frac{4A^2(B+C)^2}{(B\rho+C)^4} > \frac{4AB(B+C)}{(B\rho+C)^3}.$$

Using (D.5) and (D.6), this condition will be satisfied if and only if the following condition is satisfied:

$$B_T > \frac{B}{\ln 2 (B+C)} = \frac{1}{\ln 2} \left[1 - \frac{\eta_a + \eta_t}{2\eta_a + 3\eta_t} \right]. \tag{D.9}$$

The term on the right-hand side of (D.9) is clearly less than $1/\ln 2$. This means that $\frac{d^2\xi_q}{d\rho^2}$ will be positive if $B_T > 1/\ln 2$. This condition is clearly true under the assumption $B_T \geq 2$. Therefore, ξ_q is once again a convex function of ρ. To solve for ρ^{**} we set $\frac{d\xi_q}{d\rho}$ equal to zero. Using (D.7) we have

$$\alpha_c \; 2^{-2(1-\rho^{**}) \cdot \frac{A}{B\rho^{**}+C}} \cdot \frac{2A(B+C)}{(B\rho^{**}+C)^2} = \alpha_d \; 2^{-2\rho^{**} \cdot \frac{A}{B\rho^{**}+C}} \cdot \frac{2AC}{(B\rho^{**}+C)^2}.$$

This implies

$$\rho^{**} = \frac{2A + C \log_2 \left(\frac{\alpha_d}{\alpha_c} \cdot \frac{C}{B+C} \right)}{4A - B \log_2 \left(\frac{\alpha_d}{\alpha_c} \cdot \frac{C}{B+C} \right)}. \tag{D.10}$$

Equations (D.5) and (D.10) lead to (2.23).

APPENDIX E

Derivation of Asymptotic Discrimination Losses

E.1 Asymptotic Loss in Discrimination Between Two Sources

To derive the asymptotic loss in discrimination (3.28) between q_0 and q_1, we follow the "sequence approach" used in [13, 14, 46]. Consider a sequence of quantizers $Q_N = (\mathcal{S}_N, \mathcal{C}_N)$ where the Nth quantizer contains the N cells $\mathcal{S}_N = \{S_{N,1}, \ldots, S_{N,N}\}$ and the N codebook points $\mathcal{C}_N = \{x_{N,1}, \ldots, x_{N,N}\}$. Define the cell probability sequences

$$\begin{aligned} \bar{q}_{0,N,i} &= \int_{S_{N,i}} q_0(y)dy \\ \bar{q}_{1,N,i} &= \int_{S_{N,i}} q_1(y)dy \end{aligned}$$

for $i = 1, \ldots, N$. Note that for each N, the sets $\bar{q}_{0,N} = \{\bar{q}_{0,N,1}, \ldots, \bar{q}_{0,N,N}\}$ and $\bar{q}_{1,N} = \{\bar{q}_{1,N,1}, \ldots, \bar{q}_{1,N,N}\}$ form the quantized source probability mass functions.

The log-likelihood ratio $\Lambda(x)$ is defined as

$$\Lambda(x) = \log \frac{q_0(x)}{q_1(x)}$$

and the sequence of log-likelihood ratios after quantization is

$$\bar{\Lambda}_{N,i} = \log \frac{\bar{q}_{0,N,i}}{\bar{q}_{1,N,i}}.$$

The discrimination before quantization can be written in terms of the cells of the Nth quantizer:

$$L \triangleq L(q_0 \| q_1) = \sum_{i=1}^{N} \int_{S_{N,i}} q_0(y)\Lambda(y)dy.$$

The discrimination after quantization by the Nth quantizer can be written as

$$\hat{L}_N \triangleq L(\bar{q}_{0,N} \| \bar{q}_{1,N}) = \sum_{i=1}^{N} \bar{q}_{0,N,i} \bar{\Lambda}_{N,i}.$$

Since our goal is to maximize the discrimination after quantization, we will refer to the loss in discrimination as distortion. It is well known that discrimination can not increase with processing (i.e. quantization). Thus, the distortion is nonnegative. The distortion resulting from the Nth quantizer is thus

$$\Delta L_N \triangleq L - \hat{L}_N = \sum_{i=1}^{N} \int_{S_{N,i}} q_0(y)\Lambda(y)dy - \bar{q}_{0,N,i}\bar{\Lambda}_{N,i}. \tag{E.1}$$

Note that (E.1) is independent of the codebook $\mathcal{C}_N$. Therefore, we lose no generality by assuming that the codebook points are the centroids of their cells. That is, for each N

$$x_{N,i} = \frac{\int_{S_{N,i}} y dy}{V_{N,i}}, \ i = 1, \ldots, N \tag{E.2}$$

where $V_{N,i}$ is the volume of the ith cell in the Nth quantizer. Note that (E.2) implies

$$\int_{S_{N,i}} (y - x_{N,i}) dy = 0, \ i = 1, \ldots, N.$$

E.1.1 Sequence Definitions

We define a few more sequences that will be necessary in analyzing the asymptotic behavior of the quantizer sequence.

1. The sequence of diameter functions is $d_N(x)$.

2. The sequence of specific inertial profile functions is $m_N(x)$.

3. The sequence of specific covariation profile functions is $M_N(x)$. We will write $M_{N,i} = M_N(x)$ for $x \in S_{N,i}$.

4. The sequence of specific point density functions is $\zeta_N(x) = \zeta_{N,i} = 1/(NV_{N,i})$ for $x \in S_{N,i}$.

E.1.2 Assumptions

We make several assumptions regarding the quantizers and the convergence of the sequences defined in the previous section.

1. For all N, each codebook point lies in the centroid of its cell. That is, equation (E.2) holds.

2. The sequence of diameter functions $d_N(x)$ converges uniformly to zero.

3. The sequence of specific inertial profile functions $m_N(x)$ converges uniformly to a function $m(x)$, called the inertial profile, that is uniformly bounded by m_B.

4. The sequence of specific covariation profile matrix functions $M_N(x)$ converges uniformly to a full-rank matrix function $M(x)$, called the covariation profile.

5. The sequence of specific point density functions $\zeta_N(x)$ converges uniformly to a function $\zeta(x)$, called the point density, that satisfies $\int \zeta(x)dx = 1$.

E.1.3 Notation

To facilitate the analysis, we define some simplifying notation here. The density functions evaluated at codebook point $x_{N,i}$ will be denoted

$$\begin{aligned} q_{0,N,i} &= q_0(x_{N,i}) \\ q_{1,N,i} &= q_1(x_{N,i}). \end{aligned}$$

Similarly, the gradients and Hessians of q_0 and q_1 evaluated at $x_{N,i}$ will be denoted

$$\begin{aligned} \nabla_{0,N,i} &= \nabla q_0(x_{N,i}) \\ \nabla_{1,N,i} &= \nabla q_1(x_{N,i}) \\ \nabla^2_{0,N,i} &= \nabla^2 q_0(x_{N,i}) \\ \nabla^2_{1,N,i} &= \nabla^2 q_1(x_{N,i}) \end{aligned}$$

and the log-likelihood ratio evaluated at $x_{N,i}$ is

$$\Lambda_{N,i} = \Lambda(x_{N,i}).$$

The following matrix functions will be useful in our analysis. The "Fisher" matrix function is defined to be the outer product of the log-likelihood ratio gradient:

$$F(x) = \nabla\Lambda(x)\nabla\Lambda(x)^T$$

and the matrix function $G(x)$ is

$$G(x) = \frac{\nabla^2 q_0(x)}{q_0(x)} - \frac{\nabla^2 q_1(x)}{q_1(x)}. \tag{E.3}$$

In keeping with the convention set forth above, we define

$$\begin{aligned} F_{N,i} &= F(x_{N,i}) \\ G_{N,i} &= G(x_{N,i}). \end{aligned} \tag{E.4}$$

E.1.4 Taylor Expansions

For all N, we can expand the function $q_0(x)$ in a Taylor series about the codebook points of quantizer Q_N. Therefore, for all N we can write

$$\begin{aligned} q_0(x) &= q_{0,N,i} + \nabla_{0,N,i}^T(x - x_{N,i}) + \frac{1}{2}(x - x_{N,i})^T\nabla_{0,N,i}^2(x - x_{N,i}) \\ &\quad + o(\|x - x_{N,i}\|^2),\ \forall x \in S_{N,i}. \end{aligned} \tag{E.5}$$

A similar expansion can be done for $q_1(x)$ and $\Lambda(x)$ as shown below:

$$\begin{aligned} \Lambda(x) &= \Lambda_{N,i} + \nabla\Lambda_{N,i}^T(x - x_{N,i}) + \frac{1}{2}(x - x_{N,i})^T\nabla^2\Lambda_{N,i}(x - x_{N,i}) \\ &\quad + o(\|x - x_{N,i}\|^2),\ \forall x \in S_{N,i}. \end{aligned} \tag{E.6}$$

The "o" terms in (E.5) and (E.6) are explained as follows. From the definition of the diameter function, we have $\|x - Q_N(x)\| \leq d_N(x)$ for all N and from Assumption 2 we have $\|x - Q_N(x)\| \to 0$ uniformly. Therefore, given $\epsilon > 0$ there is an integer N_0 such that for all $N \geq N_0$ and for all $x \in S_{N,i}$

$$\frac{o(\|x - x_{N,i}\|^2)}{\|x - x_{N,i}\|^2} < \epsilon.$$

E.1.5 Single-Cell Distortion

The distortion of the Nth quantizer given by (E.1) is a sum over the N quantizer cells of the quantity $\int_{S_{N,i}} q_0(y)\Lambda(y)dy - \bar{q}_{0,N,i}\bar{\Lambda}_{N,i}$. We call this term the single-cell distortion of cell $S_{N,i}$. The bulk of the analysis required to determine the distortion involves studying the single-cell distortion, which we do in this section.

Using (E.5) and (E.6) along with Assumption 1 we have

$$\begin{aligned}\int_{S_{N,i}} q_0(y)\Lambda(y)dy &= q_{0,N,i}\Lambda_{N,i}V_{N,i} + \int_{S_{N,i}} (y-x_{N,i})^T A_{N,i}(y-x_{N,i})dy \\ &\quad + \int_{S_{N,i}} o(\|y-x_{N,i}\|^2)dy \end{aligned} \tag{E.7}$$

where

$$A_{N,i} = \frac{1}{2}\left[\Lambda_{N,i}\nabla^2_{0,N,i} + q_{0,N,i}\nabla^2\Lambda_{N,i} + \nabla_{0,N,i}\nabla\Lambda^T_{N,i} + \nabla\Lambda_{N,i}\nabla^T_{0,N,i}\right]. \tag{E.8}$$

The last two terms in (E.8) arise due to the fact that the matrix in a quadratic form may be transposed without affecting the result [36]. After some algebra, (E.8) can be written

$$A_{N,i} = \frac{1}{2}\left[\Lambda_{N,i}\nabla^2_{0,N,i} + q_{0,N,i}(F_{N,i} + G_{N,i})\right] \tag{E.9}$$

where $F_{N,i}$ and $G_{N,i}$ are given in (E.4).

To simplify (E.7), we first focus on the last term. For $\epsilon > 0$ there is an integer N_0 such that for all $N \geq N_0$, the following two conditions hold:

$$\frac{o(\|y-x_{N,i}\|^2)}{\|y-x_{N,i}\|^2} \leq \frac{\epsilon}{2m_B},\ \forall y \in S_{N,i}$$

and

$$\begin{aligned}|m_N(y) - m(y)| &\leq m_B, \\ \Rightarrow m_N(y) &\leq m(y) + m_B \leq 2m_B,\ \forall y \in S_{N,i}.\end{aligned}$$

Therefore, for all $N \geq N_0$,

$$\begin{aligned}\left|\int_{S_{N,i}} o(\|y - x_{N,i}\|^2)dy\right| &\leq \int_{S_{N,i}} \left|o(\|y - x_{N,i}\|^2)\right| dy \\ &\leq \int_{S_{N,i}} \frac{\epsilon}{2m_B}\|y - x_{N,i}\|^2 dy \\ &= \frac{\epsilon}{2m_B} \cdot m_N(x) V_{N,i}^{1+2/k}, \ \forall x \in S_{N,i} \\ &\leq \epsilon \cdot V_{N,i}^{1+2/k}.\end{aligned}$$

Therefore, the sequence

$$\frac{\left|\int_{S_{N,i}} o(\|y - x_{N,i}\|^2)dy\right|}{V_{N,i}^{1+2/k}}$$

converges to zero and we will thus write

$$\int_{S_{N,i}} o(\|y - x_{N,i}\|^2)dy = o\left(V_{N,i}^{1+2/k}\right).$$

Next, we rewrite the second term on the right-hand side of (E.7) as

$$\int_{S_{N,i}} (y - x_{N,i})^T A_{N,i}(y - x_{N,i})dy = \mathrm{tr}(A_{N,i}M_{N,i})V_{N,i}^{1+2/k}.$$

Therefore (E.7) becomes

$$\int_{S_{N,i}} q_0(y)\Lambda(y)dy = q_{0,N,i}\Lambda_{N,i}V_{N,i} + \mathrm{tr}(A_{N,i}M_{N,i})V_{N,i}^{1+2/k} + o\left(V_{N,i}^{1+2/k}\right). \tag{E.10}$$

We now turn our attention to the term $\bar{q}_{0,N,i}\bar{\Lambda}_{N,i}$ found in (E.1). From (E.5) and (E.6) we have

$$\bar{q}_{0,N,i}\bar{\Lambda}_{N,i} = q_{0,N,i}\bar{\Lambda}_{N,i}V_{N,i} + \mathrm{tr}(\hat{A}_{N,i}M_{N,i})V_{N,i}^{1+2/k} + o\left(V_{N,i}^{1+2/k}\right) \tag{E.11}$$

where

$$\hat{A}_{N,i} = \frac{1}{2}\bar{\Lambda}_{N,i}\nabla_{0,N,i}^2. \tag{E.12}$$

Combining (E.10) and (E.11) yields

$$\begin{aligned}\int_{S_{N,i}} q_0(y)\Lambda(y)dy - \bar{q}_{0,N,i}\bar{\Lambda}_{N,i} &= q_{0,N,i}\left(\Lambda_{N,i}-\bar{\Lambda}_{N,i}\right)V_{N,i} + \\ &\quad \frac{1}{2}\left(\Lambda_{N,i}-\bar{\Lambda}_{N,i}\right)\mathrm{tr}\left(\nabla^2_{0,N,i}M_{N,i}\right)V_{N,i}^{1+2/k} + \\ &\quad \frac{1}{2}q_{0,N,i}\ \mathrm{tr}\left([F_{N,i}+G_{N,i}]M_{N,i}\right)V_{N,i}^{1+2/k} + \\ &\quad o\left(V_{N,i}^{1+2/k}\right). \end{aligned} \tag{E.13}$$

From the definitions of $\Lambda_{N,i}$ and $\bar{\Lambda}_{N,i}$ we have

$$\Lambda_{N,i}-\bar{\Lambda}_{N,i} = \log\left(\frac{q_{0,N,i}\cdot\bar{q}_{1,N,i}}{q_{1,N,i}\cdot\bar{q}_{0,N,i}}\right).$$

Using the Taylor expansion

$$\log a = (a-1)-\frac{1}{2}(a-1)^2+o(|a-1|^2)$$

we have

$$\Lambda_{N,i}-\bar{\Lambda}_{N,i} = (l-1)-\frac{1}{2}(l-1)^2+o(|l-1|^2)$$

where

$$l = \frac{q_{0,N,i}\cdot\bar{q}_{1,N,i}}{q_{1,N,i}\cdot\bar{q}_{0,N,i}}.$$

Next, using (E.5)

$$l = \frac{q_{0,N,i}q_{1,N,i}V_{N,i}+\frac{1}{2}q_{0,N,i}\mathrm{tr}(\nabla^2_{1,N,i}M_{N,i})V_{N,i}^{1+2/k}+o\left(V_{N,i}^{1+2/k}\right)}{q_{0,N,i}q_{1,N,i}V_{N,i}+\frac{1}{2}q_{1,N,i}\mathrm{tr}(\nabla^2_{0,N,i}M_{N,i})V_{N,i}^{1+2/k}+o\left(V_{N,i}^{1+2/k}\right)}$$

and

$$l-1 = \frac{1}{2q_{1,N,i}}\mathrm{tr}(\nabla^2_{1,N,i}M_{N,i})V_{N,i}^{2/k}-\frac{1}{2q_{0,N,i}}\mathrm{tr}(\nabla^2_{0,N,i}M_{N,i})V_{N,i}^{2/k}+o\left(V_{N,i}^{2/k}\right). \tag{E.14}$$

Therefore, $(l-1)^2 = o\left(V_{N,i}^{2/k}\right)$ and using (E.14) and (E.3) we get

$$\Lambda_{N,i}-\bar{\Lambda}_{N,i} = -\frac{1}{2}\mathrm{tr}(G_{N,i}M_{N,i})V_{N,i}^{2/k}+o\left(V_{N,i}^{2/k}\right). \tag{E.15}$$

Finally, (E.13) and (E.15) give

$$\begin{aligned}\int_{S_{N,i}} q_0(y)\Lambda(y)dy - \bar{q}_{0,N,i}\bar{\Lambda}_{N,i} &= \frac{1}{2}q_{0,N,i}\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{1+2/k} + o\left(V_{N,i}^{1+2/k}\right) \\ &= \frac{1}{2}q_{0,N,i}\mathrm{tr}(F_{N,i}M_{N,i})\frac{V_{N,i}}{N^{2/k}\zeta_{N,i}^{2/k}} + \\ &\quad o\left(V_{N,i}^{1+2/k}\right). \end{aligned} \tag{E.16}$$

E.1.6 Total Distortion

Having calculated the single-cell distortion (E.16), the total distortion is obtained by summing over all quantizer cells. Using (E.1) and (E.16), the total distortion of quantizer Q_N is

$$\Delta L_N = \frac{1}{2N^{2/k}}\sum_{i=1}^{N} q_{0,N,i}\mathrm{tr}(F_{N,i}M_{N,i})\frac{1}{\zeta_{N,i}^{2/k}}V_{N,i} + o\left(\frac{1}{N^{2/k}}\right)V_{N,i}.$$

Multiplying by $N^{2/k}$ and taking the limit, we obtain (3.28).

E.2 Asymptotic Loss in Discrimination Between Each Source and the Tilted Source

In this section, we once again use a sequence approach to determine the asymptotic loss in discrimination – this time between q_λ and q_0 – due to quantization. Many of the quantities defined in Section E.1 will be used in this section as well. The sequence definitions and assumptions made in Sections E.1.1 and E.1.2, respectively, will all apply here, as well as the notation defined in Section E.1.3. In addition, we will define several new quantities throughout this section.

We begin by writing the loss in discrimination between the tilted source q_λ and source q_0 due to quantization with an N-point vector quantizer as

$$\begin{aligned}\Delta L_{0,N} &\triangleq L(q_\lambda\|q_0) - L(\hat{q}_{\lambda,N}\|\bar{q}_{0,N}) \\ &= \sum_{i=1}^{N}\int_{S_{N,i}} q_\lambda(x)\Lambda_0(x)dx - \hat{q}_{\lambda,N,i}\hat{\Lambda}_{0,N,i} \end{aligned} \tag{E.17}$$

where

$$\Lambda_0(x) = \log \frac{q_\lambda(x)}{q_0(x)}, \qquad \hat{\Lambda}_{0,N,i} = \log \frac{\hat{q}_{\lambda,N,i}}{\bar{q}_{0,N,i}}.$$

E.2.1 Notation

In keeping with the notational convention of Section E.1.3 we define

$$\begin{aligned} q_{\lambda,N,i} &= q_\lambda(x_{N,i}) \\ \nabla_{\lambda,N,i} &= \nabla q_\lambda(x_{N,i}) \\ \nabla^2_{\lambda,N,i} &= \nabla^2 q_\lambda(x_{N,i}) \end{aligned}$$

and

$$\Lambda_{0,N,i} = \Lambda_0(x_{N,i}).$$

Next we define

$$\begin{aligned} \mu &= \int q_0(x)^{1-\lambda} q_1(x)^\lambda dx = \sum_{i=1}^N \mu_{N,i} \\ \mu_{N,i} &= \int_{S_{N,i}} q_0(x)^{1-\lambda} q_1(x)^\lambda dx = \mu \int_{S_{N,i}} q_\lambda(x) dx \\ d_{N,i} &= \bar{q}^{1-\lambda}_{0,N,i} \cdot \bar{q}^{\lambda}_{1,N,i} - \mu_{N,i} \\ d_N &= \sum_{i-1}^N d_{N,i}. \end{aligned} \tag{E.18}$$

Thus we can write

$$\hat{q}_{\lambda,N,i} = \frac{\mu_{N,i} + d_{N,i}}{\mu + d_N}. \tag{E.19}$$

E.2.2 Expansions of $\mu_{N,i}$ and $d_{N,i}$

Expanding $q_\lambda(x)$ in a Taylor series about $x_{N,i}$ we get the following representation for $\mu_{N,i}$:

$$\mu_{N,i} = \mu q_{\lambda,N,i} V_{N,i} + \frac{\mu}{2} \int_{S_{N,i}} (x - x_{N,i})^T \nabla^2_{\lambda,N,i} (x - x_{N,i}) dx + o\left(V_{N,i}^{1+2/k}\right). \tag{E.20}$$

It can be straightforwardly shown that the Hessian of the tilted density is

$$\nabla^2 q_\lambda(x) = q_\lambda(x)\left[\lambda\frac{\nabla^2 q_1(x)}{q_1(x)} + (1-\lambda)\frac{\nabla^2 q_0(x)}{q_0(x)} - \lambda(1-\lambda)F(x)\right]. \tag{E.21}$$

Next, using the centroid assumption, we write

$$\begin{aligned}
\bar{q}_{0,N,i} &= q_{0,N,i}V_{N,i} + \frac{1}{2}\int_{S_{N,i}} (x-x_{N,i})^T \nabla^2_{0,N,i}(x-x_{N,i})dx + o\left(V_{N,i}^{1+2/k}\right) \\
\bar{q}_{1,N,i} &= q_{1,N,i}V_{N,i} + \frac{1}{2}\int_{S_{N,i}} (x-x_{N,i})^T \nabla^2_{1,N,i}(x-x_{N,i})dx + o\left(V_{N,i}^{1+2/k}\right)
\end{aligned} \tag{E.22}$$

and using the Taylor expansion

$$(x+y)^a = x^a + ax^{a-1}y + \frac{1}{2}a(a-1)x^{a-2}y^2 + o(y^2) \tag{E.23}$$

we obtain

$$\begin{aligned}
\bar{q}_{0,N,i}^{1-\lambda} &= q_{0,N,i}^{1-\lambda}V_{N,i}^{1-\lambda} + \frac{1}{2}(1-\lambda)q_{0,N,i}^{-\lambda}V_{N,i}^{-\lambda}\int_{S_{N,i}} (x-x_{N,i})^T \nabla^2_{0,N,i}(x-x_{N,i})dx + \\
&\quad o\left(V_{N,i}^{2/k+1-\lambda}\right)
\end{aligned}$$

and

$$\bar{q}_{1,N,i}^{\lambda} = q_{1,N,i}^{\lambda}V_{N,i}^{\lambda} + \frac{1}{2}\lambda q_{1,N,i}^{\lambda-1}V_{N,i}^{\lambda-1}\int_{S_{N,i}} (x-x_{N,i})^T \nabla^2_{1,N,i}(x-x_{N,i})dx + o\left(V_{N,i}^{2/k+\lambda}\right).$$

Multiplying the two formulas above yields

$$\begin{aligned}
\bar{q}_{0,N,i}^{1-\lambda}\cdot\bar{q}_{1,N,i}^{\lambda} &= \mu q_{\lambda,N,i}\left(\frac{\lambda}{2q_{1,N,i}}\int_{S_{N,i}} (x-x_{N,i})^T \nabla^2_{1,N,i}(x-x_{N,i})dx\right. \\
&\quad \left. + \frac{1-\lambda}{2q_{0,N,i}}\int_{S_{N,i}} (x-x_{N,i})^T \nabla^2_{0,N,i}(x-x_{N,i})dx + V_{N,i}\right) \\
&\quad + o\left(V_{N,i}^{1+2/k}\right).
\end{aligned} \tag{E.24}$$

Finally, using (E.20), (E.21), and (E.24) we get

$$d_{N,i} = \frac{\mu}{2}\lambda(1-\lambda)q_{\lambda,N,i}\int_{S_{N,i}} (x-x_{N,i})^T F_{N,i}(x-x_{N,i})dx + o\left(V_{N,i}^{1+2/k}\right). \tag{E.25}$$

We shall find the following formulas for $\mu_{N,i}$ and $d_{N,i}$ useful:

$$\mu_{N,i} = \mu q_{\lambda,N,i}V_{N,i} + \frac{\mu}{2}\mathrm{tr}\left(\nabla^2_{\lambda,N,i}M_{N,i}\right)V_{N,i}^{1+2/k} + o\left(V_{N,i}^{1+2/k}\right) \tag{E.26}$$

$$d_{N,i} = \frac{\mu}{2}\lambda(1-\lambda)q_{\lambda,N,i}\mathrm{tr}\left(F_{N,i}M_{N,i}\right)V_{N,i}^{1+2/k} + o\left(V_{N,i}^{1+2/k}\right). \tag{E.27}$$

E.2.3 Asymptotic Values of $\Delta L_{0,N}$ and $\Delta L_{1,N}$

From (E.7) and (E.8) we can write

$$\begin{aligned}\int_{S_{N,i}} q_\lambda(x)\Lambda_0(x)dx &= q_{\lambda,N,i}\Lambda_{0,N,i}V_{N,i} + \frac{1}{2}\Lambda_{0,N,i}\mathrm{tr}(\nabla^2_{\lambda,N,i}M_{N,i})V_{N,i}^{1+2/k} + \\ &\quad \frac{1}{2}q_{\lambda,N,i}\mathrm{tr}\left((F'_{N,i}+G'_{N,i})M_{N,i}\right)V_{N,i}^{1+2/k} + o\left(V_{N,i}^{1+2/k}\right)\end{aligned}$$

where

$$\begin{aligned}F'_{N,i} &= \nabla\Lambda_{0,N,i}\nabla\Lambda_{0,N,i}^T \\ G'_{N,i} &= \frac{\nabla^2_{\lambda,N,i}}{q_{\lambda,N,i}} - \frac{\nabla^2_{0,N,i}}{q_{0,N,i}}.\end{aligned} \tag{E.28}$$

Note that $F'_{N,i}$ can be written in terms of $F_{N,i}$:

$$F'_{N,i} = \lambda^2 F_{N,i}.$$

From (E.19), (E.26), and (E.27) we can write

$$\begin{aligned}\hat{q}_{\lambda,N,i} &= t_N\left(q_{\lambda,N,i}V_{N,i} + \frac{1}{2}\mathrm{tr}(\nabla^2_{\lambda,N,i}M_{N,i})V_{N,i}^{1+2/k} + \frac{1}{2}\lambda(1-\lambda)q_{\lambda,N,i}\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{1+2/k}\right) \\ &\quad + o\left(V_{N,i}^{1+2/k}\right)\end{aligned} \tag{E.29}$$

where

$$t_N = \frac{\mu}{\mu + d_N}.$$

Thus (E.17) becomes

$$\begin{aligned}\Delta L_{0,N} &= \sum_{i=1}^{N} q_{\lambda,N,i}V_{N,i}\left(\Lambda_{0,N,i} - t_N\hat{\Lambda}_{0,N,i}\right) + \\ &\quad \frac{1}{2}\mathrm{tr}(\nabla^2_{\lambda,N,i}M_{N,i})V_{N,i}^{1+2/k}\left(\Lambda_{0,N,i} - t_N\hat{\Lambda}_{0,N,i}\right) + \\ &\quad \frac{1}{2}q_{\lambda,N,i}\mathrm{tr}\left((\lambda^2 F_{N,i} + G'_{N,i})M_{N,i}\right)V_{N,i}^{1+2/k} - \\ &\quad \frac{\lambda(1-\lambda)}{2}t_N q_{\lambda,N,i}\hat{\Lambda}_{0,N,i}\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{1+2/k} + o\left(V_{N,i}^{1+2/k}\right).\end{aligned} \tag{E.30}$$

Next we use the Taylor expansion

$$\log(x+y) = \log x + \frac{y}{x} - \frac{y^2}{2x^2} + o(y^2)$$

to write

$$\hat{\Lambda}_{0,N,i} = \Lambda_{0,N,i} + 2r_{0,N,i} - \frac{1}{2}r_{0,N,i}^2 - \frac{3}{2} + o\left(\left(\frac{\hat{q}_{\lambda,N,i}}{\bar{q}_{0,N,i}} - \frac{q_{\lambda,N,i}}{q_{0,N,i}}\right)^2\right) \tag{E.31}$$

where

$$r_{0,N,i} = \frac{q_{0,N,i}\hat{q}_{\lambda,N,i}}{\bar{q}_{0,N,i}q_{\lambda,N,i}}.$$

To see that the last term in (E.31) is small, note that

$$\frac{\hat{q}_{\lambda,N,i}}{\bar{q}_{0,N,i}} - \frac{q_{\lambda,N,i}}{q_{0,N,i}} = \left(\frac{\bar{q}_{1,N,i}}{\bar{q}_{0,N,i}}\right)^{\lambda}\frac{1}{\mu + d_N} - \left(\frac{q_{1,N,i}}{q_{0,N,i}}\right)^{\lambda}\frac{1}{\mu}.$$

Using the Taylor expansions (E.22), after some algebra this becomes

$$\begin{aligned}
\frac{\hat{q}_{\lambda,N,i}}{\bar{q}_{0,N,i}} - \frac{q_{\lambda,N,i}}{q_{0,N,i}} &= \left(\frac{q_{1,N,i}}{q_{0,N,i}} + o\left(V_{N,i}^{2/k}\right)\right)^{\lambda}\frac{1}{\mu + d_N} - \left(\frac{q_{1,N,i}}{q_{0,N,i}}\right)^{\lambda}\frac{1}{\mu} \\
&= \left[\left(\frac{q_{1,N,i}}{q_{0,N,i}}\right)^{\lambda} + o\left(V_{N,i}^{2/k}\right)\right]\frac{1}{\mu + d_N} - \left(\frac{q_{1,N,i}}{q_{0,N,i}}\right)^{\lambda}\frac{1}{\mu} \\
&= -\left(\frac{q_{1,N,i}}{q_{0,N,i}}\right)^{\lambda}\frac{d_N}{\mu(\mu + d_N)} + o\left(V_{N,i}^{2/k}\right)
\end{aligned}$$

where the second equality follows from (E.23). From (E.27) it is easily seen that

$$o\left(\left(\frac{\hat{q}_{\lambda,N,i}}{\bar{q}_{0,N,i}} - \frac{q_{\lambda,N,i}}{q_{0,N,i}}\right)^2\right) = o\left(V_{N,i}^{2/k}\right).$$

Now, using (E.22) and (E.29), $r_{0,N,i}$ becomes

$$\begin{aligned}
&r_{0,N,i} = \\
&\frac{t_N q_{0,N,i}\left(q_{\lambda,N,i} + \frac{1}{2}\mathrm{tr}(\nabla^2_{\lambda,N,i}M_{N,i})V_{N,i}^{2/k} + \frac{1}{2}\lambda(1-\lambda)q_{\lambda,N,i}\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{2/k}\right) + o\left(V_{N,i}^{2/k}\right)}{q_{\lambda,N,i}q_{0,N,i} + \frac{1}{2}q_{\lambda,N,i}\mathrm{tr}(\nabla^2_{0,N,i}M_{N,i})V_{N,i}^{2/k} + o\left(V_{N,i}^{2/k}\right)} \\
&= t_N\left(1 + \frac{\mathrm{tr}(\nabla^2_{\lambda,N,i}M_{N,i})}{2q_{\lambda,N,i}}V_{N,i}^{2/k} - \frac{\mathrm{tr}(\nabla^2_{0,N,i}M_{N,i})}{2q_{0,N,i}}V_{N,i}^{2/k} + \frac{1}{2}\lambda(1-\lambda)\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{2/k}\right) \\
&\quad + o\left(V_{N,i}^{2/k}\right) \\
&= t_N\left(1 + \frac{1}{2}\mathrm{tr}(G'_{N,i}M_{N,i})V_{N,i}^{2/k} + \frac{1}{2}\lambda(1-\lambda)\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{2/k}\right) + o\left(V_{N,i}^{2/k}\right)
\end{aligned} \tag{E.32}$$

and

$$r_{0,N,i}^2 = t_N^2\left(1 + \mathrm{tr}(G'_{N,i}M_{N,i})V_{N,i}^{2/k} + \lambda(1-\lambda)\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{2/k}\right) + o\left(V_{N,i}^{2/k}\right).$$

Thus (E.31) becomes

$$\begin{aligned}\hat{\Lambda}_{0,N,i} &= \Lambda_{0,N,i} + 2t_N - \frac{1}{2}t_N^2 - \frac{3}{2} + \\ &\left(\mathrm{tr}(G'_{N,i}M_{N,i})V_{N,i}^{2/k} + \lambda(1-\lambda)\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{2/k}\right)\left(t_N - \frac{1}{2}t_N^2\right) + o\left(V_{N,i}^{2/k}\right).\end{aligned}$$

Therefore

$$\begin{aligned}\Lambda_{0,N,i} - t_N\hat{\Lambda}_{0,N,i} &= \Lambda_{0,N,i}(1-t_N) + \frac{3}{2}t_N - 2t_N^2 + \frac{1}{2}t_N^3 - \\ &\left(\mathrm{tr}(G'_{N,i}M_{N,i})V_{N,i}^{2/k} + \lambda(1-\lambda)\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{2/k}\right)\left(t_N^2 - \frac{1}{2}t_N^3\right) \\ &+o\left(V_{N,i}^{2/k}\right). \qquad \text{(E.33)}\end{aligned}$$

Next, using (E.27), we note that

$$\lim_{N\to+\infty} N^{2/k}\frac{d_N}{\mu} = \frac{1}{2}\lambda(1-\lambda)\int \frac{q_\lambda(x)\mathcal{F}(x)}{\zeta(x)^{2/k}}dx$$

and thus

$$\begin{aligned}t_N &= 1 - \frac{d_N}{\mu + d_N} = 1 - \frac{d_N}{\mu} + o\left(\frac{1}{N^{2/k}}\right) \\ t_N^2 &= 1 - \frac{2d_N}{\mu} + o\left(\frac{1}{N^{2/k}}\right) \\ t_N^3 &= 1 - \frac{3d_N}{\mu} + o\left(\frac{1}{N^{2/k}}\right).\end{aligned}$$

Using this in (E.33) gives

$$\begin{aligned}\Lambda_{0,N,i} - t_N\hat{\Lambda}_{0,N,i} &= \Lambda_{0,N,i}\frac{d_N}{\mu} + \frac{d_N}{\mu} - \\ &\frac{1}{2}\left(\mathrm{tr}(G'_{N,i}M_{N,i})V_{N,i}^{2/k} + \lambda(1-\lambda)\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{2/k}\right) + \\ &o\left(V_{N,i}^{2/k}\right) + o\left(\frac{1}{N^{2/k}}\right). \qquad \text{(E.34)}\end{aligned}$$

Next, (E.30) and (E.34) give

$$\begin{aligned}\Delta L_{0,N} &= \sum_{i=1}^{N} q_{\lambda,N,i}\Lambda_{0,N,i}V_{N,i}\frac{d_N}{\mu} - \frac{1}{2}\lambda(1-\lambda)q_{\lambda,N,i}\Lambda_{0,N,i}\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{1+2/k} + \\ &q_{\lambda,N,i}V_{N,i}\frac{d_N}{\mu} - \frac{1}{2}\lambda(1-\lambda)q_{\lambda,N,i}\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{1+2/k} + \\ &\frac{1}{2}\lambda^2 q_{\lambda,N,i}\mathrm{tr}(F_{N,i}M_{N,i})V_{N,i}^{1+2/k} + \\ &o\left(V_{N,i}^{1+2/k}\right) + o\left(\frac{V_{N,i}}{N^{2/k}}\right). \qquad \text{(E.35)}\end{aligned}$$

Finally, by multiplying (E.35) by $N^{2/k}$ and passing to the limit, we obtain (3.33). By symmetry arguments, (3.34) can easily be obtained.

APPENDIX F

Procedure for Generating Vector Quantizers

In this appendix, we describe the procedure and algorithm used to generate the one and two-dimensional quantizers analyzed in Section 3.7. For the estimation-optimal quantizers, we use the well-known LBG algorithm, also known as the generalized Lloyd algorithm [27, 40, 60]. This algorithm produces a quantizer optimal with respect to MSE given the source density and is described below. We then show that a congruent-cell quantizer possessing any arbitrary point density may also be obtained using the generalized Lloyd algorithm.

F.1 Generalized Lloyd (LBG) Algorithm

The generalized Lloyd algorithm utilizes the following two facts regarding MSE-optimal quantizers [27].

1. If an N-cell quantizer is MSE optimal, then it is a nearest-neighbor mapping. That is, if x lies in cell S_i, then $\|x - x_i\| \leq \|x - x_j\|$ for all $j \in \{1, \ldots, N\}$, $j \neq i$.

2. If an N-cell quantizer is MSE optimal, then each of its codebook points lies in the centroid of its cell with respect to the source density. This is equivalent to

$$x_i = \frac{\int_{S_i} xq(x)dx}{\int_{S_i} q(x)dx}$$

for source density $q(x)$.

The steps of the generalized Lloyd algorithm are as follows:

1. Choose an initial codebook. Initialize the distortion to $+\infty$.

2. Compute the cell boundaries such that the resulting quantizer is a nearest-neighbor mapping.

3. Compute a new set of codebook points such that each point lies in the centroid of its cell, with respect to $q(x)$.

4. Compute the distortion (MSE) of the quantizer resulting from step 3.

5. Compute the percent change in distortion. If this percent change falls below a prespecified threshold, stop. Otherwise go to step 2.

It is clear that if the threshold is chosen sufficiently small, the algorithm will produce a quantizer that satisfies the nearest-neighbor property as well as the centroid property. The resulting quantizer will be MSE optimal assuming a local minimum is not reached.

F.2 Generating Vector Quantizers from Point Densities

The estimation-optimal quantizer may be generated using the generalized Lloyd algorithm as described in the previous section. Here we show how to generate a congruent-cell quantizer given an arbitrary point density. To be more specific, we describe the procedure for generating a quantizer whose cells are k-dimensional tessellating polytopes with minimum moment of inertia and whose point density is arbitrarily chosen.

First we note that the estimation-optimal point density corresponding to source density $q(x)$, as given by (3.19), is $\zeta^e(x) = T(q)$ where

$$T(q) = \frac{q(x)^{\frac{k}{k+2}}}{\int q(y)^{\frac{k}{k+2}} dy}. \tag{F.1}$$

Inverting (F.1), it is easy to see that the source density is given by $q(x) = T^{-1}(\zeta^e)$ where

$$T^{-1}(\zeta) = \frac{\zeta(x)^{\frac{k+2}{k}}}{\int \zeta(y)^{\frac{k+2}{k}} dy}. \tag{F.2}$$

Next we note that since the quantizer generated by the generalized Lloyd algorithm is known to be estimation optimal (with respect to source density $q(x)$), we can conclude that its point density is the estimation-optimal point density $T(q)$. Now suppose the generalized Lloyd algorithm is given the source density $T^{-1}(\tilde{\zeta})$ where $\tilde{\zeta}$ is any point density. The VQ that is generated by the algorithm will clearly have point density $T\left(T^{-1}(\tilde{\zeta})\right) = \tilde{\zeta}$. Furthermore, for large N, the cells of the resulting quantizer will be k-dimensional polytopes with minimum moment of inertia. This follows from Gersho's conjecture and the MSE-optimality of the generalized Lloyd algorithm. Therefore, the congruent-cell quantizer with point density $\tilde{\zeta}$ may be obtained by feeding the generalized Lloyd algorithm with the source density $T^{-1}(\tilde{\zeta})$.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] R. Ahlswede and I. Csiszár, "Hypothesis Testing with Communication Constraints," *IEEE Trans. on Inform. Theory*, vol. 32, no. 4, pp. 533–542, July 1986.

[2] S. T. Alexander, "Transient Weight Misadjustment Properties for the Finite Precision LMS Algorithm," *IEEE Trans. Acoust., Speech, and Sig. Proc.*, vol. ASSP-35, no. 9, pp. 1250–1258, Sep. 1987.

[3] H. H. Barrett, C. K. Abbey, and E. Clarkson, "Objective Assessment of Image Quality. III. ROC Metrics, Ideal Observers, and Likelihood-Generating Functions," *J. Opt. Soc. Am.*, vol. 15, no. 6, pp. 1520–1535, June 1998.

[4] H. H. Barrett, J. L. Denny, R. F. Wagner, and K. J. Meyers, "Objective Assessment of Image Quality. II. Fisher Information, Fourier Crosstalk, and Figures of Merit for Task Performance," *J. Opt. Soc. Am.*, vol. 12, no. 5, pp. 834–852, May 1995.

[5] G. R. Benitz and J. A. Bucklew, "Asymptotically Optimal Quantizers for Detection of I.I.D. Data," *IEEE Trans. on Inform. Theory*, vol. 35, no. 2, pp. 316–325, Mar. 1989.

[6] J. C. M. Bermudez and N. J. Bershad, "A Nonlinear Analytical Model for the Quantized LMS Algorithm – the Arbitrary Step Size Case," *IEEE Trans. on Signal Processing*, vol. SP-44, no. 5, pp. 1175–1183, May 1996.

[7] N. J. Bershad and J. C. M. Bermudez, "A Nonlinear Analytical Model for the Quantized LMS Algorithm – the Power-of-Two Step Size Case," *IEEE Trans. on Signal Processing*, vol. SP-44, no. 11, pp. 2895–2900, Nov. 1996.

[8] N. J. Bershad and J. C. M. Bermudez, "New Insights on the Transient and Steady-State Behavior of the Quantized LMS Algorithm," *IEEE Trans. on Signal Processing*, vol. SP-44, no. 10, pp. 2623–2625, Oct. 1996.

[9] T. G. Birdsall, *The Theory of Signal Detectability: ROC Curves and Their Character*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 1966.

[10] R. Blahut, *Principles and Practice of Information Theory*, Addison-Wesley, 1987.

[11] T. A. Brubaker and J. C. Becker, "Multiplication using Logarithms implemented with Read Only Memory," *IEEE Trans. on Computers*, vol. C-24, pp. 761–765, 1975.

[12] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*, Wiley, New York, 1990.

[13] J. A. Bucklew, "Two Results on the Asymptotic Performance of Quantizers," *IEEE Trans. on Inform. Theory*, vol. IT-30, pp. 341–348, March 1984.

[14] J. A. Bucklew and G. L. Wise, "Multidimensional Asymptotic Quantization Theory with rth Power Distortion Measures," *IEEE Trans. on Inform. Theory*, vol. IT-28, pp. 239–247, March 1982.

[15] C. Caraiscos and B. Liu, "A Roundoff Error Analysis of the LMS Adaptive Algorithm," *IEEE Trans. Acoust., Speech, and Sig. Proc.*, vol. ASSP-32, no. 1, pp. 34–41, Feb. 1984.

[16] A. P. Chandrakasan and R. W. Brodersen, "Minimizing Power Consumption in Digital CMOS Circuits," *IEEE Proceedings*, vol. 83, pp. 498–523, April 1995.

[17] Y.-H. Chang, C.-K. Tzou, and N. J. Bershad, "Postsmoothing for the LMS Algorithm and a Fixed Point Roundoff Error Analysis," *IEEE Trans. on Signal Processing*, vol. 39, no. 4, pp. 959–961, April 1991.

[18] C. Chien, *Digital Radio Systems on a Chip: A Systems Approach*, Kluwer, Norwell, MA, 2001.

[19] J. M. Cioffi and M. Ho, "A Finite Precision Analysis of the Block-Gradient Adaptive Data-Driven Echo Canceller," *IEEE Trans. on Communications*, vol. 40, no. 5, pp. 940–946, May 1992.

[20] J. M. Cioffi and J. J. Werner, "The Tap-Drifting Problem in Digitally Implemented Data-Driven Echo Cancellers," *AT&T Bell Lab. Tech. J.*, vol. 64, no. 1, pp. 115–138, 1985.

[21] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, New York, 1987.

[22] S. Douglas and T.-Y. Meng, "Optimum Error Quantization for LMS Adaptation," in *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, May 1991.

[23] E. Eweda, N. R. Yousef, and S. H. El-Ramly, "Effect of Finite Wordlength on the Performance of an Adaptive Filter Equipped with the Signed Regressor Algorithm," in *Proc. of IEEE Global Telecommunications Conf.*, pp. 1325–1329, London, UK, 1996.

[24] T. J. Flynn and R. M. Gray, "Encoding of Correlated Observations," *IEEE Trans. on Inform. Theory*, vol. 33, no. 6, pp. 773–787, Nov. 1987.

[25] W. F. Gabriel, "Adaptive Arrays – an Introduction," *IEEE Proceedings*, vol. 64, pp. 239–272, Feb. 1976.

[26] A. Gersho, "Asymptotically Optimal Block Quantization," *IEEE Trans. on Inform. Theory*, vol. IT-25, pp. 373–380, July 1979.

[27] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer, Boston MA, 1992.

[28] R. D. Gitlin, J. E. Mazo, and M. G. Taylor, "On the Design of Gradient Algorithms for Digitally Implemented Adaptive Filters," *IEEE Trans. Circuit Theory*, vol. CT-20, pp. 125–136, Mar. 1973.

[29] R. M. Gray, "Quantization, Classification, and Density Estimation," in *Proc. of IEEE Information Theory Workshop on Detection, Estimation, Classification, and Imaging*, Santa Fe, NM, Feb. 1999.

[30] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. on Inform. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[31] R. Gupta and A. O. Hero, "Transient Behavior of Fixed-Point LMS Adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, Istanbul, Turkey, June 2000.

[32] R. Gupta and A. O. Hero, "Power versus Performance Tradeoffs for Reduced Resolution LMS Adaptive Filters," *IEEE Trans. on Signal Processing*, vol. 48, no. 10, pp. 2772–2784, Oct. 2000.

[33] T. S. Han, "Hypothesis Testing with Multiterminal Data Compression," *IEEE Trans. on Inform. Theory*, vol. 33, no. 6, pp. 759–772, Nov. 1987.

[34] T. S. Han and S. Amari, "Statistical Inference Under Multiterminal Data Compression," *IEEE Trans. on Inform. Theory*, vol. 44, no. 6, pp. 2300–2324, Oct. 1998.

[35] A. O. Hero, *Statistical Methods for Signal Processing*, Coursepack, University of Michigan, Ann Arbor, 2000.

[36] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge, 1985.

[37] S. A. Kassam, "Optimum Quantization for Signal Detection," *IEEE Trans. on Communications*, vol. COM-25, pp. 479–484, May 1977.

[38] R. H. Katz, *Contemporary Logic Design*, Benjamin-Cummings, 1994.

[39] E. L. Lehmann, *Testing Statistical Hypotheses*, Wiley, New York, 1959.

[40] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. on Communications*, vol. COM-28, pp. 84–95, Jan. 1980.

[41] M. Longo, T. D. Lookabaugh, and R. M. Gray, "Quantization for Decentralized Hypothesis Testing under Communication Constraints," *IEEE Trans. on Inform. Theory*, vol. 36, pp. 241–255, March 1990.

[42] T. D. Lookabaugh and R. M. Gray, "High-Resolution Quantization Theory and the Vector Quantizer Advantage," *IEEE Trans. on Inform. Theory*, vol. 35, no. 5, pp. 1020–1033, Sep. 1989.

[43] D. G. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[44] J. McChesney, "Notes for MURI Seminar on Low Energy Electronics Design for Mobile Platforms," University of Michigan, Ann Arbor, Oct. 1997.

[45] P. W. Moo, *Asymptotic Analysis of Lattice-Based Quantization*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 1998.

[46] S. Na and D. L. Neuhoff, "Bennet's Integral for Vector Quantizers," *IEEE Trans. on Inform. Theory*, vol. 41, no. 4, pp. 886–900, July 1995.

[47] D. L. Neuhoff, Department of Electrical Engineering and Computer Science, University of Michigan, personal communication, 2001.

[48] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Phil. Trans. Roy. Soc. Ser. A.*, pp. 289–337, 1933.

[49] C. J. Nicol, P. Larsson, K. Azadet, and J. H. O'Neill, "A Low-Power 128-Tap Digital Adaptive Equalizer for Broadband Modems," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 11, pp. 1777–1789, Nov. 1997.

[50] K. Oehler and R. M. Gray, "Combining Image Compression and Classification using Vector Quantization," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 17, no. 5, pp. 461–473, May 1995.

[51] K. K. Parhi, "Algorithm Transformation Techniques for Concurrent Processors," *IEEE Proceedings*, vol. 77, pp. 1879–1895, Dec. 1989.

[52] A. Peled and B. Liu, *Digital Signal Processing: Theory, Design, and Implementation*, Wiley, New York, 1976.

[53] K. O. Perlmutter, S. M. Perlmutter, *et al.*, "Bayes Risk Weighted Vector Quantization with Posterior Estimation for Image Compression and Classification," *IEEE Trans. on Image Processing*, vol. 5, no. 2, pp. 347–360, Feb. 1996.

[54] B. Picinbono and P. Duvaut, "Optimum Quantization for Detection," *IEEE Trans. on Communications*, vol. 36, no. 11, pp. 1254–1258, Nov. 1988.

[55] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag, New York, 1988.

[56] H. V. Poor, "A Companding Approximation for the Statistical Divergence of Quantized Data," in *Proc. of IEEE Conf. Decision and Control*, San Antonio, TX, Dec. 1983.

[57] H. V. Poor, "Fine Quantization in Signal Detection and Estimation – Part 1," *IEEE Trans. on Inform. Theory*, vol. 34, pp. 960–972, Sep. 1988.

[58] H. V. Poor and J. B. Thomas, "Applications of Ali-Silvey Distance Measures in the Design of Generalized Quantizers," *IEEE Trans. on Communications*, vol. COM-25, pp. 893–900, Sep. 1977.

[59] J. G. Proakis, *Digital Communications –3rd. ed.*, McGraw Hill, Boston, MA, 1995.

[60] K. Sayood, *Introduction to Data Compression*, Morgan-Kaufman, 1996.

[61] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. Journ.*, vol. 27, pp. 379–423, 623–656, 1948.

[62] H. Shimokawa, T. S. Han, and S. Amari, "Error Bound of Hypothesis Testing with Data Compression," in *Proc. of IEEE Symposium on Information Theory*, 1994.

[63] D. Singh, J. M. Rabaey, M. Pedram, F. Catthoor, S. Rajgopal, N. Sehgal, and T. J. Mozden, "Power Conscious CAD Tools and Methodologies: a Perspective," *IEEE Proceedings*, vol. 83, pp. 5701–594, April 1995.

[64] D. Slepian and J. K. Wolf, "Noiseless Coding of Correlated Information Sources," *IEEE Trans. on Inform. Theory*, vol. 19, pp. 471–480, July 1973.

[65] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[66] J. Stewart, *Calculus: Early Transcendentals*, Brooks-Cole, Pacific Grove, CA, 1987.

[67] J. R. Treichler, C. R. Johnson, and M. G. Larimore, *Theory and Design of Adaptive Filters*, Wiley, New York, 1987.

[68] J. N. Tsitsiklis, "Extremal Properties of Likelihood Ratio Quantizers," *IEEE Trans. on Communications*, vol. 41, no. 4, pp. 550–558, Apr. 1993.

[69] H. L. Van-Trees, *Detection, Estimation, and Modulation Theory: Part I*, Wiley, New York, 1968.

[70] A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication*, Addison-Wesley, Reading, MA, 1995.

[71] B. Widrow and M. Hoff, "Adaptive Switching Circuits," *IRE Wescon Convention Record Part IV*, pp. 96–104, 1960.

[72] B. Widrow, P. E. Mantey, *et al.*, "Adaptive Antenna Systems," *IEEE Proceedings*, vol. 55, no. 12, pp. 2143–2159, Dec. 1967.

[73] B. Widrow, J. McCool, and M. Ball, "Complex LMS Algorithm," *IEEE Proceedings*, vol. 63, pp. 719–720, 1975.

[74] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice Hall, Englewood Cliffs NJ, 1985.

[75] P. Zador, *Development and Evaluation of Procedures for Quantizing Multivariate Distributions*, Ph.D. thesis, Stanford University, Stanford, CA, 1963.

[76] P. Zador, "Asymptotic Quantization Error of Continuous Signals and Quantization Dimension," *IEEE Trans. on Inform. Theory*, vol. IT-28, pp. 139–149, 1982.

[77] R. Zamir and M. Feder, "On Lattice Quantization Noise," *IEEE Trans. on Inform. Theory*, vol. 42, no. 4, pp. 1152–1159, July 1996.