UNIVERSITY OF CALIFORNIA, SAN DIEGO

Channel Estimation and Feedback for Multiple Antenna Communication

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in
Electrical and Computer Engineering
(Communication Theory and Systems)

by

Chandra Ramabhadra Murthy

Committee in charge:

Professor Bhaskar D. Rao, Chair
Professor Philip E. Gill
Professor Alon Orlitsky
Professor Paul H. Siegel
Professor Kenneth Zeger

2006

The dissertation of Chandra Ramabhadra Murthy is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2006

To my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am grateful to my advisor Professor Bhaskar D. Rao for his continual support and encouragement throughout the duration of my PhD study. His interest in research was infectious and motivated much of this dissertation. It has been a pleasure working under his supervision. I would also like to thank my committee members, Professors Philip E. Gill, Alon Orlitsky, Paul H. Siegel and Kenneth Zeger for their time and helpful comments.

I would also like to thank the UCSD CoRe research granting agency and the affiliated companies for supporting me throughout my PhD program. This work was supported by CoRe Research Grant Com02-10105.

The text of Chapter 2, in part, is a reprint of the material as it appears in C. R. Murthy, A. K. Jagannatham, and B. D. Rao, "Training-only and semi-blind channel estimation for maximum ratio transmission based MIMO systems," *IEEE Transactions on Sig. Proc.*, vol. 54, pp. 2546–2558, July 2006. Chapter 3, in part, is a reprint of a paper which has been accepted for publication in the *IEEE Transactions on Signal Processing* as C. R. Murthy and B. D. Rao, "Quantization methods for equal gain transmission with finite rate feedback". The text of Chapter 4, in part, has appeared in C. R. Murthy and B. D. Rao, "High-Rate Analysis of Source Coding for Symmetric Error Channels", *Data Compression Conference (DCC)*, Snowbird, UT, Mar. 2006, and C. R. Murthy, E. R. Duni and B. D. Rao, "High-rate analysis of vector quantization for noisy channels", *Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, Tolouse, France, May 2006. Chapter 5, in part, is a reprint of the material which has appeared as C. R. Murthy and B. D. Rao, "Effect of feedback errors on quantized equal gain transmission", *Int. Conf. on Communications (ICC)*, Istanbul, Turkey, Jun. 2006. The dissertation author was the primary researcher and author, and the co-authors listed in these publications contributed to or supervised the research which forms the basis for this dissertation.

My lab-mates and co-authors Ethan Duni, Adityakiran Jagannatham,

and Jun Zheng deserve special mention for the invaluable technical exchange. I am grateful to Zhongren Cao, Yogananda Isukapalli, June Chul Roh, Cécile Levasseur, Joseph Murray, Shankar Shivappa, Anand Subramaniam, Thomas Svantesson, Yeliz Tokgoz, Chengjin Zhang and Wenyi Zhang from the DSP lab for the many hours of discussions, both technical and non-technical. In particular, I thank Cécile Levasseur for being so much more than a friend and lab-mate during the past few years. Also, I dearly miss the lively, scintillating discussions and the amazing sense of humor of Anand Subrahmaniam: may peace be with him.

The apartment at Mesa residential has been a comfortable home for me during my stay in San Diego. I was not allowed to keep a pet, so I had to keep a room-mate instead. I would like to thank my room-mates Anand Balachandran, Rajiv Bharadwaja and Christopher Ellison for being wonderful companions. I would also like to thank long-distance friends Sunil Gopinath and Anand Ilango for supporting me through seemingly interminable phone calls. My surf-buddies Patrick Amihood, Luke Barrington, Shay Har-Noy and David Wipf made life outside work seem almost more fun than doing research. Shankar Shivappa has been a great sailing partner, I thank him for the time spent exploring Mission Bay.

Without the countless lunch-breaks with Azadeh Bozorgzadeh, Elizabeth Gire, Jittra Jootar, Athanasios Leontaris, Periklis Liaskovitis, Maziar Nezhad, Yiannis Spyropoulos and Kostas Stamatiou my life here would have been significantly duller, hence I am grateful to them. Other people I have interacted regularly outside work include Ofer Achler, Ramesh Annavajjala, Supratik Bhattacharjee, Mohamed Jalloh, Ashok Mantravadi, Jens Mühle, Kiran Mukkavilli, Shiauhe Tsai, Patrick Verkaik and Helen Yapura, and these people contributed in no small way by being good friends and counselors.

My local family and friends Deepa and Keshava Datta, Uma and Balakrishna Rao, Anu and Sudarshan Keshava deserve thanks for numerous meals that fueled my work. Acting in Kannada dramas with Aparna and Sundaram Nagaraj, Vijay Balakrishna, Rashmi and Madhukara was a memorable experience.

Finally, I owe my deepest gratitude to my family. I would like to thank my parents for their love. They have always supported me in my decisions, and encouraged me to pursue my dreams. My sister Rashmi is blossoming into a world class doctor from being my childhood fighting partner. My late grand-father Ramabhadra Shastry continues to be a source of strength and discipline. My grandmother Subbalakshmi Shastry is the embodiment of unconditional love. My maternal grand-parents H.C. Subba Rao and Lalitha have taught me the value of simple living and dedication to work. This thesis is dedicated to all of them.

VITA

1976	Born, Bangalore, INDIA
1998	B.S., Electrical Engineering Indian Institute of Technology Madras
1998–2000	Research Assistant Department of Electrical and Computer Engineering Purdue University
2000	M.S., Electrical and Computer Engineering Purdue University
2000–2002	Systems Engineer Qualcomm, Inc.
2002–2006	Research Assistant Department of Electrical and Computer Engineering University of California, San Diego
2006	Ph.D., Electrical and Computer Engineering University of California, San Diego

PUBLICATIONS

C. R. Murthy and B. D. Rao, "High-Rate Analysis of Source Coding for Symmetric Error Channels", *Data Compression Conference (DCC)*, Snowbird, UT, Mar. 2006.

C. R. Murthy and B. D. Rao, "Effect of feedback errors on quantized equal gain transmission", *Int. Conf. on Communications (ICC)*, Istanbul, Turkey, Jun. 2006.

C. R. Murthy, E. R. Duni and B. D. Rao, "High-rate analysis of vector quantization for noisy channels", *Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, Tolouse, France, May 2006

C. R. Murthy and B. D. Rao, "Quantization methods for equal gain transmission with finite rate feedback", *IEEE Transactions on Sig. Proc., accepted for publication*, 2006.

C. R. Murthy and B. D. Rao,"A vector quantization based approach for equal gain transmission", *Proc. IEEE Global Telecommunications Conference (Globecom)*, St. Louis, MO, Nov. 2005.

C. R. Murthy, A. K. Jagannatham, and B. D. Rao, "Training-only and semi-blind channel estimation for maximum ratio transmission based MIMO systems," *IEEE Transactions on Sig. Proc.*, vol. 54, pp. 2546–2558, July 2006.

C. R. Murthy, J. Zheng and B. D. Rao, "Multiple Antenna Systems with Finite Rate Feedback", in *Proc. MILCOM*, Atlantic City, NJ, Oct. 2005.

C. R. Murthy and B. D. Rao, "On antenna selection with maximum ratio transmission," *Conf. Record of the 37th Asilomar Conf. on Signals, Systems and Computers*, Nov. 2003, vol. 1, pp. 228 – 232.

C. R. Murthy, J. C. Roh, and B. D. Rao, "Optimality of extended maximum ratio transmission," *6th Baiona Workshop on Signal Processing in Communications*, Baiona, Spain, Sept. 2003, pp. 47–50.

ABSTRACT OF THE DISSERTATION

Channel Estimation and Feedback for Multiple Antenna Communication

by

Chandra Ramabhadra Murthy

Doctor of Philosophy in Electrical and Computer Engineering
(Communication Theory and Systems)

University of California, San Diego, 2006

Professor Bhaskar D. Rao, Chair

This dissertation studies several aspects of feedback-based communication with multiple antennas, such as the estimation of the Channel State Information (CSI), the quantization of the CSI with a finite number of bits to enable its feedback to the transmitter, as well as the effect of errors in the feedback channel on the performance of the communication system.

Channel estimation is doubly important in feedback-based communication because inaccurate CSI affects not only the receiver performance, but also results in sub-optimal transmission. In this context, Multiple Input Multiple Output (MIMO) flat-fading channel estimation when the transmitter employs Maximum Ratio Transmission (MRT) is studied. Two competing schemes for estimating the transmit and receive beamforming vectors of the channel matrix are analyzed: a training based conventional least squares estimation (CLSE) scheme and a closed-form semi-blind (CFSB) scheme that employs training followed by information-bearing spectrally white data symbols. Employing matrix perturbation theory, expressions for the mean squared error (MSE) in the beamforming vector, the average received SNR and the symbol error rate (SER) performance of both the semi-blind and the conventional schemes are derived.

Another important issue in beamforming-based communication with multiple antennas is the feedback of CSI. Hence, the design and analysis of quantizers for Equal Gain Transmission (EGT) systems with finite rate feedback-based communication in flat-fading Multiple Input Single Output (MISO) systems is considered. Two popular approaches for quantizing the phase angles are contrasted: vector quantization (VQ) and scalar quantization (SQ). Closed-form expressions are derived for the performance of quantized feedback in terms of capacity loss and outage probability in the case of i.i.d. Rayleigh flat-fading channels.

In the work described above, the feedback channel is assumed to be free of delay and noise. With the view to understand the effect of errors on quantization, this dissertation considers the more general problem of characterizing the high-rate performance of source coding for noisy discrete symmetric channels with random index assignment. Theoretical expressions for the performance of source coding for noisy channels are derived for a large class of distortion measures. The theoretical expressions are used to derive new results for two specific applications. The first is the quantization of the CSI for MISO systems with beamforming at the transmitter. The second application is in the wideband speech compression problem, i.e., that of quantizing the linear predictive coding parameters in speech coding systems with the log spectral distortion as performance metric.

1 Introduction

In the past decade, Multiple-Input, Multiple-Output (MIMO) systems have enjoyed a renewed interest both in academics (starting from the seminal works of Winters [1], Foschini [2], Telatar [3] and others) and in the industry (wireless communication standards such as 802.11 and 802.16 for local area networks, and CDMA 2000 (3GPP) and WCDMA (3GPP2) for cellular telephony). Although antenna array communication has been known since the 1930s, the renewed interest in the past decade or so has primarily to do with the dramatic increase in center-frequency of signal transmission. Physically, it is known that when antenna elements are placed about 10 wavelengths apart, the channel gains, or fade-values, from a single common source in a rich-scattering environment are uncorrelated. Therefore, in classical radio frequency (RF) communication, the antenna separation needed to attain decorrelation of the channel gains between different antennas would be of the order of several meters. However, current wireless communication standards operate in the Giga-Hertz range, which allows the channel gains to be decorrelated with an antenna separation of just a few centimeters, making it possible to fit multiple antennas even on small hand-held devices. In addition, it was recognized that having independent fade-values between antennas is a means to achieve diversity or multiplexing benefits. For example, if one of the transmit-receive antenna pairs is in a deep fade, perhaps another pair has a good channel condition, thus enabling a significantly more reliable communication. Also, if there are several independent channels from the transmitter to the receiver, it allows for the possibility of sending independent data on the different

paths or beams, thereby increasing the total data rate. Finally, the availability of cheaper hardware has made it feasible to implement complicated DSP algorithms economically. Due to these reasons, there has been an explosion of research and development in MIMO systems.

1.1 Preliminaries

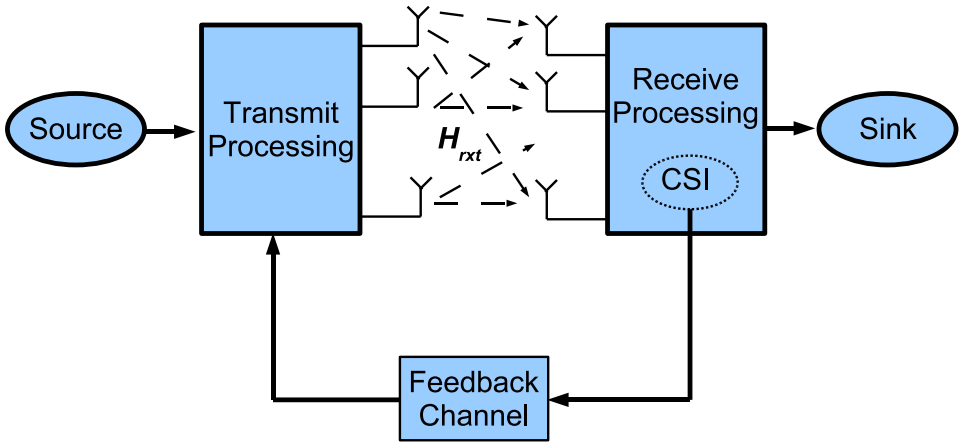


Figure 1.1: A simple MIMO system model.

A simple model of a point-to-point, narrowband MIMO wireless communication system with t transmit antennas and r receive antennas is shown in Fig. 1.1. Under the block flat-fading model, the multiple-antenna channel is represented by the channel matrix $H \in \mathbb{C}^{r \times t}$ which remains constant for the duration of a block, and changes independently according to some statistical distribution from block to block. For simplicity of notation, therefore, the time index can be omitted and the relationship between the channel input $\mathbf{x} \in \mathbb{C}^t$ and the channel output $\mathbf{y} \in \mathbb{C}^r$ can be expressed as

$$\mathbf{y} = H\mathbf{x} + \mathbf{n} \tag{1.1}$$

where $\mathbf{n} \in \mathbb{C}^r$ is the Additive White Gaussian Noise (AWGN) at the receiver. $\mathbf{x}$ is

typically obtained by preprocessing the source symbol to prepare it for transmission across the channel with maximum reliability and data rate.

1.1.1 Role of the Availability of Channel State Information

It is known that the capacity of the above MIMO wireless link depends on the availability of accurate Channel State Information (CSI) at the transmitter and the receiver. More specifically, the capacity (and how to achieve the capacity) of a MIMO system with with perfect CSI at the receiver and no CSI at the transmitter, as well as with perfect CSI at both transmitter and receiver are well known [3].

Several problems remain to be solved in order to achieve the higher capacity promised by MIMO systems, some of which are addressed in this thesis. First, the channel needs to be estimated at the receiver. Then, for systems that require the CSI at the transmitter, the CSI needs to be quantized with a finite number of bits, since typically there is a bandwidth constraint on the feedback link. There may be a delay in the feedback link, in which case, the transmitter would have an outdated copy of the CSI. Also, there may be errors in the feedback link. For wideband communication, the channel becomes a tapped-delay line filter (in the discrete-time representation), which further complicates the problem. With multi-user communication, new problems associated with multi-user diversity and optimum signalling schemes arise. In addition, there are a multitude of issues such as synchronization, multipath, jitter and so on, which need to be addressed in any practical system. Recently, several papers have appeared quantifying the effects of the partial CSI at the transmitter coupled with these impairments.

Channel feedback is also under consideration in 3rd generational mobile and wireless LAN standards, for example in the closed-loop mode specification in 3GPP High Speed Downlink Packet Access (HSDPA) [4] and in the eigenbeamforming mode specification in IEEE 802.11 [5] and IEEE 802.16 [6].

Transmit-Receive Beamforming

Beamforming is an attractive technique for data transmission and reception when using multiple antennas, wherein the transmitted vector $\mathbf{x}$ is given by $\mathbf{x} = \mathbf{v}s$, where $\mathbf{v} \in \mathbb{C}^t$ is a *beamforming vector* and s is the data symbol to be transmitted. Typically, to ensure that the data power is not amplified, there exists a 2-norm constraint on $\mathbf{v}$. At the receiver, a receive beamforming vector $\mathbf{z} \in \mathbb{C}^r$ is used to compute $\mathbf{z}^H\mathbf{y}$, from which the transmitted symbol s is recovered. Beamforming is the optimum method (in terms of maximizing the capacity) of transmission when the transmitter has perfect CSI and there is only one receive antenna, or at low transmit powers with multiple receive antennas.

Two popular beamforming methods are Maximum Ratio Transmission (MRT) [7] and Equal Gain Transmission (EGT) [8]. When beamforming is employed at the transmitter, MRT maximizes the channel capacity when a constraint is imposed on the total power from all the transmit antennas. In general, an MRT beamforming vector is denoted $\mathbf{v} \in \mathbb{C}^t$, where t is the number of transmit antennas, and the constant total power constraint can be expressed as $\|\mathbf{v}\|_2^2 = t$, where $\|\mathbf{v}\|_2$ denotes Euclidean norm (or L_2-norm) of $\mathbf{v}$. It can be shown that the MRT beamforming vector contains t complex parameters and two real constraints, i.e, it can be completely described by $(t-1)$ complex parameters, which need to be made available at the transmitter to enable optimum MRT.

EGT (see, e.g., [8] and the references therein), on the other hand, is the optimum beamforming vector for maximizing the capacity of beamforming-based flat fading systems with an equal power per-antenna constraint. A per-antenna power constraint, rather than a total power constraint, is more practically meaningful in the design of transmit beamforming vectors multiple antenna systems as they impose much less fidelity requirements on the the transmit RF power amplifiers. In general, an EGT beamforming vector is given by $\mathbf{w} = [1, \exp(j\theta_2), \exp(j\theta_3), \ldots, \exp(j\theta_t)]^T$, where θ_i denotes the phase rotation applied at antenna element i. Thus, the EGT vector contains exactly $(t-1)$ real

parameters that need to be made available at the transmitter to enable optimum EGT, which is half the number of parameters needed to enable optimum MRT.

1.1.2 Channel Estimation for Feedback-based Communication

In order for feedback-based beamforming schemes such as MRT or EGT to work well, it is necessary for the receiver to first have an accurate estimate of the channel matrix H. One standard technique to estimate the channel is to transmit a sequence of training symbols (also called pilot symbols) at the beginning of each frame. This training symbol sequence is known at the receiver, and thus the channel is estimated from the measured outputs to training symbols. Training based schemes usually have very low complexity making them ideally suited for implementation in systems (e.g., mobile stations) where the available computational capacity is limited. However, such training based schemes are transmission scheme agnostic. Semi-blind techniques, on the other hand, can be tailored to enhance the accuracy of channel estimation by efficiently utilizing not only the known training symbols but also the unknown data symbols to specifically estimate the parameters of interest. Hence, they can be used to reduce the amount of training data required to achieve the desired system performance, or equivalently, achieve better accuracy of estimation for a given number of training symbols, thereby improving the spectral efficiency and channel throughput. Work on semi-blind techniques for the design of fractional semi-blind equalizers in multi-path channels has been reported earlier by Pal in [9, 10], and in [11, 12], error bounds and asymptotic properties of blind and semi-blind techniques are analyzed.

1.1.3 Channel State Information Feedback Models

Two of the popular models for studying the effect of partial CSI at the transmitter are statistical feedback and instantaneous feedback. In the statistical feedback approach, it is assumed that the channel coherence time is too small to feedback every channel instantiation. However, the channel statistics vary suffi-

ciently slowly, so that the mean and/or the covariance of the channel can be made available to the transmitter accurately. The channel is then modeled as Gaussian distributed with the given mean and covariance, and the system performance is optimized with respect to the input distribution and analytically characterized. Examples of works that employ statistical feedback include [13] - [19].

In the instantaneous feedback approach, which is the feedback model employed in this thesis, the receiver attempts to convey to the transmitter the current channel instantiation, typically through a bandwidth-constrained feedback link. That is, given B bits of feedback, the receiver maps the current channel instantiation H to one of $N = 2^B$ integer indices, with each index corresponding to a particular mode of the channel. The transmitter has knowledge of the N-mode codebook, and therefore, it is able to optimize its transmission strategy based on the feedback information. Thus, it is a challenging problem to design optimal quantization schemes and the associated transmission strategies for multiple-antenna systems with finite rate feedback. The design and analysis of the optimum quantizer that takes advantage of both the underlying channel distribution as well as the performance metric (received SNR, outage probability, mutual information rate, bit error rate, etc) has received much attention in the past few years, notably in [20] - [33]. An overview of recent work in this area is provided in [34].

1.1.4 Channel Quantization for Feedback

Another aspect of instantaneous feedback with MRT and EGT is that in practical communication systems, due to feedback channel bit rate constraints, the beamforming vector has to be quantized with a finite number of bits, and only the quantized version can be made available to the transmitter. The problem of quantizing the MRT beamforming vector in a vector quantization (VQ) framework has been considered in [35]. Several results quantifying the performance of finite-rate feedback systems with MRT have also appeared in [20] - [36]. Other works on the performance of quantized-feedback based multiple antenna systems include

[32] and [33], where the authors use quantized CSI obtained from a feedback link to determine a weighting matrix or precoding matrix to improve the performance of an orthogonal space-time block (OSTB) code. The problem of quantizing the EGT beamforming vector was proposed in [37]. In [8], the authors considered the problem of designing quantized EGT beamforming vectors for the case of i.i.d. Rayleigh fading channels and proposed a design criterion based on Grassmannian beamforming. Another recent work that considers per-antenna power constrained transmission is [38], where the authors derive an random search based algorithm to design equal gain codebooks. In [39], the authors employ the minimum value of the maximum magnitude of the inner product between any two code vectors as the performance metric, and derive several nontrivial families of codebooks for which, imposing the per-antenna power constraint rather than the total power constraint results in no performance loss. Finally, one simple method of reducing the feedback overhead and the hardware cost is antenna selection, discussed in [40], where the idea is to have more antenna elements than the transmit and receive chains, and select a subset of antennas that yields the best performance based on the current channel instantiation.

1.2 Outline of the Thesis

Chapter 2 of this dissertation is a comparative study of training-based and semi-blind MIMO flat-fading channel estimation schemes when the transmitter employs Maximum Ratio Transmission (MRT). Two competing schemes for estimating the transmit and receive beamforming vectors of the channel matrix are presented: a training based conventional least squares estimation (CLSE) scheme and a closed-form semi-blind (CFSB) scheme that employs training followed by information-bearing spectrally white data symbols. Employing matrix perturbation theory, expressions for the mean squared error (MSE) in the beamforming vector, the average received SNR and the symbol error rate (SER) performance

of both the semi-blind and the conventional schemes are derived. A weighted linear combiner of the CFSB and CLSE estimates for additional improvement in performance is also proposed.

In the third chapter, the design and analysis of quantizers for EGT systems with finite rate feedback-based communication in flat-fading Multiple-Input, Single-Output (MISO) systems is considered. Two popular approaches for quantizing the phase angles are contrasted: vector quantization (VQ) and scalar quantization (SQ). On the VQ side, using the capacity loss with respect to EGT with perfect CSI at transmitter as performance metric, a criterion for designing the beamforming codebook for quantized EGT (Q-EGT) is developed. An iterative algorithm based on the well-known generalized Lloyd algorithm is proposed for computing the beamforming vector codebook. On the analytical side, closed-form expressions are derived for the performance of quantized feedback in terms of capacity loss and outage probability in the case of i.i.d. Rayleigh flat-fading channels.

In the preceding work, the feedback channel is assumed to be free of delay and noise. Errors in the feedback channel can adversely affect the performance of a quantized-feedback based transmission scheme. With the view to understand the effect of errors on channel quantization, the fourth chapter considers the more general problem of characterizing the high-rate performance of source coding for noisy discrete symmetric channels with random index assignment. Theoretical expressions for the performance of source coding are derived for a large class of distortion measures. It is shown that when the point density is continuous, the high-rate distortion can be approximately expressed as the sum of the source quantization distortion and the channel-error induced distortion, result known previously only for the case of the mean-squared error distortion. Optimization of the point density is also considered. For general distortion functions, assuming that the point density is continuous, expressions are derived for the point density that minimizes the expected distortion. For the mean squared error distortion, an upper bound on the asymptotic (i.e., high-rate) distortion is derived by assuming a certain structure

on the codebook. This structure enables the extension of the analysis to source coders with *singular* point densities. It shown that, for channels with small errors, the point density that minimizes the upper bound is continuous, while as the error rate increases, the point density becomes singular, and the extent of the singularity can be analytically characterized.

In the fifth chapter of the thesis, new results on the performance of the high-rate vector quantization of random sources when the quantized index is transmitted over a noisy channel are derived for two specific applications. The first is the quantization of the CSI for MISO systems with beamforming at the transmitter. Here, it is assumed that there exists a per-antenna power constraint at the transmitter, hence, the EGT beamforming vector is quantized and sent from the receiver to the transmitter over a noisy discrete symmetric channel with random index assignment. The loss in received SNR is analytically characterized, and it is shown that at high rates, the overall distortion can be expressed as the sum of the quantization-induced distortion and the channel error-induced distortion. The optimum density of codepoints (also known as the point density) that minimizes the overall distortion subject to a boundedness constraint is shown to be the uniform density. Also, it is found that the high-rate performance depends on the behavior of the noisy feedback channel as the number of codepoints gets large. The binary symmetric channel with random index assignment is a special case of the analysis, and it is shown that as the number of quantized bits gets large, the distortion approaches that obtained with random beamforming, i.e., feedback is useless if no error control coding is employed. The second application is in the wideband speech compression problem, i.e., that of quantizing the linear predictive coding parameters in speech coding systems with the log spectral distortion as performance metric. It is shown that the theory is able to correctly predict the channel error rate that is permissible for operation at a particular distortion level.

2 Training-Based and Semi-Blind Channel Estimation for MIMO Systems with Maximum Ratio Transmission

2.1 Introduction

MIMO and smart antenna systems have gained popularity due to the promise of a linear increase in achievable data rate with the number of antennas, and because they inherently benefit from effects such as channel fading. Maximum Ratio Transmission (MRT) is a particularly attractive beamforming scheme for MIMO communication systems because of its low implementation complexity. It is also known that MRT coupled with maximum ratio combining (MRC) leads to SNR maximization at the receiver and achieves a performance close to capacity in low SNR scenarios. However, in order to realize these benefits, an accurate estimate of the channel is necessary. One standard technique to estimate the channel is to transmit a sequence of training symbols (also called pilot symbols) at the beginning of each frame. This training symbol sequence is known at the receiver and thus the channel is estimated from the measured outputs to training symbols. Training based schemes usually have very low complexity making them ideally suited for implementation in systems (e.g., mobile stations) where the available

computational capacity is limited.

However, the above training-based technique for channel estimation in MRT based MIMO systems is transmission scheme agnostic. For example, channel estimation algorithms when MRT is employed at the transmitter only need to estimate $\mathbf{v}_1$ and $\mathbf{u}_1$, where $\mathbf{v}_1$ and $\mathbf{u}_1$ are the dominant eigenvectors of $H^H H$ and HH^H respectively, H is the $r \times t$ channel transfer matrix, and $r\,/\,t$ are the number of receive / transmit antennas. Hence, techniques that estimate the entire H matrix from a set of training symbols and use the estimated H to compute $\mathbf{v}_1$ and $\mathbf{u}_1$ may be inefficient, compared to techniques designed to use the training data specifically for estimating the beamforming vectors. Moreover, as r increases, the mean squared error (MSE) in estimation of $\mathbf{v}_1 \in \mathbb{C}^t$ remains constant since the number of unknown parameters in $\mathbf{v}_1$ does not change with r, while that of H increases since the number of elements, rt, grows linearly with r. Added to this, the complexity of reliably estimating the channel increases with its dimensionality. The channel estimation problem is further complicated in MIMO systems because the SNR per bit required to achieve a given system throughput performance decreases as the number of antennas is increased. Such low SNR environments call for more training symbols, lowering the effective data rate.

For the above reasons, semi-blind techniques can enhance the accuracy of channel estimation by efficiently utilizing not only the known training symbols but also the unknown data symbols. Hence, they can be used to reduce the amount of training data required to achieve the desired system performance, or equivalently, achieve better accuracy of estimation for a given number of training symbols, thereby improving the spectral efficiency and channel throughput. Work on semi-blind techniques for the design of fractional semi-blind equalizers in multipath channels has been reported earlier by Pal in [9, 10]. In [11, 12] error bounds and asymptotic properties of blind and semi-blind techniques are analyzed. In [41–43], an orthogonal pilot based maximum likelihood (OPML) semi-blind estimation scheme is proposed, where the channel matrix H is factored into the prod-

uct of a whitening matrix W and a unitary rotation matrix Q. W is estimated from the data using a blind algorithm, while Q is estimated exclusively from the training data using the OPML algorithm. However, feedback-based transmission schemes such as MRT pose new challenges for semi-blind estimation, because employment of the precoder (beamforming vector) corresponding to an erroneous channel estimate precludes the use of the received data symbols to improve the channel estimate. This necessitates the development of new transmission schemes to enable implementation of semi-blind estimation, as shown in Section 2.2-2.2.3. Furthermore, the proposed techniques specifically estimate the MRT beamforming vector and hence can potentially achieve better estimation accuracy compared to techniques that are independent of the transmission scheme.

The contributions of this chapter are as follows. We describe the training-only based conventional least squares estimation (CLSE) algorithm, and derive analytical expressions for the MSE in the beamforming vector, the mean received SNR and the symbol error rate (SER) performance. For improved spectral efficiency (reduced training overhead), we propose a closed-form semi-blind (CFSB) algorithm that estimates $\mathbf{u}_1$ from the data using a blind algorithm, and estimates $\mathbf{v}_1$ exclusively from the training. This necessitates the introduction of a new signal transmission scheme that involves transmission of information-bearing spectrally white data symbols to enable semi-blind estimation of the beamforming vectors. Expressions are derived for the performance of the proposed CFSB scheme. We show that given perfect knowledge of $\mathbf{u}_1$ (which can be achieved when there are a large number of white data symbols), the error in estimating $\mathbf{v}_1$ using the semi-blind scheme asymptotically achieves the theoretical Cramer-Rao lower bound (CRB), and thus the CFSB scheme outperforms the CLSE scheme. However, there is a trade-off in transmission of white data symbols in semi-blind estimation, since the SER for the white data is frequently greater than that for the beamformed data. Thus, we show that there exist scenarios where for a reasonable number of white data symbols, the gains from beamformed data for this improved estimate in CFSB outweigh

the loss in performance due to transmission of white data. As a more general estimation method when a given number of blind data symbols are available, we propose a new scheme that judiciously combines the above described CFSB and CLSE estimates based on a heuristic criterion. Through Monte-Carlo simulations, we demonstrate that this proposed linearly-combined semi-blind (LCSB) scheme outperforms the CLSE and CFSB scheme in terms of both estimation accuracy as well as SER and thus achieves good performance.

The rest of this chapter is organized as follows. In Section 2.2, we present the problem setup and notation. We also present both the CFSB and CLSE schemes in detail. The MSE and the received SNR performance of the CLSE scheme are derived using a first order perturbation analysis in Section 2.3 and the performance of the CFSB scheme is analyzed in Section 2.4. In Section 2.5, to conduct an end-to-end system comparison, we derive the performance of Alamouti space-time coded data with training-based channel estimation, and present the proposed LCSB algorithm. We compare the different schemes through Monte-Carlo simulations in Section 2.6 and present our conclusions in Section 2.7.

2.2 Preliminaries

2.2.1 System Model and Notation

Fig. 2.1 shows the MIMO system model with beamforming at the transmitter and the receiver. We model a flat-fading channel by a complex-valued channel matrix $H \in \mathbb{C}^{r\times t}$. We assume that H is quasi-static and constant over the period of one transmission block. We denote the singular value decomposition (SVD) of H by $H = U\Sigma V^H$, and $\Sigma \in \mathbb{R}^{r\times t}$ contains singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m > 0$, along the diagonal, where $m = \text{rank}(H)$. Let $\mathbf{v}_1$ and $\mathbf{u}_1$ denote the first columns of V and U, respectively.

The channel input-output relation at time instant k is

$$\mathbf{y}_k = H\mathbf{x}_k + \mathbf{n}_k, \tag{2.1}$$

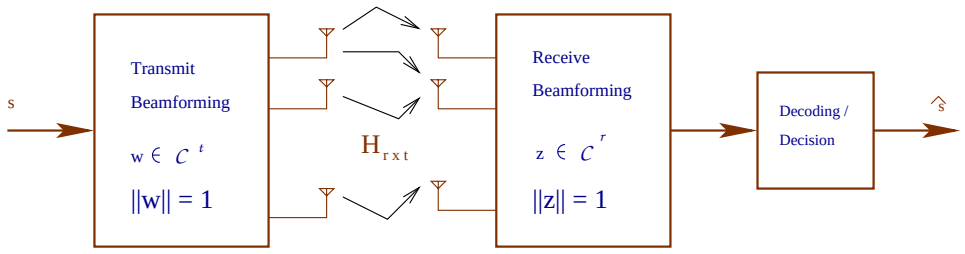


Figure 2.1: MIMO system model, with beamforming at the transmitter and receiver.

where $\mathbf{x}_k \in \mathbb{C}^t$ is the channel input, $\mathbf{y}_k \in \mathbb{C}^r$ is the channel output, and $\mathbf{n}_k \in \mathbb{C}^r$ is the spatially and temporally white noise vector with *i.i.d.* zero mean circularly symmetric complex Gaussian (ZMCSCG) entries. The input $\mathbf{x}_k$ could denote either be data or training symbols. Also, we let the noise power in each receive antenna be unity, that is, $\mathbf{E}\left\{\mathbf{n}_k\mathbf{n}_k^H\right\} = I_r$, where $\mathbf{E}\left\{\cdot\right\}$ denotes the expectation operation, and I_r is the $r \times r$ identity matrix.

Let L training symbol vectors be transmitted at an average power P_T per vector (T stands for 'training'). The training symbols are stacked together to form a training symbol matrix $X_p \in \mathbb{C}^{t\times L}$ as $X_p = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L]$ (p stands for 'pilot'). We employ orthogonal training sequences because of their optimality properties in channel estimation [44]. That is, $X_p X_p^H = \gamma_p I_t$, where $\gamma_p \triangleq LP_T/t$, thus maintaining the training power of P_T. The data symbols $\mathbf{x}_k$ could either be spatially-white (i.e., $\mathbf{E}\left\{\mathbf{x}_k\mathbf{x}_k^H\right\} = (P_D/t)\, I_t$), or it could be the result of using beamforming at the transmitter with unit-norm weight vector $\mathbf{w} \in \mathbb{C}^{t\times 1}$ $\left(\text{i.e., } \mathbf{E}\{\mathbf{x}_k\mathbf{x}_k^H\} = P_D\mathbf{w}\mathbf{w}^H\right)$, where the data transmit power is $\mathbf{E}\left\{\mathbf{x}_k^H\mathbf{x}_k\right\} = P_D$ (D stands for 'data'). We let N denote the number of spatially-white data symbols transmitted, that is, a total of $N + L$ symbols are transmitted prior to transmitting beamformed-data. Note that the N white data symbols carry (unknown) information bits, and hence are not a waste of available bandwidth.

In this chapter, we restrict our attention to the case where the transmitter employs MRT to send data, that is, a *single* data stream is transmitted over t

transmit antennas after passing through a beamformer $\mathbf{w}$. Given the channel matrix H, the optimum choice of $\mathbf{w}$ is $\mathbf{v}_1$ [7]. Thus, MRT only needs an accurate estimate of $\mathbf{v}_1$ to be fed-back to the transmitter. We assume that $t \geq 2$, since when $t = 1$, estimation of the beamforming vector has no relevance. Finally, we will compare the performance of different estimation techniques using several different measures, namely, the MSE in the estimate of $\mathbf{v}_1$, the gain (rather, the power amplification/attenuation), and the symbol error rate (SER) of the one-dimensional channel resulting from beamforming with the estimated vector $\hat{\mathbf{v}}_1$ assuming uncoded M-ary QAM transmission. The performance of a practical communication would also be affected by factors such as quantization error in $\hat{\mathbf{v}}_1$, errors in the feedback channel, feedback delay in time-varying environments, etc., and a detailed study of these factors warrant separate treatment.

2.2.2 Conventional Least Squares Estimation (CLSE)

Here, an ML estimate of the channel matrix, $\hat{H}_c$, is first obtained from the training data as the solution to the following least squares problem:

$$\hat{H}_c = \arg \min_{G \in \mathbb{C}^{r \times t}} \|Y_p - GX_p\|_F^2, \tag{2.2}$$

where $\|\cdot\|_F$ represents the Frobenius norm, Y_p is the $r \times L$ matrix of received symbols given by $Y_p = HX_p + \eta_p$, where $\eta_p \in \mathbb{C}^{r \times L}$ is the set of AWGN (spatially and temporally white) vectors. From [45], the solution to this least squares estimation problem can be shown to be $\hat{H}_c = Y_p X_p^\dagger$, where $X_p^\dagger$ is the Moore-Penrose generalized inverse of X_p. Since orthogonal training sequences are employed, we have $X_p^\dagger = \frac{1}{\gamma_p} X_p^H$, and consequently

$$\hat{H}_c = \frac{1}{\gamma_p} Y_p X_p^H. \tag{2.3}$$

The ML estimate of $\mathbf{v}_1$ and $\mathbf{u}_1$, denoted $\hat{\mathbf{v}}_c$ and $\hat{\mathbf{u}}_c$ respectively, is now obtained via an SVD of the estimated channel matrix $\hat{H}_c$. Since $\hat{H}_c$ is the ML estimate of H, from properties of ML estimation of principal components [46], the $\hat{\mathbf{v}}_c$ obtained by this technique is also the ML estimate of $\mathbf{v}_1$ given only the training data.

2.2.3 Semi-Blind Estimation

In the scenario that the transmitted data symbols are spatially-white, the ML estimate of $\mathbf{u}_1$ is the dominant eigenvector of the output correlation matrix $\hat{R}_y$, which is estimated as $\hat{R}_y = \sum_{i=1}^{N} \mathbf{y}_i \mathbf{y}_i^H$. Now, the estimate of $\mathbf{u}_1$ is obtained by computing the following SVD

$$\hat{U}\tilde{\Sigma}^2\hat{U}^H = \hat{R}_y. \tag{2.4}$$

Note that it is possible to use the entire received data to compute $\hat{R}_y$ in (2.4) rather than just the data symbols, in this case, N should be changed to $N + L$. The estimate of $\mathbf{u}_1$, denoted $\hat{\mathbf{u}}_s$ (the subscript 's' stands for semi-blind), is thus computed blind from the received data as the first column of $\hat{U}$. As N grows, a near perfect estimate of $\mathbf{u}_1$ can be obtained.

In order to estimate $\mathbf{u}_1$ as described above, it is necessary that the transmitted symbols be *spatially-white*. If the transmitter uses any (single) beamforming vector $\mathbf{w}$, the expected value of the correlation at the receiver is $H\mathbf{w}(H\mathbf{w})^H = H\mathbf{w}\mathbf{w}^H H^H \neq HH^H$, and hence, the estimated eigenvector will be a vector proportional to $H\mathbf{w}$ instead of $\mathbf{u}_1$. Fig. 2.2 shows a schematic representation of the CLSE and the CFSB schemes. Thus, the CFSB scheme involves a two-phase data transmission: spatially-white data followed by beamformed data. White data transmission could lead to a loss of performance relative to beamformed data, but this performance loss can be compensated for by the gain obtained from the improved estimate of the MRT beamforming vector. Thus, the semi-blind scheme can have an overall better performance than the CLSE scheme. Section 2.5 presents an overall SER comparison in a practical scenario, after accounting for the performance of the white data as well as for the beamformed data.

Having obtained the estimate of $\mathbf{u}_1$ from the white data, the training symbols are now exclusively used to estimate $\mathbf{v}_1$. Since the vector $\mathbf{v}_1$ has fewer real parameters $(2t-1)$ than the channel matrix H $(2rt)$, it is expected to achieve a greater accuracy of estimation for the same number of training symbols, compared

Figure 2.2: Comparison of the transmission scheme for conventional least squares (CLSE) and closed-form semi-blind (CFSB) estimation.

to the CLSE technique which requires an accurate estimate of the full H matrix in order to estimate $\mathbf{v}_1$ accurately. If $\mathbf{u}_1$ is estimated perfectly from the blind data, the received training symbols can be filtered by ${\mathbf{u}_1}^H$ to obtain

$$ {\mathbf{u}_1}^H Y_p = \sigma_1 {\mathbf{v}_1}^H X_p + {\mathbf{u}_1}^H \eta_p. \tag{2.5} $$

Since $\|\mathbf{u}_1\| = 1$, (here $\|\cdot\|$ represents the 2-norm) the statistics of the Gaussian noise η_p are unchanged by the above operation. We seek the estimate of $\mathbf{v}_1$ as the solution to the following least squares problem

$$ \hat{\mathbf{v}}_s = \arg \min_{\mathbf{v} \in \mathbb{C}^t,\, \|\mathbf{v}\|=1} \|{\mathbf{u}_1}^H Y_p - \mathbf{v}^H X_p \sigma_1\|^2, \tag{2.6} $$

where $\hat{\mathbf{v}}_s$ denotes the semi-blind estimate of $\mathbf{v}_1$. The following lemma establishes the solution.

Lemma 1. *If X_p satisfies $X_p X_p^H = \gamma_p I_t$, the least squares estimate of $\mathbf{v}_1$ (under $\|\mathbf{v}_1\| = 1$) given perfect knowledge of $\mathbf{u}_1$ is*

$$ \hat{\mathbf{v}}_s = \frac{X_p Y_p^H \mathbf{u}_1}{\|X_p Y_p^H \mathbf{u}_1\|}. \tag{2.7} $$

Proof. See Appendix 2.8.1. □

Closed-Form Semi-Blind Estimation Algorithm (CFSB)

Based on the above observations, the proposed CFSB algorithm is as follows. First, we obtain $\hat{\mathbf{u}}_s$, the estimate of $\mathbf{u}_1$, from (2.4). Then, we estimate $\mathbf{v}_1$

from the L training symbols by substituting $\hat{\mathbf{u}}_s$ for $\mathbf{u}_1$ in (2.7). This requires $L+N$ symbols to actually estimate $\mathbf{v}_1$, however, N of these symbols are data symbols. Hence, we can potentially achieve the desired accuracy of estimation of $\mathbf{v}_1$ using fewer training symbols compared to the CLSE technique.

An alternative to employing $\hat{\mathbf{u}}_1$ at the receiver is employ maximum ratio combining (MRC), i.e., to use an estimate of $H\hat{\mathbf{v}}_1/\|H\hat{\mathbf{v}}_1\|$ (which can be accurately estimated as the dominant eigenvector of the sample covariance matrix of the beamformed data). The performance of such a scheme can be derived using the same techniques presented in this chapter.

In the next section, we present a theoretical analysis of the MSE in $\hat{\mathbf{v}}_s$ and the received SNR with both the CLSE and the CFSB techniques, which will give us insight into the trade-offs involved in implementing the two methods, and suggest strategies to further improve the CFSB algorithm.

2.3 Conventional Least Squares Estimation (CLSE)

2.3.1 Perturbation of Eigenvectors

We recapitulate a result from matrix perturbation theory [47] that we will use frequently in the sequel. Consider a first order perturbation of a hermitian symmetric matrix R by an error matrix ΔR to get $\hat{R}$, that is, $\hat{R} = R + \Delta R$. Then, if the eigenvalues of R are distinct, for small perturbations, the eigenvectors $\hat{\mathbf{s}}_k$ of $\hat{R}$ can be approximately expressed in terms of the eigenvectors $\mathbf{s}_k$ of R as

$$\hat{\mathbf{s}}_k = \mathbf{s}_k + \sum_{\substack{r=1 \\ r \neq k}}^{n} \frac{\mathbf{s}_r^H \Delta R \mathbf{s}_k}{\lambda_k - \lambda_r} \mathbf{s}_r, \tag{2.8}$$

where n is the rank of R, λ_k is the k-th eigenvalue of R, and $\lambda_k \neq \lambda_j,\ k \neq j$.

When $k = 1$, we have $\hat{\mathbf{s}}_\mathbf{1} = S\check{\mathbf{d}}$, where $S = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n]$ is the matrix of eigenvectors and $\check{\mathbf{d}} = [1, \frac{\mathbf{s}_2^H \Delta R \mathbf{s}_1}{\lambda_1 - \lambda_2}, \ldots, \frac{\mathbf{s}_n^H \Delta R \mathbf{s}_1}{\lambda_1 - \lambda_n}]^T$. One could scale the vector $\hat{\mathbf{s}}_1$ to construct a unit-norm vector as $\tilde{\mathbf{s}}_1 = \hat{\mathbf{s}}_1/\|\hat{\mathbf{s}}_1\|$. Then, $\tilde{\mathbf{s}}_1 = S\mathbf{d}$, where $\mathbf{d} = \check{\mathbf{d}}/\|\check{\mathbf{d}}\| = [1 + \Delta d_1, \Delta d_2, \ldots, \Delta d_n]^T$. Following an approach similar to [48], if Δd_i are small,

since $\|\mathbf{d}\| = 1$, the components Δd_i are approximately given by

$$\begin{aligned} \Delta d_i &\simeq \frac{\mathbf{s}_i^H \Delta R \mathbf{s}_1}{\lambda_1 - \lambda_i},\ i = 2, \ldots, n \\ \Delta d_1 &\simeq -\frac{1}{2} \sum_{i=1}^{n} |\Delta d_i|^2 . \end{aligned} \tag{2.9}$$

Note that Δd_1 is real, and is a higher-order term compared to $\Delta d_i,\ i \geq 2$. We will use this fact in our first-order approximations to ignore terms such as $|\Delta d_1|^2, |\Delta d_1|^3, \ldots$ and $|\Delta d_i|^3, |\Delta d_i|^4, \ldots,\ i \geq 2$. In the sequel, we assume that the dominant singular value of H is distinct, so the conditions required for the above result are valid.

2.3.2 MSE in $\hat{\mathbf{v}}_c$

To compute the MSE in $\hat{\mathbf{v}}_c$, we use (2.3) to write the matrix $\hat{H}_c^H \hat{H}_c$ as a perturbation of $H^H H$ and use the above matrix perturbation result to derive the desired expressions.

$$\hat{H}_c^H \hat{H}_c = V \Sigma^2 V^H + E_t, \tag{2.10}$$

where $E_t \approx \left[V \Sigma U^H E_p + {E_p}^H U \Sigma V^H \right]$ with $E_p = \frac{1}{\gamma_p} \eta_p X_p^H$. Here, we have ignored the ${E_p}^H E_p$ term in writing the expression for E_t, since it is a second order term due to the $\frac{1}{\gamma_p}$ factor in E_p. Now, we can regard E_t as a perturbation of the matrix $H^H H$. As seen in Section 2.2(2.2.2), $\hat{\mathbf{v}}_c$ is estimated from the SVD of $\hat{H}_c$. Since the basis vectors V span $\mathbb{C}^t$, we can let $\hat{\mathbf{v}}_c = V\mathbf{d}$, and write $\mathbf{d} = [1 + \Delta d_1, \Delta d_2, \ldots, \Delta d_t]^T$ as a perturbation of $[1, 0, \ldots, 0]^T$.

For $i \geq 2$, Δd_i is obtained from (2.9) as

$$\Delta d_i = \frac{{\mathbf{v}_i}^H E_t \mathbf{v}_1}{\sigma_1^2 - \sigma_i^2} = \frac{\sigma_i \mathbf{u}_i^H E_p \mathbf{v}_1 + \sigma_1 \mathbf{v}_i^H {E_p}^H \mathbf{u}_1}{\sigma_1^2 - \sigma_i^2}. \tag{2.11}$$

Note that, if $r < t$, we have $\sigma_i = 0,\ i > r$, hence, $\Delta d_i = \mathbf{v}_i^H {E_p}^H \mathbf{u}_1 / \sigma_1$, for $i > r$. Therefore, to simplify notation, we can define $\mathbf{u}_i \triangleq \mathbf{0}_{r \times 1}$ and $\mathbf{v}_j \triangleq \mathbf{0}_{t \times 1}$, for $i > r$ and $j > t$ respectively. The following result is used to find $\mathbf{E}\left\{ |\Delta d_i|^2 \right\}$.

Lemma 2. *Let $\mu_1, \mu_2 \in \mathbb{C}$ be fixed complex numbers. Let $\sigma_p^2 = \frac{1}{\gamma_p}$ denote the variance of one of the elements of E_p. Then,*

$$\mathbf{E}\left\{\left|\mu_1 \mathbf{u}_i{}^H E_p \mathbf{v}_j + \mu_2 \mathbf{v}_i{}^H E_p{}^H \mathbf{u}_j\right|^2\right\} = \sigma_p^2 \left(|\mu_1|^2 + |\mu_2|^2\right),$$

for any $1 \leq i \leq r,\ 1 \leq j \leq t$.

Proof. Let $a \triangleq \mathbf{u}_i{}^H E_p \mathbf{v}_j$ and $b \triangleq \mathbf{v}_i{}^H E_p{}^H \mathbf{u}_j$. Then, from lemma 6 in Section 2.8.5 of the Appendix, a and b are circularly symmetric random variables. Since E_p is circularly symmetric ($\mathbf{E}\left\{E_p(i,j)\, E_p(k,l)\right\} = 0, \forall\, i,j,k,l$) and a and b^* are both linear combinations of elements of E_p, we have $\mathbf{E}\left\{ab^*\right\} = 0$. Finally, since $\|\mathbf{u}_i\| = \|\mathbf{v}_j\| = 1$, the variance of a and b are equal, and $\sigma_a^2 = \sigma_b^2 = \sigma_p^2$. Substituting,

$$\begin{aligned}\mathbf{E}\left\{\left|\mu_1 \mathbf{u}_i{}^H E_p \mathbf{v}_j + \mu_2 \mathbf{v}_i{}^H E_p{}^H \mathbf{u}_j\right|^2\right\} &= |\mu_1|^2 \sigma_a^2 + |\mu_2|^2 \sigma_b^2 \\ &= \sigma_p^2 \left(|\mu_1|^2 + |\mu_2|^2\right).\end{aligned}$$

□

Using the above lemma with $\mu_1 = \sigma_i, \mu_2 = \sigma_1$ and $j = 1$, for $i \geq 2$,

$$\mathbf{E}\left\{|\Delta d_i|^2\right\} = \sigma_p^2 \frac{\sigma_1^2 + \sigma_i^2}{\left(\sigma_1^2 - \sigma_i^2\right)^2}, \tag{2.12}$$

where the expectation is taken with respect to the AWGN term η_p. The following lemma helps simplify the expression further.

Lemma 3. *If $\hat{\mathbf{v}}_c = V\mathbf{d}$, then*

$$\|\hat{\mathbf{v}}_c - \mathbf{v}_1\|^2 = 2\left(1 - \mathrm{Re}(d_1)\right) = -\left(\Delta d_1 + \Delta d_1^*\right), \tag{2.13}$$

where $d_1 = 1 + \Delta d_1$ is the first element of $\mathbf{d}$.

Using (2.12) in (2.9) and substituting into in (2.13), the final estimation error is

$$\mathbf{E}\left\{\|\hat{\mathbf{v}}_c - \mathbf{v}_1\|^2\right\} \;=\; \frac{1}{\gamma_p} \sum_{i=2}^{t} \frac{\sigma_1^2 + \sigma_i^2}{\left(\sigma_1^2 - \sigma_i^2\right)^2}. \tag{2.14}$$

2.3.3 Received SNR and Symbol Error Rate (SER)

In this section, we derive the expression for the received SNR when beamforming using $\hat{\mathbf{v}}_c$ at the transmitter and filtering using $\hat{\mathbf{u}}_c$ at the receiver. Since the unitary matrices V and U span $\mathbb{C}^t$ and $\mathbb{C}^r$, $\hat{\mathbf{v}}_c$ and $\hat{\mathbf{u}}_c$ can be expressed as $\hat{\mathbf{u}}_c = U\mathbf{c}$ and $\hat{\mathbf{v}}_c = V\mathbf{d}$ respectively. Borrowing notation from Section 2.3-2.3.2, let $\mathbf{c} = [1 + \Delta c_1, \Delta c_2, \ldots, \Delta c_r]^T \in \mathbb{C}^r$ and $\mathbf{d} = [1 + \Delta d_1, \Delta d_2, \ldots, \Delta d_t]^T \in \mathbb{C}^t$ respectively. Then, $\mathbf{c}$ can be derived by a perturbation analysis on $\hat{H}_c\hat{H}_c^H$ analogous to that in (2.10) in Section 2.3-2.3.2. We obtain

$$\Delta c_i = \frac{\sigma_i \mathbf{v}_i^H {E_p}^H \mathbf{u}_1 + \sigma_1 \mathbf{u}_i^H E_p \mathbf{v}_1}{\sigma_1^2 - \sigma_i^2},$$

where, as before, we define $\sigma_i = 0$, and $\mathbf{u}_i \triangleq \mathbf{0}_{r\times 1}$, $\mathbf{v}_j \triangleq \mathbf{0}_{t\times 1}$, for $i > r$ and $j > t$ respectively, so that $\Delta c_i = 0; i > r$, as expected. The channel gain is given by

$$\hat{\mathbf{u}}_c^H H \hat{\mathbf{v}}_c = \mathbf{c}^H \Sigma \mathbf{d} = \sigma_1(1 + \Delta d_1)(1 + \Delta c_1^*) + \sum_{i=2}^{t} \sigma_i \Delta c_i^* \Delta d_i.$$

Ignoring higher order terms (cf. Section 2.2(2.3.1)), the power amplification $\rho_c \triangleq \mathbf{E}\left\{\left|\hat{\mathbf{u}}_c^H H \hat{\mathbf{v}}_c\right|^2\right\}$ is

$$\begin{aligned}\rho_c \approx \sigma_1^2 \mathbf{E}\Big\{ & 1 + (\Delta d_1 + \Delta d_1^*) + (\Delta c_1 + \Delta c_1^*) \\ & + \sum_{i=2}^{t} \frac{\sigma_i}{\sigma_1} \left(\Delta c_i \Delta d_i^* + \Delta c_i^* \Delta d_i\right) \Big\}. \end{aligned} \tag{2.15}$$

From (2.9) and (2.12), we have

$$\begin{aligned} \mathbf{E}\left\{\Delta d_1 + \Delta d_1^*\right\} &= -\frac{1}{\gamma_p} \sum_{i=2}^{t} \frac{\sigma_1^2 + \sigma_i^2}{\left(\sigma_1^2 - \sigma_i^2\right)^2}; \\ \mathbf{E}\left\{\Delta c_1 + \Delta c_1^*\right\} &= -\frac{1}{\gamma_p} \sum_{i=2}^{r} \frac{\sigma_1^2 + \sigma_i^2}{\left(\sigma_1^2 - \sigma_i^2\right)^2}. \end{aligned}$$

Now, $\Delta c_i \Delta d_i^*$ can be written as

$$\begin{aligned}
\mathbf{E}\{ \quad \Delta c_i \quad \Delta d_i^* \} &= \mathbf{E}\left\{ \left(\frac{\sigma_i \mathbf{v}_i^H {E_p}^H \mathbf{u}_1 + \sigma_1 \mathbf{u}_i^H E_p \mathbf{v}_1}{\sigma_1^2 - \sigma_i^2} \right) \right. \\
&\qquad \left. \left(\frac{\sigma_i \mathbf{v}_1{}^H {E_p}^H \mathbf{u}_i + \sigma_1 \mathbf{u}_1{}^H E_p \mathbf{v}_i}{\sigma_1^2 - \sigma_i^2} \right) \right\}, \\
&= \mathbf{E}\left\{ \frac{\sigma_1 \sigma_i \left(\left| \mathbf{u}_i^H E_p \mathbf{v}_1 \right|^2 + \left| \mathbf{v}_i^H {E_p}^H \mathbf{u}_1 \right|^2 \right)}{\left(\sigma_1^2 - \sigma_i^2\right)^2} \right\}, \\
&= \frac{2\sigma_p^2 \sigma_1 \sigma_i}{\left(\sigma_1^2 - \sigma_i^2\right)^2} = \frac{2\sigma_1 \sigma_i}{\gamma_p \left(\sigma_1^2 - \sigma_i^2\right)^2}.
\end{aligned}$$

And likewise for $\Delta c_i^* \Delta d_i$. Denoting $m \triangleq \mathrm{rank}(H)$, the power amplification is

$$\begin{aligned}
\rho_c &= \sigma_1^2 \left(1 - \frac{1}{\gamma_p} \sum_{i=2}^{t} \frac{\sigma_1^2 + \sigma_i^2}{\left(\sigma_1^2 - \sigma_i^2\right)^2} - \frac{1}{\gamma_p} \sum_{i=2}^{r} \frac{\sigma_1^2 + \sigma_i^2}{\left(\sigma_1^2 - \sigma_i^2\right)^2} \right. \\
&+ \left. \frac{4}{\gamma_p} \sum_{i=2}^{m} \frac{\sigma_i^2}{\left(\sigma_1^2 - \sigma_i^2\right)^2} \right), \\
&= \sigma_1^2 - \frac{2}{\gamma_p} \sum_{i=2}^{m} \frac{\sigma_1^2}{\sigma_1^2 - \sigma_i^2} - \frac{1}{\gamma_p} (r + t - 2m). \qquad (2.16)
\end{aligned}$$

In obtaining (2.16), we have used the fact that $\sigma_i = 0$ for $i > m$, where $m = \mathrm{rank}(H)$. Finally, the received SNR is

$$\mathrm{SNR} = \rho_c P_D, \qquad (2.17)$$

where P_D is the power per data symbol. The power amplification with perfect knowledge of H at the transmitter and the receiver is $\rho_p \triangleq \sigma_1^2$. As $\gamma_p = LP_T/t$ increases, ρ_c approaches ρ_p. Note that, when $r = 1$, the above expression simplifies to $\rho_c = \rho_p - \frac{1}{\gamma_p}(t-1)$. Also, $\sum_{i=2}^{m} \frac{\sigma_1^2}{\sigma_1^2 - \sigma_i^2} \geq (m-1)$ since $\sigma_1 \geq \sigma_i$. Hence, if $r = t$, the CLSE performs best when the channel is spatially single dimensional (for example, in keyhole channels or highly correlated channels), that is, $\sigma_i = 0,\ i \geq 2$. In this case, we have $\rho_c = \rho_p - \frac{2}{\gamma_p}(t-1)$. At the other extreme, if the dominant singular values are very close to each other such that $(\sigma_1^2 - \sigma_2^2) < 2/\gamma_p$, the analysis is incorrect because it requires that the dominant singular values of H be sufficiently separated. For Rayleigh fading channels, i.e., H has i.i.d. ZMCSCG entries of unit

variance, we can numerically evaluate the probability $Pr\{\sigma_1^2 - \sigma_2^2 < 2/\gamma_p\}$ to be approximately 1.7×10^{-4}, with $r = t = 4$ and a typical value of $\gamma_p = 10$dB. Thus, the above analysis is valid for most channel instantiations.

Having determined the expected received SNR for a given channel instantiation, assuming uncoded M-ary QAM transmission, the corresponding SER P_M is given as [49]

$$P_{\sqrt{M}}(\rho_c) = 2\left[1 - \frac{1}{\sqrt{M}}\right] Q\left(\sqrt{\frac{3\rho_c P_T}{M-1}}\right) \tag{2.18}$$

$$P_M(\rho_c) = 1 - \left(1 - P_{\sqrt{M}}(\rho_c)\right)^2, \tag{2.19}$$

where $Q(\cdot)$ is the Gaussian Q-function, and ρ_c is given by (2.16). The above expression can now be averaged over the probability density function of σ_i^2 through numerical integration.

2.4 Closed-Form Semi-Blind estimation (CFSB)

First, recall that the first order Taylor expansion of a function of two variables $g(x, y)$ is given by

$$\begin{aligned} g(x + \Delta x, y + \Delta y) - g(x, y) &= \frac{\partial g(x,y)}{\partial x}\Delta x + \frac{\partial g(x,y)}{\partial y}\Delta y \\ &\quad + O\left(\Delta x^2\right) + O\left(\Delta y^2\right) \\ &\approx \left[g\left(x + \Delta x, y\right) - g\left(x, y\right)\right] + \left[g\left(x, y + \Delta y\right) - g\left(x, y\right)\right]. \end{aligned}$$

Now, in CFSB, the error in $\mathbf{v}_1$ (or loss in SNR) occurs due to two reasons: first, the noise in the received training symbols, and second, the use of an imperfect estimate of $\mathbf{u}_1$ (from the noise in the data symbols and availability of only a finite number N of unknown white data). More precisely, let the estimator of $\mathbf{v}_1$ be expressed as a function $\hat{\mathbf{v}}_s = f(Y_p, \hat{\mathbf{u}}_s)$ of the two variables Y_p and $\hat{\mathbf{u}}_s$. Using the above expansion, we have

$$\begin{aligned} f\left(Y_p, \hat{\mathbf{u}}_s\right) - f\left(HX_p, \mathbf{u}_1\right) &\approx \left[f\left(Y_p, \mathbf{u}_1\right) - f\left(HX_p, \mathbf{u}_1\right)\right] \\ &\quad + \left[f\left(HX_p, \hat{\mathbf{u}}_s\right) - f\left(HX_p, \mathbf{u}_1\right)\right] \end{aligned} \tag{2.20}$$

where $\hat{\mathbf{v}}_s = f\left(Y_p, \hat{\mathbf{u}}_s\right)$ and from (2.7), $\mathbf{v}_1 = f\left(HX_p, \mathbf{u}_1\right)$. Since the training noise η_p and the error in the estimate $\hat{\mathbf{u}}_s$ are mutually independent, we get

$$\begin{aligned}\mathbf{E}\left\{\|\hat{\mathbf{v}}_s - \mathbf{v}_1\|^2\right\} \approx & \underbrace{\mathbf{E}\left\{\|f(Y_p, \mathbf{u}_1) - f(HX_p, \mathbf{u}_1)\|^2\right\}}_{T_1} \\ & + \underbrace{\mathbf{E}\left\{\|f(HX_p, \hat{\mathbf{u}}_s) - f(HX_p, \mathbf{u}_1)\|^2\right\}}_{T_2}. \end{aligned} \tag{2.21}$$

Note that the term T_1 represents the MSE in $\hat{\mathbf{v}}_s$ as if the receiver had perfect knowledge of $\hat{\mathbf{u}}_s$ (i.e., $\hat{\mathbf{u}}_s = \mathbf{u}_1$), and the term T_2 represents the MSE in $\hat{\mathbf{v}}_s$ when the training symbols are noise-free (i.e., $Y_p = HX_p$). Hence, the error in $\hat{\mathbf{v}}_s$ can be thought of as the sum of two terms: the first one being the error due to the noise in the white (unknown) data, and the second being the error due to the noise in the training data. A similar decomposition can be used to express the loss in channel gain (relative to σ_1).

2.4.1 MSE in $\hat{\mathbf{v}}_s$ with Perfect $\hat{\mathbf{u}}_s$

In this section we consider the error arising exclusively from the training noise, by setting $\hat{\mathbf{u}}_s = \mathbf{u}_1$. Let $\tilde{\mathbf{v}}_s$ be defined as $\tilde{\mathbf{v}}_s \triangleq \frac{X_p Y_p^H \mathbf{u}_1}{\sigma_1 \gamma_p}$. Then, from (2.5)

$$\tilde{\mathbf{v}}_s = \mathbf{v}_1 + \frac{{E_p}^H \mathbf{u}_1}{\sigma_1},$$

where, $E_p \triangleq \eta_p X_p^H / \gamma_p$, as before. Recall from (2.7) that $\hat{\mathbf{v}}_s = \frac{\tilde{\mathbf{v}}_s}{\|\tilde{\mathbf{v}}_s\|}$. Now, $\|\tilde{\mathbf{v}}_s\|$ can be simplified as $\|\tilde{\mathbf{v}}_s\|^2 \simeq 1 + \left({\mathbf{u}_1}^H E_p \mathbf{v}_1 + {\mathbf{v}_1}^H {E_p}^H \mathbf{u}_1\right)/\sigma_1$, whence we get

$$\hat{\mathbf{v}}_s \simeq \left(\mathbf{v}_1 + \frac{{E_p}^H \mathbf{u}_1}{\sigma_1}\right)\left[1 - \frac{1}{2\sigma_1}\left({\mathbf{u}_1}^H E_p \mathbf{v}_1 + {\mathbf{v}_1}^H {E_p}^H \mathbf{u}_1\right)\right]$$

Ignoring terms of order ${E_p}^H E_p$ and simplifying, the MSE in $\hat{\mathbf{v}}_s$ is

$$\begin{aligned}\hat{\mathbf{v}}_s - \mathbf{v}_1 &\simeq \frac{{E_p}^H \mathbf{u}_1}{\sigma_1} - \frac{1}{2\sigma_1}\left({\mathbf{u}_1}^H E_p \mathbf{v}_1 + {\mathbf{v}_1}^H {E_p}^H \mathbf{u}_1\right)\mathbf{v}_1 \\ \|\hat{\mathbf{v}}_s - \mathbf{v}_1\|^2 &= \frac{\|{E_p}^H \mathbf{u}_1\|^2}{\sigma_1^2} - \frac{1}{4\sigma_1^2}\left|{\mathbf{u}_1}^H E_p \mathbf{v}_1 + {\mathbf{v}_1}^H {E_p}^H \mathbf{u}_1\right|^2 \end{aligned} \tag{2.22}$$

Taking expectation and simplifying the above expression using lemma 2, we get

$$\mathbf{E}\left\{\|\hat{\mathbf{v}}_s - \mathbf{v}_1\|^2\right\} = \frac{1}{2\gamma_p\sigma_1^2}\left(2t-1\right). \tag{2.23}$$

Interestingly, the above expression is the Cramer-Rao lower bound (CRB) for the estimation of $\mathbf{v}_1$ assuming perfect knowledge of $\mathbf{u}_1$, as shown below.

Theorem 1. *The error given in (2.23) is the CRB for the estimation of* $\mathbf{v}_1$ *under perfect knowledge of* $\mathbf{u}_1$.

Proof. From (2.35), the effective SNR for estimation of $\mathbf{v}_1$ is $\gamma_s = \gamma_p\sigma_1^2$. From the results derived for the CRB with constrained parameters [42, 50], since $\tilde{X_p}\tilde{X_p}^H = I_t/\gamma_p$, the estimation error in $\mathbf{v}_1$ is proportional to the number of parameters, which equals $2t-1$ as $\mathbf{v}_1$ is a t-dimensional complex vector with one constraint ($\|\mathbf{v}_1\| = 1$). The estimation error is given by

$$\begin{aligned}\mathbf{E}\left\{\|\hat{\mathbf{v}}_s - \mathbf{v}_1\|^2\right\} &= \frac{1}{2\gamma_s}\{\text{Num. Parameters}\} \\ &= \frac{1}{2\gamma_p\sigma_1^2}\left(2t-1\right),\end{aligned} \tag{2.24}$$

which agrees with the ML error derived in (2.23). □

2.4.2 Received SNR with Perfect $\hat{\mathbf{u}}_s$

We start with the expression for the channel gain when using $\hat{\mathbf{u}}_s$ and $\hat{\mathbf{v}}_s$ as the transmit and receive beamforming vectors. When we have perfect knowledge of $\mathbf{u}_1$ at the receiver, $\hat{\mathbf{u}}_s = \mathbf{u}_1$ and $\hat{\mathbf{v}}_s = \tilde{\mathbf{v}}_s/\|\tilde{\mathbf{v}}_s\|$, where $\tilde{\mathbf{v}}_s = \mathbf{v}_1 + E_u\mathbf{u}_1$ and $E_u \triangleq {E_p}^H/\sigma_1$. The power amplification with perfect knowledge of $\mathbf{u}_1$, denoted by $\rho_u \triangleq \mathbf{E}\left\{\left|{\mathbf{u}_1}^H H\hat{\mathbf{v}}_s\right|^2\right\} = \mathbf{E}\left\{\frac{\left|{\mathbf{u}_1}^H H\tilde{\mathbf{v}}_s\right|^2}{\|\tilde{\mathbf{v}}_s\|^2}\right\}$. As shown in the Appendix 2.8.2, this can be simplified to

$$\rho_u = \sigma_1^2 - \frac{t-1}{\gamma_p}. \tag{2.25}$$

Finally, the received SNR is given by $P_D\rho_u$, as before. Comparing the above expression with the power amplification with CLSE (2.16), we see that when $r = t$, even in the best case of a spatially single-dimensional channel $\rho_c = \rho_p - \frac{2}{\gamma_p}(t-1) <$

ρ_u. Next, when $r = 1$, CLSE and CFSB techniques perform exactly the same: $\rho_c = \rho_u = \sigma_1^2 - \frac{t-1}{\gamma_p}$ since $\mathbf{u}_1 = 1$ (that is, no receive beamforming is needed). Thus, if perfect knowledge of $\mathbf{u}_1$ is available at the receiver, CFSB is guaranteed to perform as well as CLSE, regardless of the training symbol SNR.

2.4.3 MSE in $\hat{\mathbf{v}}_s$ with Noise-Free Training

We now present analysis to compute the second term in (2.21), the MSE in $\hat{\mathbf{v}}_s$ solely due to the use of the erroneous vector $\hat{\mathbf{u}}_s$ in (2.7), and hence let $\eta_p = 0$, or $Y_p = HX_p$. As in Section 2.3-2.3.3, we can express $\hat{\mathbf{u}}_s$ as a linear combination $\mathbf{c}$ of the columns of U as $\hat{\mathbf{u}}_s = U\mathbf{c}$. We slightly abuse notation from Section 2.4-2.4.1 and redefine $\tilde{\mathbf{v}}_s$ as $\tilde{\mathbf{v}}_s \triangleq X_p Y_p^H \hat{\mathbf{u}}_s/\gamma_p = V\Sigma\mathbf{c}$. Hence,

$$\|\tilde{\mathbf{v}}_s\|^2 = \mathbf{c}^H \Sigma^2 \mathbf{c}.$$

Thus, from (2.7), we have, $\hat{\mathbf{v}}_s = V\tilde{\mathbf{c}}$, where $\tilde{\mathbf{c}} = \frac{\Sigma\mathbf{c}}{\sqrt{c^H\Sigma^2\mathbf{c}}}$. From lemma (3),

$$\|\hat{\mathbf{v}}_s - \mathbf{v}_1\|^2 = 2\left(1 - \mathrm{Re}(\tilde{c}_1)\right). \tag{2.26}$$

Let $\mathbf{c} = [1 + \Delta c_1, \Delta c_2, \ldots, \Delta c_r]^T$. Then, as shown in the Appendix 2.8.3, $\tilde{c}_1$, the first element of $\tilde{\mathbf{c}}$, is given by

$$\tilde{c}_1 \simeq 1 - \frac{1}{2}\sum_{i=2}^{r} \frac{\sigma_i^2}{\sigma_1^2} |\Delta c_i|^2, \tag{2.27}$$

and hence $\|\hat{\mathbf{v}}_s - \mathbf{v}_1\|^2 = \sum_{i=2}^{r} \frac{\sigma_i^2}{\sigma_1^2} |\Delta c_i|^2$. Let γ_d be defined as $\gamma_d \triangleq NP_D/t$. Then, from Appendix 2.8.3, $\mathbf{E}\left\{|\Delta c_i|^2\right\}$ is given by

$$\mathbf{E}\left\{|\Delta c_i|^2\right\} = \frac{1}{\left(\sigma_1^2 - \sigma_i^2\right)^2}\left(\frac{\sigma_1^2\sigma_i^2}{N} + \frac{\sigma_i^2 + \sigma_1^2}{\gamma_d} + \frac{N}{\gamma_d^2}\right). \tag{2.28}$$

Substituting, we get the final expression for the MSE as

$$\mathbf{E}\left\{\|\hat{\mathbf{v}}_s - \mathbf{v}_1\|^2\right\} = \sum_{i=2}^{r} \frac{\sigma_i^2}{\sigma_1^2\left(\sigma_1^2 - \sigma_i^2\right)^2}\left(\frac{\sigma_1^2\sigma_i^2}{N} + \frac{\sigma_i^2 + \sigma_1^2}{\gamma_d} + \frac{N}{\gamma_d^2}\right). \tag{2.29}$$

Note that the above expression decreases as $O(1/N)$ since γ_d depends linearly on N, and therefore the MSE asymptotically approaches the bound in (2.23).

2.4.4 Received SNR with Noise-Free Training

The power amplification with noise-free training, denoted ρ_w, is given by $\rho_w = \left|\hat{\mathbf{u}}_s^H H \hat{\mathbf{v}}_s\right|^2$. We also have $\hat{\mathbf{u}}_s = U\mathbf{c}$ and $\hat{\mathbf{v}}_s = V\tilde{\mathbf{c}}$, where $\tilde{\mathbf{c}} = \frac{\Sigma \mathbf{c}}{\sqrt{\mathbf{c}^H \Sigma^2 \mathbf{c}}}$. Then, $\hat{\mathbf{u}}_s^H H \hat{\mathbf{v}}_s = \mathbf{c}^H \Sigma \tilde{\mathbf{c}} = \sqrt{\mathbf{c}^H \Sigma^2 \mathbf{c}}$, and thus

$$\begin{aligned} \rho_w &= \mathbf{c}^H \Sigma^2 \mathbf{c} = \sigma_1^2 \left(1 + \Delta c_1^*\right)\left(1 + \Delta c_1\right) + \sum_{i=2}^{r} \sigma_i^2 \Delta c_i^* \Delta c_i \\ &\simeq \sigma_1^2 \left(1 + \Delta c_1 + \Delta c_1^*\right) + \sum_{i=2}^{r} \sigma_i^2 \left|\Delta c_i\right|^2. \end{aligned}$$

Substituting for Δc_1 from (2.9) and Δc_i from (2.28), we obtain the power amplification with noise-free training as

$$\rho_w = \sigma_1^2 - \sum_{i=2}^{r} \frac{1}{\left(\sigma_1^2 - \sigma_i^2\right)} \left(\frac{\sigma_1^2 \sigma_i^2}{N} + \frac{\sigma_1^2 + \sigma_i^2}{\gamma_d} + \frac{N}{\gamma_d^2} \right). \tag{2.30}$$

As before, the received SNR is given by $P_D \rho_w$. Note that ρ_w approaches $\rho_p = \sigma_1^2$ for large values of length N and SNR γ_d.

2.4.5 Semi-blind Estimation: Summary

Recall that $\gamma_p = LP_T/t$ and $\gamma_d = NP_D/t$. The final expressions for the MSE in $\hat{\mathbf{v}}_s$ and the power amplification, from (2.21), are:

$$\begin{aligned} \mathbf{E}\left\{\left\|\mathbf{v}_1 - \hat{\mathbf{v}}_s\right\|^2\right\} &= \frac{(2t-1)}{2\gamma_p \sigma_1^2} \\ &+ \sum_{i=2}^{r} \frac{\sigma_i^2}{\sigma_1^2 \left(\sigma_1^2 - \sigma_i^2\right)^2} \left(\frac{\sigma_1^2 \sigma_i^2}{N} + \frac{\sigma_i^2 + \sigma_1^2}{\gamma_d} + \frac{N}{\gamma_d^2} \right), \end{aligned} \tag{2.31}$$

$$\begin{aligned} \rho_s &= \sigma_1^2 - \frac{t-1}{\gamma_p} \\ &- \sum_{i=2}^{r} \frac{1}{\left(\sigma_1^2 - \sigma_i^2\right)} \left(\frac{\sigma_1^2 \sigma_i^2}{N} + \frac{\sigma_1^2 + \sigma_i^2}{\gamma_d} + \frac{N}{\gamma_d^2} \right). \end{aligned} \tag{2.32}$$

The SER with semi-blind estimation is given by $P_M\left(\rho_s\right)$, with $P_M\left(\cdot\right)$ as in (2.18).

2.5 Comparison of CLSE and Semi-blind Schemes

In order to compare the CFSB and CLSE techniques, one needs to account for the performance of the white data versus beamformed data, an issue we address now. Generic comparison of the semi-blind and conventional schemes for any arbitrary system configuration is difficult, so we consider an example to illustrate the trade-offs involved. We consider the 2×2 system with the Alamouti scheme [51] employed for white data transmission, and with uncoded 4-QAM symbol transmission. The choice of the Alamouti scheme enables us to present a fair comparison of the two estimation algorithms since it has an effective data rate of 1 bit per channel use, the same as that of MRT. Additionally, it is possible to employ a simple receiver structure, which makes the performance analysis tractable.

Let the beamformed data and the white data be statistically independent, and a zero-forcing receiver based on the conventional estimate of the channel (2.3) be used to detect the white data symbols. In Appendix 2.8.4, we derive the average SNR of this system as

$$\rho_w = \frac{\left(\|H\|_F^4 + \frac{\|H\|_F^2}{\gamma_p} \right) P_x}{\frac{\|H\|_F^2}{\gamma_p} P_x + \|H\|_F^2 + \frac{2r}{\gamma_p}}, \tag{2.33}$$

where $\| \cdot \|_F$ is the Frobenius norm, P_x is the per-symbol transmit power and $\gamma_p = LP_T/t$ as defined before. From (2.33), we can also obtain the symbol error rate performance of the Alamouti coded white data by using (2.18) with $\rho_c P_D$ replaced by ρ_w. The resulting expression can be numerically averaged over the pdf of $\|H\|_F^2$, which is Gamma distributed with $2rt$ degrees of freedom, to obtain the SER. The analysis of the beamformed data with the CFSB estimation when the Alamouti scheme is employed to transmit spatially white data remains largely the same as that presented in the previous section, where we had assumed that X_d satisfies $\mathbf{E}\left\{X_d X_d^H\right\} = \gamma_d I_t$. With Alamouti white-data transmission, we have that $X_d X_d^H = \gamma_d I_t$, which causes the E_χ term to drop out in (2.39) of Appendix 2.8.3.

2.5.1 Performance of a 2×2 System with CLSE and CFSB

In order to get a more concrete feel for the expressions obtained in the preceding, let us consider a 2×2 system with $L = 2$, $N = 8$, $P_D = 6$dB and 110 total symbols per frame, i.e., 2 training symbols, 8 white data symbols and 100 beamformed data symbols in the semi-blind case, and 2 training symbols and 108 beamformed data symbols in the conventional case. The average channel power gain ρ versus training symbol SNR (P_T), obtained under different CSI and signal transmission conditions are shown in Fig. 2.3. When the receiver has perfect channel knowledge (labelled perfect $\mathbf{u}_1$, $\mathbf{v}_1$), the average power gain ρ is $\mathbf{E}\{\sigma_1^2\} = 5.5$dB, independent of the training symbol SNR. The ρ with CLSE as well as the semi-blind techniques asymptotically tend to this gain of 5.5dB as the SNR becomes large, since the loss due to estimation error becomes negligible. The channel power gain with only white (Alamouti) data transmission asymptotically approaches 3dB (the gain per symbol of the 2×2 system with Alamouti encoding).

The channel power gain at any P_T is given by (2.33), which is validated in Fig. 2.3 through simulation. Observe that at a given training SNR, there is a loss of approximately $P_a = -3$dB in terms of the channel gain performance for the Alamouti scheme compared to the beamforming with conventional estimation. The results of the channel power gain obtained by employing the CFSB technique with $N = 8$ Alamouti-coded data symbols are shown in Fig. 2.3, and show the improved performance of CFSB. By transmitting a few ($N = 8$) Alamouti-coded symbols, the CFSB scheme obtains a better estimate of $\mathbf{v}_1$, thereby gaining about $P_{sb} = 0.8$dB per symbol over the CLSE scheme, at a training SNR of 2dB.

If the frame length is 110 symbols, we have L_d =100 beamformed data symbols in the semi-blind case and $L_d + N = 108$ beamformed data symbols in the conventional case. Using the beamforming vectors estimated by the CFSB algorithm, we then have a net power gain ρ_g given by $\rho_g = \frac{L_d+N}{L_d/P_{sb}+N/P_a}$, or about 0.4dB per frame. Thus, this simple example shows that CFSB estimation can

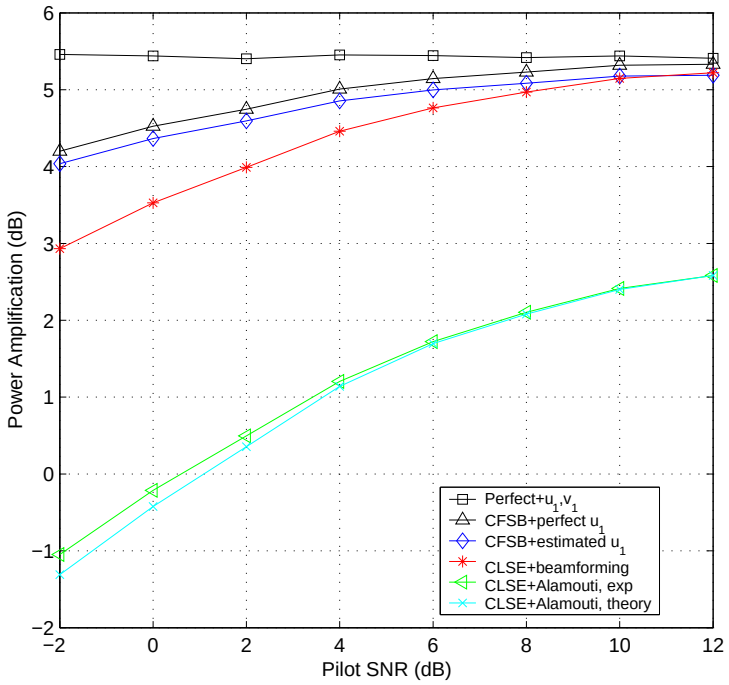


Figure 2.3: Average channel gain of a $t = r = 2$ MIMO channel with $L = 2$, $N = 8$ and $P_D = 6$dB, for the CLSE and beamforming, CFSB and beamforming (with and without knowledge of $\mathbf{u}_1$), CLSE and white data (Alamouti-coded), and perfect beamforming at transmitter and receiver. Also plotted is the theoretical result for the performance of Alamouti-coded data with channel estimation error

.

potentially offer an overall better performance compared to the CLSE. Although we have considered uncoded modulation here, in more practical situations a channel code will be used with interleaving both between the white and beamformed symbols as well as across multiple frames. In this case, burst errors can be avoided and the errors in the white data symbols corrected. Furthermore, the performance of the white data symbols can also be improved by employing an MMSE receiver or other more advanced multi-user detectors rather than the zero-forcing receiver, leading to additional improvements in the CFSB technique.

2.5.2 Discussion

We are now in a position to discuss the merits of the conventional estimation and the semi-blind estimation. Clearly, the CLSE enjoys the advantages of being simple and easy to implement. As with any semi-blind technique, CFSB being a second-order method requires the channel to be relatively slowly time-varying. If not, the CLSE can still estimate the channel quickly from a few training symbols, whereas the CFSB may not be able to converge to an accurate estimate of $\mathbf{u}_1$ from the second order statistics computed using just a few received vectors. Another disadvantage of the CFSB is that it requires the implementation of two separate receivers, one for detecting the white data and the other for the beamformed data. However, the CFSB estimation could outperform the CLSE in channels where the loss due to the transmission of spatially-white data is not too great, i.e., in full column-rank channels. Given the parameters N, L, P_T and P_D, the theory developed in this chapter can be used to decide if the CFSB technique would offer any performance benefits versus the CLSE technique. If the CFSB technique is to perform comparably or better than the CLSE, two things need to be satisfied:

1. The *estimation* performance of CLSE and CFSB should be comparable, i.e., the number of white data symbols N and the data power P_D should be large enough to ensure that the estimate $\hat{\mathbf{u}}_s$ is accurate, so that the resulting $\hat{\mathbf{v}}_s$ can perform comparably to the conventional estimate. For example, since the channel gain with semi-blind estimation is given by (2.31), N should be chosen to be of the same order as γ_d; and both N and γ_d should be of the order higher than γ_p. With such a choice, the $(t-1)/\gamma_p$ term will dominate the SNR loss in the CFSB, thus enabling the beamformed data with CFSB estimation to outperform the beamformed data with CLSE.

2. The block length should be sufficiently long to ensure that after sending $L+N$ symbols, there is sufficient room to send as many beamformed symbols as is

necessary for the CFSB technique to be able to make up for the performance lost during the white data transmission. In the above example, after having obtained the appropriate value of N, one can use (2.33) to determine the loss due to the white data symbols (for the $t = 2$ case), and then finally determine whether the block length is long enough for the CFSB to be able to outperform the CLSE method.

In Section 2.6, we demonstrate through additional simulations that the CFSB technique does offer performance benefits relative to the CLSE, for an appropriately designed system.

2.5.3 Semi-blind Estimation: Limitations and Alternative Solutions

The CFSB algorithm requires a sufficiently large number of spatially-white data (N) to guarantee a near perfect estimate of $\mathbf{u}_1$ and this error cannot be overcome by increasing the white-data SNR. It is therefore desirable to find an estimation scheme that performs at least as well as the CLSE algorithm, regardless of the value of N and L. Formal fusion of the estimates obtained from the CLSE and CFSB techniques is difficult, hence we adopt an intuitive approach and consider a simple weighted linear combination of the estimated beamforming vectors as follows:

$$\hat{\mathbf{u}}_1 = \frac{\beta_{\mathbf{u}}\gamma_p\hat{\mathbf{u}}_c + \gamma_d\hat{\mathbf{u}}_s}{\|\beta_{\mathbf{u}}\gamma_p\hat{\mathbf{u}}_c + \gamma_d\hat{\mathbf{u}}_s\|_2} \quad , \quad \hat{\mathbf{v}}_1 = \frac{\beta_{\mathbf{v}}\gamma_p\hat{\mathbf{v}}_c + \gamma_d\hat{\mathbf{v}}_s}{\|\beta_{\mathbf{v}}\gamma_p\hat{\mathbf{v}}_c + \gamma_d\hat{\mathbf{v}}_s\|_2}. \tag{2.34}$$

The above estimates will be referred to as the *linear combination semi-blind* (LCSB) estimates. The weights $\gamma_p = LP_T/t$ and $\gamma_d = NP_D/t$ are a measure of the accuracy of the vectors estimated from the CLSE and CFSB schemes respectively. The scaling factor of $\beta_{\mathbf{u}}$ and $\beta_{\mathbf{v}}$ is introduced because $\hat{\mathbf{u}}_c$ obtained from known training symbols is more reliable than the blind estimate $\hat{\mathbf{u}}_s$ when $L = N$ and $\gamma_p = \gamma_d$. In our simulations, for $t = r = 4$, the choice $\beta_{\mathbf{u}} = \beta_{\mathbf{v}} = 4$ was found to perform well. Analysis of the impact of $\beta_{\mathbf{u}}$ and $\beta_{\mathbf{v}}$ is a topic for future research.

2.6 Simulation Results

In this section, we present simulation results to illustrate the performance of the different estimation schemes. The simulation setup consists of a Rayleigh flat fading channel with 4 transmit antennas and 4 receive antennas ($t = r = 4$). The data (and training) are drawn from a 16-QAM constellation. 10,000 random instantiations of the channel were used in the averaging.

Measuring the error between singular vectors

In the simulations, $\mathbf{v}_1$ and $\hat{\mathbf{v}}_1$ are obtained by computing the SVD of two different matrices H and $\hat{H}$ respectively. However, the SVD involves an unknown phase factor, that is, if $\mathbf{v}_1$ is a singular vector, so is $\mathbf{v}_1 e^{j\phi}$ for any $\phi \in (-\pi, \pi]$. Hence, for computational consistency in measuring the MSE in $\mathbf{v}_1$, we use the following dephased norm in our simulations, similar to [8]: $\|\mathbf{v}_1 - \hat{\mathbf{v}}_1\|_{DN}^2 \triangleq 2\left(1 - \left|\mathbf{v}_1{}^H \hat{\mathbf{v}}_1\right|\right)$; which satisfies $\|\mathbf{v}_1 - \hat{\mathbf{v}}_1\|_{DN}^2 = \min_{\phi \in (-\pi,\pi]} \|\mathbf{v}_1 - \hat{\mathbf{v}}_1 e^{j\phi}\|^2$. The norm considered in our analysis is implicitly consistent with the above dephased norm. For example, the norm in (2.13) is the same as the dephased norm, since the perturbation term Δd_1 is real (as noted in Section 2.3-2.3.1). Also, for small additive perturbations, it can easily be shown that (for example) in (2.22), the dephased norm reduces to the Euclidean norm.

Experiment 1

In this experiment, we compute the MSE of conventional estimation and the MSE of the semi-blind estimation with perfect $\mathbf{u}_1$, which serves as a benchmark for the performance of the proposed semi-blind scheme. Fig.2.4 shows the MSE in $\hat{\mathbf{v}}_1$ versus L, for two different values of pilot SNR (or γ_p), with perfect $\mathbf{u}_1$. CFSB performs better than the CLSE technique by about 6dB, in terms of the training symbol SNR for achieving the same MSE in $\hat{\mathbf{v}}_1$. The experimental curves agree well with the theoretical curves from (2.14), (2.23). Also, the results for

the performance of the semi-blind OPML technique proposed in [43] are plotted in Fig. 2.4. In the OPML technique, the channel matrix H is factored into the product of a whitening matrix W $(= U\Sigma)$ and a unitary rotation matrix Q. A blind algorithm is used to estimate W, while the training data is used exclusively to estimate Q. Thus, the OPML technique outperforms the CFSB because it assumes perfect knowledge of the entire U and Σ matrices (and is computationally more expensive). The CFSB technique, on the other hand, only needs an accurate estimate of $\mathbf{u}_1$ from the spatially-white data.

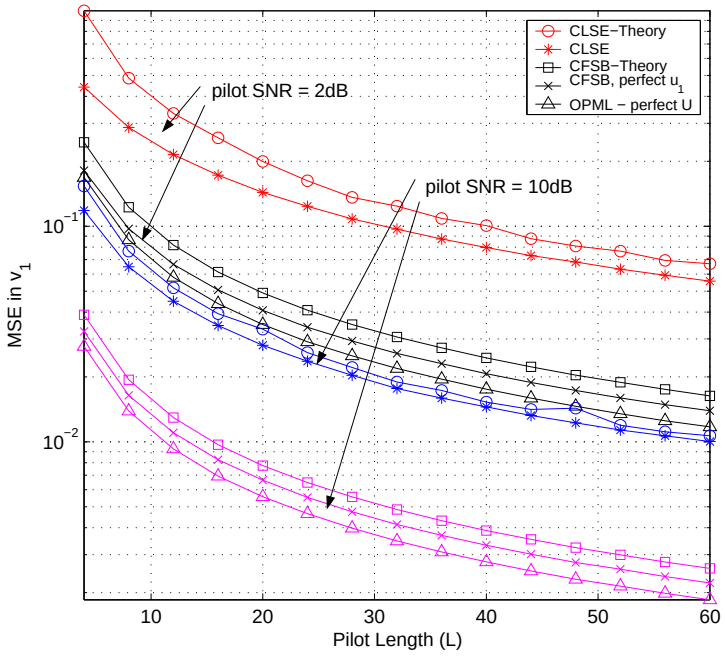


Figure 2.4: MSE in $\mathbf{v}_1$ vs training data length L, for a $t = r = 4$ MIMO system. Curves for CLSE, CFSB and OPML with perfect $\mathbf{u}_1$ are plotted. The top five curves correspond to a training symbol SNR of 2dB, and the bottom five curves 10dB.

Experiment 2

Next, we relax the perfect $\mathbf{u}_1$ assumption. Fig.2.5 shows the SER performance of the CLSE, OPML and the CFSB schemes at two different values of

N, as well as the $N = \infty$ (perfect knowledge of U) case. At $N = 50$ white data symbols, the CLSE technique outperforms the CFSB for $L \geq 24$, as the error in $\mathbf{u}_1$ dominates the error in the semi-blind technique. As white data length increases, the CFSB performs progressively better than the CLSE. Also, in the presence of a finite number (N) of white data, the CFSB marginally outperforms the OPML scheme as CFSB only requires an accurate estimate of the dominant eigenvector $\mathbf{u}_1$ from the white data. In Fig. 2.6, we plot both the theoretical and experimental curves for the CFSB scheme when $N = 100$, as well as the simulation result for the LCSB scheme defined in Section 2.5-2.5.3. The LCSB outperforms the CLSE and the CFSB technique at both $N = 50$ and $N = 100$. Thus, the theory developed in this chapter can be used to compare the performance of CFSB and CLSE techniques for any choice of N and L.

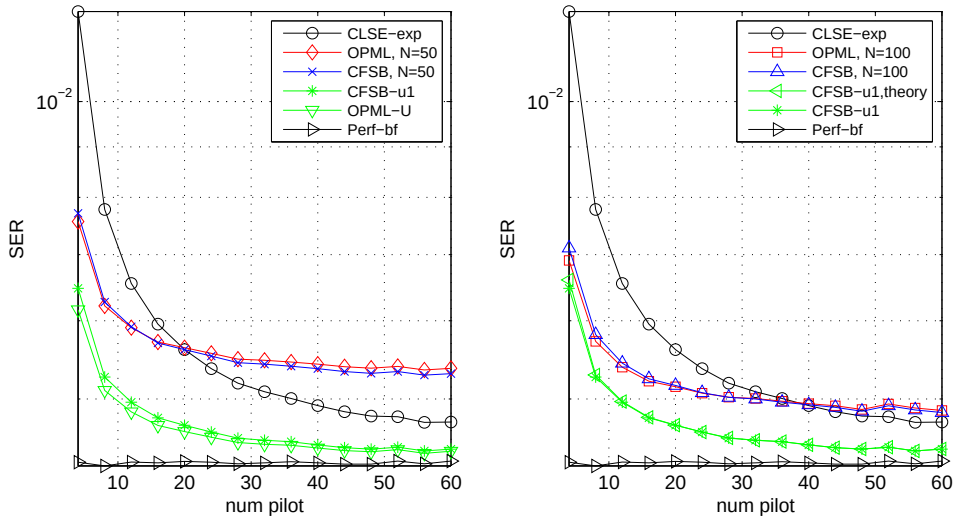


Figure 2.5: SER of beamformed-data vs number of training symbols L, $t = r = 4$ system, for two different values of white-data length N, and data and training symbol SNR fixed at $P_T = P_D = 6$dB. The two competing semi-blind techniques, OPML and CFSB, are plotted. CFSB marginally outperforms OPML for $N = 50$, as it only requires an accurate estimate of $\mathbf{u}_1$ from the blind data.

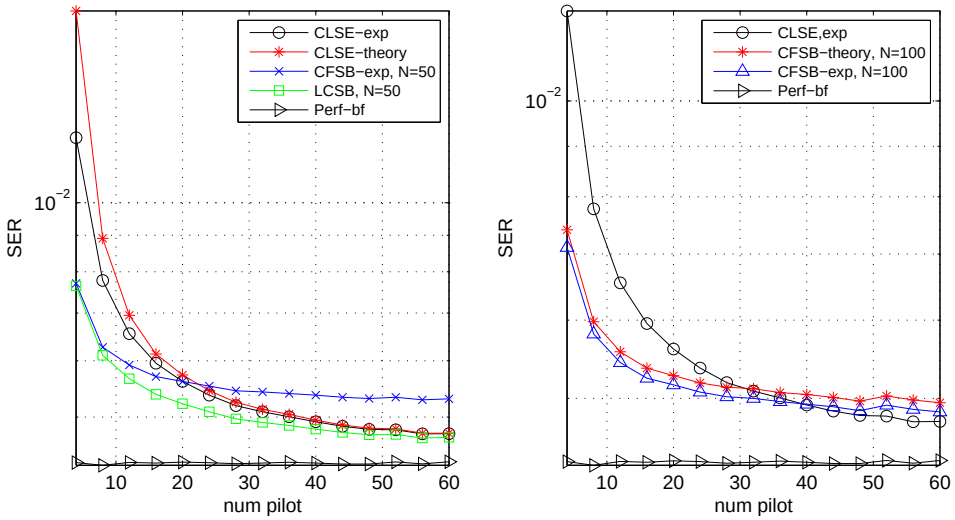


Figure 2.6: SER vs L, $t = r = 4$ system, for two different values of N, and data and training symbol SNR fixed at $P_T = P_D = 6$dB. The theoretical and experimental curves are plotted for the CFSB estimation technique. Also, the LCSB technique outperforms both the conventional (CLSE) and semi-blind (CFSB) techniques.

Experiment 3

Finally, as an example of overall performance comparison, Fig. 2.7 shows the SER performance versus the data SNR of the different estimation schemes for a 2×2 system, with uncoded 4-QAM transmission, $L = 2$ training symbols, $N = 16$ white data symbols (for the semi-blind technique) and a frame size $L_d = 500$ symbols. The parameter values are chosen for illustrative purposes, and as L and P_T increase, the gap between the CLSE and CFSB reduces. From the graph, it is clear that the LCSB scheme outperforms the CLSE scheme in terms of its SER performance, including the effect of white data transmission.

2.7 Conclusion

In this chapter, we have investigated training-only and semi-blind channel estimation for MIMO flat-fading channels with MRT, in terms of the MSE in the

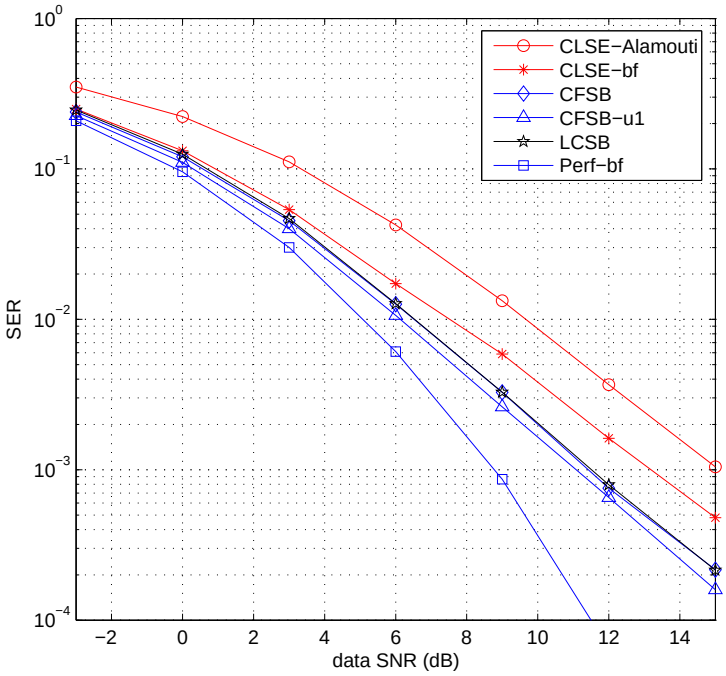


Figure 2.7: SER versus data SNR for the $t = r = 2$ system, with $L = 2, N = 16, \gamma_p = 2dB$. 'CLSE-Alamouti' refers to the performance of the spatially-white data with conventional estimation, 'CLSE-bf' is the performance of the beamformed data with $\hat{\mathbf{v}}_c$, 'CFSB' and 'LCSB' refer to the performance of the corresponding techniques after accounting for the loss due to the white data. 'CFSB-u1' is the performance of CFSB with perfect-$\mathbf{u}_1$, and 'Perf-bf' is the performance with the perfect $\mathbf{u}_1$ and $\mathbf{v}_1$ assumption.

beamforming vector $\mathbf{v}_1$, received SNR and the SER with uncoded M-ary QAM modulation. The CFSB scheme is proposed as a closed-form semi-blind solution for estimating the optimum transmit beamforming vector $\mathbf{v}_1$, and is shown to achieve the CRB with the perfect $\mathbf{u}_1$ assumption. Analytical expressions for the MSE, the channel power gain and the SER performance of both the CLSE and the CFSB estimation schemes are developed, which can be used to compare their performance. A novel LCSB algorithm is proposed, which is shown to outperform both the CFSB and the CLSE schemes over a wide range of training lengths and

SNRs. We have also presented Monte-Carlo simulation results to illustrate the relative performance of the different techniques.

2.8 Appendix

2.8.1 Proof of Lemma 1:

Let $\tilde{Y_p} \triangleq \frac{\mathbf{u_1}^H Y_p}{\sigma_1 \gamma_p}$, $\tilde{X_p} \triangleq \frac{X_p}{\gamma_p}$, and $\tilde{n} \triangleq \frac{\mathbf{u_1}^H \eta_p}{\sigma_1 \gamma_p}$. Then, since the training sequence is orthogonal, $X_p \tilde{X_p}^H = I_t$ holds. Substituting into (2.5), we have

$$\tilde{Y_p} = \mathbf{v_1}^H \tilde{X_p} + \tilde{n}. \tag{2.35}$$

Thus, we seek the estimate of $\mathbf{v}_1$ as the solution to the following least squares problem

$$\hat{\mathbf{v}}_s = \arg \min_{\mathbf{v} \in \mathbb{C}^t,\, \|\mathbf{v}\|=1} \|\tilde{Y_p} - \mathbf{v}^H \tilde{X_p}\|^2. \tag{2.36}$$

Note that

$$\begin{aligned}
&\arg \min_{\mathbf{v_1}:\, \|\mathbf{v_1}\|=1} \|\tilde{Y_p} - \mathbf{v_1}^H \tilde{X_p}\|^2 = \\
&\arg \min_{\mathbf{v_1}:\, \|\mathbf{v_1}\|=1} \left(\tilde{Y_p} \tilde{Y_p}^H + \frac{\|\mathbf{v_1}\|^2}{\gamma_p} - \tilde{Y_p} \tilde{X_p}^H \mathbf{v_1} - \mathbf{v_1}^H \tilde{X_p} \tilde{Y_p}^H \right) \\
&\qquad = \arg \max_{\mathbf{v_1}:\, \|\mathbf{v_1}\|=1} \left(\tilde{Y_p} \tilde{X_p}^H \mathbf{v_1} + \mathbf{v_1}^H \tilde{X_p} \tilde{Y_p}^H \right).
\end{aligned}$$

The $\mathbf{v}_1$ that maximizes the above expression is readily found to be $\hat{\mathbf{v}}_1 = \tilde{X_p} \tilde{Y_p}^H / \|\tilde{X_p} \tilde{Y_p}^H\|$. Substituting for $\tilde{X_p}$ and $\tilde{Y_p}$, the desired result is obtained.

2.8.2 Received SNR with perfect $\hat{\mathbf{u}}_s$

Here, we derive the expression in (2.25). For notational simplicity, define $x \triangleq \mathbf{v_1}^H E_u \mathbf{u_1}$ and $y \triangleq \mathbf{u_1}^H E_u^H E_u \mathbf{u_1}$. Then, we have

$$\begin{aligned}
\rho_u &= \mathbf{E} \left\{ \frac{\sigma_1^2 (1+x)(1+x^*)}{1+x+x^*+y} \right\} \\
&\simeq \sigma_1^2 \mathbf{E} \left\{ (1+x)(1+x^*) \right. \\
&\qquad \left. \left(1 - (x+x^*+y) + (x+x^*+y)^2\right) \right\} \\
&\simeq \sigma_1^2 \left(1 + \mathbf{E}\left\{x x^* - y\right\}\right),
\end{aligned} \tag{2.37}$$

where x^* is the complex conjugate of x. Also, $\mathbf{E}\{xx^*\} = \frac{\sigma_p^2}{\sigma_1^2} = \frac{1}{\gamma_p \sigma_1^2}$, and $\mathbf{E}\{y\} = \mathbf{E}\left\{\mathbf{u}_1{}^H E_u^H E_u \mathbf{u}_1\right\} = \frac{t}{\gamma_p \sigma_1^2}$. Thus, the power amplification for perfect $\mathbf{u}_1$ is given by $\rho_u = \sigma_1^2 - \frac{t-1}{\gamma_p}$.

2.8.3 Proof for equations (2.27) and (2.28)

In order to derive an expression for $\tilde{c}_1$, we write $\mathbf{c} = [1 + \Delta c_1, \Delta c_2, \ldots, \Delta c_t]^T$ as a perturbation of $[1, 0, \ldots, 0]^T$. Since $\tilde{\mathbf{c}} = \frac{\Sigma \mathbf{c}}{\sqrt{c^H \Sigma^2 \mathbf{c}}}$, equating components, we have

$$\begin{aligned}
\tilde{c}_1 &= \frac{\sigma_1 (1 + \Delta c_1)}{\sqrt{\sigma_1^2 |1 + \Delta c_1|^2 + \sum_{i=2}^{r} \sigma_i^2 |\Delta c_i|^2}} \\
&\simeq (1 + \Delta c_1) \left[1 - \frac{1}{2}\left(2\Delta c_1 + \sum_{i=2}^{r} \frac{\sigma_i^2}{\sigma_1^2} |\Delta c_i|^2\right)\right] \\
&\simeq 1 - \frac{1}{2} \sum_{i=2}^{r} \frac{\sigma_i^2}{\sigma_1^2} |\Delta c_i|^2 .
\end{aligned}$$

Substituting in (2.26), we get

$$\|\mathbf{v}_1 - \hat{\mathbf{v}}_s\|^2 = \sum_{i=2}^{r} \frac{\sigma_i^2}{\sigma_1^2} |\Delta c_i|^2 . \tag{2.38}$$

It now remains to compute Δc_i. Recall that $\hat{\mathbf{u}}_s$ is computed from the SVD in (2.4). Stacking the transmitted and received data vectors into matrices $X_d \in \mathbb{C}^{t \times N}$ and $Y_d \in \mathbb{C}^{r \times N}$ and the noise vectors into $\eta_d \in \mathbb{C}^{r \times N}$, with appropriate scaling we can rewrite (2.4) as

$$\hat{U} \hat{\Sigma}^2 \hat{U}^H = H H^H + E_s,$$

where,

$$E_s \triangleq H E_\chi H^H + H E_{\chi\eta} + E_{\chi\eta}^H H^H + E_\eta,$$

and $E_\chi \triangleq \frac{1}{\gamma_d}\left(X_d X_d^H - \gamma_d I_t\right)$, $E_{\chi\eta} \triangleq \frac{X_d \eta_d^H}{\gamma_d}$, $E_\eta \triangleq \frac{1}{\gamma_d}\left(\eta_d \eta_d^H - N I_r\right)$, and finally $\gamma_d = \frac{N P_D}{t}$, as before.

Observe that, since the white data X_d and AWGN are mutually independent, the elements of E_χ, $E_{\chi\eta}$ and E_η are pairwise uncorrelated. Also,

$\mathbf{E}\left\{|E_\chi(i,j)|^2\right\} = \left(\frac{P_D}{t}\right)^2 / \left(N\left(\frac{P_D}{t}\right)^2\right) = 1/N,$
$\mathbf{E}\left\{|E_{\chi\eta}(i,j)|^2\right\} = \left(\frac{P_D}{t}\right) / \left(N\left(\frac{P_D}{t}\right)^2\right) = 1/\gamma_d$, and
$\mathbf{E}\left\{|E_\eta(i,j)|^2\right\} = 1/\left(N\left(\frac{P_D}{t}\right)^2\right) = N/\gamma_d^2.$
Thus, from the first order perturbation analysis (2.8), $\Delta c_i = \frac{\mathbf{u}_i^H E_s \mathbf{u}_1}{\sigma_1^2 - \sigma_i^2}$, and therefore

$$\begin{aligned}\mathbf{E}\left\{|\Delta c_i|^2\right\} = \frac{1}{\left(\sigma_1^2-\sigma_i^2\right)^2}\Big(&\mathbf{E}\left\{\left|\mathbf{u}_i^H H E_\chi H^H \mathbf{u}_1\right|^2\right\} \\ &+\mathbf{E}\left\{\left|\mathbf{u}_i^H H E_{\chi\eta}\mathbf{u}_1\right|^2\right\} + \mathbf{E}\left\{\left|\mathbf{u}_i^H E_{\chi\eta}^H H^H \mathbf{u}_1\right|^2\right\} \\ &+ \mathbf{E}\left\{\left|\mathbf{u}_i^H E_\eta \mathbf{u}_1\right|^2\right\}\Big). \end{aligned} \tag{2.39}$$

Simplifying the different components in the above expression, we have
$\mathbf{E}\left\{\left|\mathbf{u}_i^H H E_\chi H^H \mathbf{u}_1\right|^2\right\} = \sigma_1^2\sigma_i^2/N,\ \mathbf{E}\left\{\left|\mathbf{u}_i^H E_\eta \mathbf{u}_1\right|^2\right\} = N/\gamma_d^2$ and
$\mathbf{E}\left\{\left|\mathbf{u}_i^H H E_{\chi\eta}\mathbf{u}_1\right|^2\right\} = \sigma_i^2/\gamma_d$. Substituting into (2.39), we get (2.28).

2.8.4 Performance of Alamouti Space-Time Coded Data with Conventional Estimation

In this section, we determine the performance of Alamouti space-time coded data for a general $r \times 2$ matrix channel with estimation error and a zero-forcing receiver. Similar results for other specific cases can be found in [52], [53]. Denote the $r \times 2$ channel matrix H in terms of its columns as $H = [\mathbf{h}_1,\ \mathbf{h}_2]$. Also, let the $2 \times L$ orthogonal training symbol matrix X_p be defined in terms of its rows as $X_p^T = [X_{p1}^T,\ X_{p2}^T]^T$. Thus, from (2.3), the channel is estimated conventionally as

$$\begin{aligned}\hat{H}_c &= \frac{1}{\gamma_p}\left[Y_p X_{p1}^H, Y_p X_{p2}^H\right] \\ \left[\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2\right] &= \left[\mathbf{h}_1 + \frac{\eta_p X_{p1}^H}{\gamma_p}, \mathbf{h}_2 + \frac{\eta_p X_{p2}^H}{\gamma_p}\right] \end{aligned} \tag{2.40}$$

The effective channel with Alamouti-coded data transmission can be represented by stacking two consecutively received $r \times 1$ vectors $\mathbf{y}_1$ and $\mathbf{y}_2^*$ vertically as follows

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2^* \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 \\ -\mathbf{h}_2^* & \mathbf{h}_1^* \end{bmatrix} \begin{bmatrix} x_1 \\ x_2^* \end{bmatrix} + \begin{bmatrix} \mathbf{n}_{w1} \\ \mathbf{n}_{w2}^* \end{bmatrix}, \tag{2.41}$$

where $\mathbf{n}_{wi},\ i = 1, 2$ is the AWGN affecting the white data symbols. When a zero-forcing receiver based on the estimated channel is employed, the received vectors are decoded using $\left[\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2\right]$ as

$$\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2^* \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{h}}_1^H & -\hat{\mathbf{h}}_2^T \\ \hat{\mathbf{h}}_2^H & \hat{\mathbf{h}}_1^T \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2^* \end{bmatrix}. \tag{2.42}$$

It is clear from symmetry that the performance of $\hat{x}_1$ and $\hat{x}_2$ will be the same; hence, we can focus on determining the SER performance of $\hat{x}_1$. Now, $\hat{x}_1$ contains three components, the signal component coming from x_1, and a leakage term coming from the symbol x_2 and the noise term coming from the white noise term $\mathbf{n}_w$ as follows

$$\begin{aligned} \hat{x}_1 = \underbrace{\left(\hat{\mathbf{h}}_1^H \mathbf{h}_1 + \mathbf{h}_2^H \hat{\mathbf{h}}_2\right)}_{\xi_{x1}} x_1 + \underbrace{\left(\hat{\mathbf{h}}_1^H \mathbf{h}_2 - \mathbf{h}_1^H \hat{\mathbf{h}}_2\right)}_{\xi_{x2}} x_2^* \\ + \underbrace{\hat{\mathbf{h}}_1^H \mathbf{n}_{w1} - \mathbf{n}_{w2}^H \hat{\mathbf{h}}_2}_{\xi_n} \end{aligned} \tag{2.43}$$

The coefficient of the x_1 term, denoted ξ_{x1} is

$$\begin{aligned} \xi_{x1} &= \left(\mathbf{h}_1 + \frac{\eta_p X_{p1}^H}{\gamma_p}\right)^H \mathbf{h}_1 + \mathbf{h}_2^H \left(\mathbf{h}_2 + \frac{\eta_p X_{p2}^H}{\gamma_p}\right), \\ &= \|H\|_F^2 + \frac{X_{p1}\eta_p^H \mathbf{h}_1 + \mathbf{h}_2^H \eta_p X_{p2}^H}{\gamma_p}. \end{aligned} \tag{2.44}$$

From the above equation, it is clear that the performance of the x_1 symbol is dependent on the training noise instantiation η_p. However, we can consider the average power gain, averaged over the training noise, as follows

$$\begin{aligned} \mathbf{E}\left\{|\xi_{x1}|^2\right\} &= \|H\|_F^4 + \frac{1}{\gamma_p^2}\mathbf{E}\left\{X_{p1}\eta_p^H \mathbf{h}_1 \mathbf{h}_1^H \eta_p X_{p1}^H \right. \\ &+ \left. \mathbf{h}_2^H \eta_p X_{p2}^H X_{p2} \eta_p^H \mathbf{h}_2\right\}, \qquad (2.45) \\ &= \|H\|_F^4 + \frac{1}{\gamma_p^2}\left(\gamma_p \|\mathbf{h}_1\|^2 + \gamma_p \|\mathbf{h}_2\|^2\right), \\ &= \|H\|_F^4 + \frac{\|H\|_F^2}{\gamma_p}, \qquad (2.46) \end{aligned}$$

where, in (2.45), the cross terms disappear since the noise η_p is zero-mean and due to the orthogonality of the training X_p. Similarly, the coefficient of the x_2^* term, denoted ξ_{x2}, can be simplified as

$$\xi_{x2} = \frac{X_{p1}\eta_p^H \mathbf{h}_2 - \mathbf{h}_1^H \eta_p X_{p2}^H}{\gamma_p}. \tag{2.47}$$

We will assume for simplicity that the x_2 term is an additive white Gaussian noise impairing the estimation of x_1, i.e., we do not perform joint detection. This noise term is independent of the AWGN component $\mathbf{n}_w$. Similar to the coefficient of x_1, we can consider the average power gain of the x_2 term, which can be obtained after a little manipulation as

$$\mathbf{E}\left\{|\xi_{x2}|^2\right\} = \frac{\|H\|_F^2}{\gamma_p}. \tag{2.48}$$

Finally, the noise term, denoted ξ_n, is

$$\begin{aligned} \xi_n &= \mathbf{h}_1^H \mathbf{n}_{w1} - \mathbf{n}_{w2}^H \mathbf{h}_2 \\ &+ \frac{X_{p1}\eta_p^H \mathbf{n}_{w1} - \mathbf{n}_{w2}^H \eta_p X_{p2}^H}{\gamma_p}, \end{aligned} \tag{2.49}$$

from which we can obtain the noise power as

$$\mathbf{E}\left\{|\xi_n|^2\right\} = \|H\|_F^2 + \frac{2r}{\gamma_p}. \tag{2.50}$$

Thus, the SNR for detection of a white data symbol is given by

$$\rho_w = \frac{\left(\|H\|_F^4 + \frac{\|H\|_F^2}{\gamma_p}\right) P_x}{\frac{\|H\|_F^2}{\gamma_p} P_x + \|H\|_F^2 + \frac{2r}{\gamma_p}}. \tag{2.51}$$

2.8.5 Other Useful Lemmas:

In this section, we present three useful lemmas without proof.

Lemma 4. *Let $X_p \in \mathbb{C}^{t\times L}$ be an orthogonal set of vectors (i.e., $X_p X_p^H = \gamma_p I_t$), and let $\eta_p \in \mathbb{C}^{r\times L}$ contain* i.i.d. *ZMCSCG entries with mean $\mu = 0$ and variance $\sigma_n^2 = 1$. Then, the elements of $E_p = X_p \eta_p^H$ are uncorrelated, and the variance of each element of E_p is $\sigma_p^2 = \gamma_p$.*

Lemma 5. *A transformation of E_p (defined in lemma 4) by any orthogonal matrix $V \in \mathbb{C}^{t\times t}$ (i.e., $VV^H = V^HV = I_t$) to get $\hat{E} = VE_p$, leaves the second order statistics of E_p unaltered, that is,*

$$\begin{aligned}
\mathbf{E}\left\{E(i,j)\right\} &= \mathbf{E}\left\{\hat{E}(i,j)\right\} = 0 \\
\mathbf{E}\left\{E(i,j)E^*(k,l)\right\} &= \mathbf{E}\left\{\hat{E}(i,j)\hat{E}^*(k,l)\right\} \\
&= \sigma_p^2\delta\left(i-k, j-l\right), \quad \forall\, i,j,k,l,
\end{aligned}$$

where $\delta\left(p,q\right) = 1$ when $p = q = 0$, and 0 otherwise.

Lemma 6. *If the random vector $X_p \in \mathbb{C}^{t\times L}$ has zero-mean circularly symmetric* i.i.d. *entries, then so does $\mathbf{v}^HX_p$, where $\mathbf{v} \in \mathbb{C}^{t\times 1}$. Further, if $\mathbf{v}$ satisfies $\|\mathbf{v}\| = 1$, then the variance of an element of X_p is the same as that of $\mathbf{v}^HX_p$.*

Acknowledgement

The text of this chapter, in part, is a reprint of the material as it appears in C. R. Murthy, A. K. Jagannatham, and B. D. Rao, "Training-only and semi-blind channel estimation for maximum ratio transmission based MIMO systems," *IEEE Transactions on Sig. Proc.*, vol. 54, pp. 2546–2558, July 2006.

3 Quantization Methods for Equal Gain Transmission With Finite Rate Feedback

3.1 Introduction

Per-antenna power constraints, rather than total power constraints, are more practically meaningful in the design of transmit beamforming vectors in multiple input, single output (MISO) systems as they impose much less fidelity requirements on the the transmit RF power amplifiers. The performance achievable by multiple antenna systems is dependent on the channel state information available at the transmitter (CSIT). When the channel state information (CSI) is known perfectly at the transmitter, *beamforming* is the optimum method of transmission to maximize the channel capacity under a per-antenna power constraint as well as under a total power constraint [7], [8]. Additionally, CSIT-based transmission such as beamforming offers the benefits of lower complexity receivers and better system throughput in a multiuser environment. However, in practical communication systems, due to feedback channel bit rate constraints, the CSI has to be quantized with a finite number of bits, and only the quantized CSI can be made available to the transmitter.

Maximum ratio transmission (MRT) [7] is the optimum beamforming vector for maximizing the capacity with a total power constraint. In general, an MRT

beamforming vector is denoted $\mathbf{v} \in \mathbb{C}^t$, where t is the number of transmit antennas, and the constant total power constraint can be expressed as $\|\mathbf{v}\|_2^2 = t$, where $\|\mathbf{v}\|_2$ denotes Euclidean norm (or L_2-norm) of $\mathbf{v}$. It can be shown that if $\mathbf{v}$ is an optimum MRT vector, so is $\mathbf{v}\exp(j\theta)$ for any angle θ [25]. Hence, we can further constrain (for example) the first element of $\mathbf{v}$ to be real without loss of performance. Therefore, the MRT beamforming vector contains t complex parameters and two real constraints, i.e, it can be completely described by $(t-1)$ complex parameters, which need to be made available at the transmitter to enable optimum MRT. The problem of quantizing the MRT beamforming vector in a vector quantization (VQ) framework has been considered in [35]. Several results quantifying the performance of finite-rate feedback systems with MRT have also appeared in [20] - [36]. Other works on the performance of quantized-feedback based multiple antenna systems include [32] and [33], where the authors use quantized CSI obtained from a feedback link to determine a weighting matrix or precoding matrix to improve the performance of an orthogonal space-time block (OSTB) code. The effect of imperfect CSI feedback has also been addressed in [13] - [19] where the authors consider transmit optimization with either channel mean feedback or covariance feedback.

Equal gain transmission (EGT) (see, e.g., [8] and the references therein), the subject of this chapter, is the optimum beamforming vector for maximizing the capacity of MISO flat fading systems with an equal power per-antenna constraint. This choice of beamforming vector also maximizes the expected received SNR given the constraints. In general, an EGT beamforming vector is given by $\mathbf{w} = [1, \exp(j\theta_2), \exp(j\theta_3), \ldots, \exp(j\theta_t)]^T$, where θ_i denotes the phase rotation applied at antenna element i. Thus, the EGT vector contains exactly $(t-1)$ real parameters that need to be made available at the transmitter to enable optimum EGT, which is half the number of parameters needed to enable optimum MRT. The quantization of these phase angles is the main focus of this chapter. We use some of the analytical tools first developed in [27] and extend the results to the

case of beamforming under a per-antenna power constraint.

The problem of quantizing the EGT beamforming vector was proposed in [37]. The solution proposed there uniformly quantized the phase angles, i.e., scalar quantization (SQ). This method has low complexity, but is sub-optimal compared to VQ based techniques, even in the case of i.i.d. Rayleigh fading channels. One of the goals of this chapter, therefore, is to quantify the performance difference between SQ and VQ for the case of i.i.d. Rayleigh fading channels. In [8], the authors considered the problem of designing quantized EGT beamforming vectors for the case of i.i.d. Rayleigh fading channels and proposed a design criterion based on Grassmannian beamforming. While the proposed criterion guarantees full diversity order, specific algorithms to generate the codebook based on the Grassmannian beamforming criterion were not developed, and it is not clear whether the codebook of quantized beamforming vectors is optimum in terms of channel capacity. Another recent work is [38], where the authors derive an random search based algorithm to design equal gain codebooks. In [39], the authors employ the minimum value of the maximum magnitude of the inner product between any two code vectors as the performance metric, and derive several families of codebooks for which, imposing the per-antenna power constraint rather than the total power constraint results in no loss.

In this chapter, we first consider VQ, and develop an algorithm based on the generalized Lloyd algorithm that converges to an optimum codebook[1] that maximizes the capacity, while imposing no restrictions on the channel statistics, in Section 3.3. We analyze the performance of this algorithm for the case of i.i.d. Rayleigh fading channels, and show that the capacity loss with quantized EGT (Q-EGT) drops off with the number of feedback bits B as $\frac{2(t-1)}{t+1}2^{-\frac{2B}{t-1}}$, in Section 3.4 and 3.5. The generality of the analytical tools developed is further demonstrated by deriving an expression for the outage probability with VQ-based feedback in

[1]The Lloyd algorithm is a standard algorithm in source coding literature, and it only guarantees local optimality of the codebook. Therefore, by 'optimum codebook', we strictly mean that the codebook is locally optimum, although we will not always make this distinction.

Section 3.6. Next, we consider the performance of SQ in Section 3.7 for the case of i.i.d. Rayleigh fading channels, and obtain analytical expressions for the capacity loss performance. The theoretical expressions are "high-rate results", i.e., they progressively become more accurate as the number of feedback bits B gets large. We compare the performance of SQ with VQ and see that while both achieve the same rate of convergence to the capacity with perfect feedback, a finite gap exists. This gap is seen to converge to a constant in terms of bits per dimension, as the number of transmit antennas t is increased. Monte-Carlo simulations confirm the accuracy of the analysis in Section 3.8.

We use the following notation. Matrices are denoted by capital letters such as R and vectors by bold-face letters such as $\mathbf{h}$. A^H denotes the conjugate transpose of A. The i-th component of a vector $\mathbf{h}$ is denoted h_i. $\|\mathbf{v}\|_p$ denotes the L_p-norm of the vector $\mathbf{v}$. The vector formed by the phase angles of $\mathbf{h}$ is denoted $\angle\mathbf{h}$. The expectation operator is denoted as $\mathbf{E}\{\cdot\}$. Finally, $\underline{1}$ and $\underline{0}$ are vectors of ones and zeros respectively, and the dimension will be clear from the context.

3.2 Preliminaries

Consider a MISO system with t antennas at the transmitter. Under the block flat-fading model, the multiple-antenna channel is represented by the channel vector $\mathbf{h} \in \mathbb{C}^t$ which remains constant for the duration of a block, and changes independently according to some statistical distribution from block to block. For simplicity of notation, therefore, we can omit the time index and express the relationship between the channel input $\mathbf{x} \in \mathbb{C}^t$ and the channel output $y \in \mathbb{C}$ as

$$y = \mathbf{h}^H \mathbf{x} + \eta, \tag{3.1}$$

where $\eta \in \mathbb{C}$ is the zero mean Gaussian noise at the receiver. The CSI $\mathbf{h}$ is assumed to be known perfectly at the receiver, and partially at the transmitter through a limited-rate feedback channel. When *beamforming* is employed at the transmitter, the data symbol $s \in \mathbb{C}$ is multiplied by a beamforming vector $\mathbf{w}$ to get

$\mathbf{x}$ as $\mathbf{x} = \mathbf{w}s$. Throughout this chapter, we will assume that $\mathbf{w}$ satisfies a constant 2-norm constraint $\|\mathbf{w}\|_2^2 \leq t$, to ensure that the total transmitted power in the symbol s, given by $P_s \triangleq \mathbf{E}\left\{|s|^2\right\}$, is not amplified. A quantized beamforming vector codebook $\mathcal{C} \triangleq \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\}$ is known to both the receiver and the transmitter, where $N = 2^B$. Based on the knowledge of $\mathbf{h}$, the receiver selects the best beamforming vector $\mathbf{w}_i \in \mathcal{C}$ and sends the corresponding index i to the transmitter through the feedback channel. We assume that the feedback channel has no delay and is error free, in order to focus on the effect of quantizing the CSI.

Given the channel instantiation $\mathbf{h}$ and the beamforming vector $\mathbf{w}$ employed at the transmitter, the mutual information is given by

$$I\left(\mathbf{h}, \mathbf{w}, P_s\right) \triangleq \log\left(1 + \left|\mathbf{h}^H\mathbf{w}\right|^2 P_s\right) \tag{3.2}$$

Therefore, given a channel instantiation $\mathbf{h}$ and with *perfect CSIT*, the optimization criterion for choosing $\mathbf{w}$ is given by

$$\mathbf{w}_o = \arg\max_{\mathbf{w}\in\mathcal{S}} \left|\mathbf{h}^H\mathbf{w}\right|^2, \tag{3.3}$$

where $\mathcal{S}$ denotes the constraint set to which $\mathbf{w}$ belongs. If $\mathbf{w}$ has a total power constraint, the constraint set is given by $\mathcal{S}_T \triangleq \{\mathbf{w} : \|\mathbf{w}\|_2 = \sqrt{t}\}$ and with a per-antenna power constraint, the constraint set is given by $\mathcal{S}_P \triangleq \{\mathbf{w} : \|\mathbf{w}\|_\infty \leq 1\}$. When we have an equal power constraint (i.e., EGT), we have $\mathcal{S} = \{\mathbf{w} : \mathbf{w} = [\exp(j\theta_1), \exp(j\theta_2), \ldots, \exp(j\theta_t)]^T\}$. Since the objective function $\left|\mathbf{w}^H\mathbf{h}\right|^2$ is unchanged by an multiplying $\mathbf{w}$ by an overall phase $\exp(j\theta)$, without loss of generality we can assume $\theta_1 = 0$, and consider
$\mathcal{S}_E \triangleq \{\mathbf{w} : \mathbf{w} = [1, \exp(j\theta_2), \exp(j\theta_3), \ldots, \exp(j\theta_t)]^T\}$. It is shown in Lemma 8 of the Appendix 3.9.1 that the solution to (3.3) with $\mathcal{S} = \mathcal{S}_P$ is the same as the solution with the smaller constraint set $\mathcal{S} = \mathcal{S}_E$.

3.3 Vector Quantization: Codebook Design

It is well known that with perfect knowledge of $\mathbf{h}$ at the transmitter, the optimum EGT vector is $\mathbf{w}_o = \exp(j\angle\mathbf{h})$, and the corresponding power gain is

$\left|\mathbf{h}^H\mathbf{w}\right|^2 = \|\mathbf{h}\|_1^2$ [8]. In this chapter, we use the capacity loss, i.e., the difference between the capacity with beamforming using $\mathbf{w}_o$ and the capacity with beamforming using the Q-EGT vector $\mathbf{w} \in \mathcal{C}$. The mutual information with quantized feedback (i.e., $\mathbf{w}$), is given by (3.2); and with perfect feedback (i.e., $\mathbf{w}_o$) it is

$$I(\mathbf{h}, \mathbf{w}_o, P_s) \;=\; \log(1 + \|\mathbf{h}\|_1^2 P_s), \tag{3.4}$$

and note that since $\mathbf{w}_o = \exp(j\angle\mathbf{h})$, the right hand side of the above equation does not explicitly depend on $\mathbf{w}_o$. Thus, the loss in mutual information, $I_L(\mathbf{h}, \mathbf{w}) \triangleq I(\mathbf{h}, \mathbf{w}_o, P_s) - I(\mathbf{h}, \mathbf{w}, P_s)$ can be simplified to obtain (for notational simplicity, we drop the dependence of I_L on P_s)

$$I_L(\mathbf{h}, \mathbf{w}) \;=\; -\log\left(1 - \frac{\|\mathbf{h}\|_1^2 P_s}{1 + \|\mathbf{h}\|_1^2 P_s}\left(1 - \frac{\left|\mathbf{h}^H\mathbf{w}\right|^2}{\|\mathbf{h}\|_1^2}\right)\right) \tag{3.5}$$

Note that $\mathbf{w}$ is implicitly a function of $\mathbf{h}$ since it is the vector in the codebook $\mathcal{C}$ that minimizes the loss in mutual information for every instantiation of $\mathbf{h}$. The goal of the quantizer is thus to minimize the loss in mutual information averaged over the channel statistics.

Using the first order approximation $-\log(1-x) \simeq x$ for the logarithm in (3.5), which is valid when $P_s \ll 1$, or when the number of feedback bits B is large, we see that a good criterion for designing the beamforming vector codebook is a maximum *mean-squared weighted inner product* (MSwIP) criterion:

$$\max_{\mathcal{Q}(\cdot)} \mathbf{E}\left\{\frac{\|\mathbf{h}\|_1^2 P_s}{1 + \|\mathbf{h}\|_1^2 P_s}\frac{\left|\mathbf{h}^H\mathbf{w}\right|^2}{\|\mathbf{h}\|_1^2}\right\}, \tag{3.6}$$

where $\mathbf{w} = \mathcal{Q}(\mathbf{h})$ is the quantized beamforming vector from the codebook $\mathcal{C}$. This is different from the criteria derived for the quantized MRT case [27] in that we have the one-norm $\|\mathbf{h}\|_1$ instead of the two-norm $\|\mathbf{h}\|_2$ and P_s is the transmit power per antenna instead of the total transmit power. Therefore, the VQ codebook design algorithm is considerably different, as we describe in the following.

The generalized Lloyd algorithm [54] can be used to generate codebooks that maximize the MSwIP, and consists of the following two steps:

- Nearest neighborhood condition *(NNC)*: Given the code vectors $\{\mathbf{w}_i; i = 1, \ldots, N\}$, the optimum partition region (Voronoi cell) $\mathcal{R}_i$ of the code vector indexed by i satisfies

$$\mathcal{R}_i = \{\mathbf{h} : \left|\varsigma \mathbf{h}^H \mathbf{w}_i\right|^2 \geq \left|\varsigma \mathbf{h}^H \mathbf{w}_j\right|^2, \forall\, j \neq i\}, \tag{3.7}$$

  where $\varsigma \triangleq \sqrt{\frac{P_s}{1+\|\mathbf{h}\|_1^2 P_s}}$. Note that ς does not impact the partitioning.

- Centroid condition *(CC)*: Given the partition regions $\{\mathcal{R}_i; i = 1, \ldots, N\}$, the optimum code-vectors $\mathbf{w}_i$ are chosen to satisfy, for $i = 1, \ldots, N$,

$$\begin{aligned} \mathbf{w}_i &= \arg\max_{\mathbf{w}\in\mathcal{S}_E} \mathbf{E}\left\{\left|\varsigma \mathbf{h}^H \mathbf{w}\right|^2 | \mathbf{h} \in \mathcal{R}_i\right\} && (3.8)\\ &= \arg\max_{\mathbf{w}\in\mathcal{S}_E} \mathbf{w}^H \mathbf{E}\left\{\varsigma^2 \mathbf{h}\mathbf{h}^H | \mathbf{h} \in \mathcal{R}_i\right\} \mathbf{w} && (3.9)\\ &= \text{(principal EGT vector of) } \mathbf{E}\left\{\varsigma^2 \mathbf{h}\mathbf{h}^H | \mathbf{h} \in \mathcal{R}_i\right\} && (3.10) \end{aligned}$$

The above two conditions are iterated till the MSwIP converges. In general, this would be implemented by using a sufficiently large number of channel realizations, and replacing the statistical correlation matrix $\mathbf{E}\left\{\varsigma^2 \mathbf{h}\mathbf{h}^H | \mathbf{h} \in \mathcal{R}_i\right\}$ by the sample average correlation matrix. In the CC step, given the quantization regions, we set the centroid to be the principal EGT vector of the conditional correlation matrix, i.e., a vector whose entries are of unit magnitude, that maximizes the expected MSwIP for that region. For computing $\mathbf{w}_i$, it is convenient to solve (3.9) over the larger region S_P instead of S_E (Lemma 8, Appendix 3.9.1). This to reduces the computational complexity to that of a standard convex optimization problem, since S_P is a convex space and $\mathbf{w}^H \mathbf{E}\left\{\varsigma^2 \mathbf{h}\mathbf{h}^H | \mathbf{h} \in \mathcal{R}_i\right\} \mathbf{w}$ is a quadratic (convex) function. The principal EGT vector can now be computed using any off-the-shelf gradient-based optimization routine (such as a Newton search). It is particularly efficient to regard the problem as an unconstrained optimization problem over the the phase angles θ_i of $\mathbf{w}$. In the Appendix 3.9.2 we derive expressions for the gradient and Hessian needed for the Newton algorithm. Moreover, the optimum solution is unique up to multiplication by a constant phase angle.

Note that the above VQ design algorithm depends on the transmitted power per antenna P_s due to the presence of the factor ς^2 in the CC step. This implies that the VQ design algorithm must be repeated and a new codebook generated every time the transmit power changes. This can be avoided by considering one of two simpler design methods. In the high-SNR regime, i.e., when P_s is large, the factor ς^2 can be approximated by $\|\mathbf{h}\|_1^{-2}$, and hence P_s drops out of the algorithm. In the low-SNR regime, i.e., when P_s is small, $P_s\|\mathbf{h}\|_1^2$ in the denominator is small compared to 1, and hence $\varsigma^2 \simeq P_s$, therefore, the factor P_s is simply an overall scaling on the conditional correlation matrix of (3.9), which can be dropped without affecting the algorithm. These two algorithms will be referred to the high-SNR VQ and the low-SNR VQ, and are identical to the algorithm described above except that ς^2 is replaced by $\|\mathbf{h}\|_1^{-2}$ and 1, respectively.

Beamforming Vector Selection (Encoding): For a given codebook $\mathcal{C} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\}$, regardless of the performance metric used, it is easy to see that the receiver encodes as follows:

$$\hat{\mathbf{w}} = \mathcal{Q}(\mathbf{h}) = \arg\max_{\mathbf{w}_i \in \mathcal{C}} \left|\mathbf{h}^H \mathbf{w}_i\right| \tag{3.11}$$

Note that the ς^2 and $\|\mathbf{h}\|_1^2$ terms in (3.6) are in fact superfluous to the encoding, and have been dropped. By this encoding scheme, the space of channel instantiations $\{\mathbf{h} : \mathbf{h} \in \mathbb{C}^t\}$ is partitioned into $\{\bar{\mathcal{R}}_i,\ i = 1, \ldots, N\}$, where

$$\bar{\mathcal{R}}_i = \{\mathbf{h} \in \mathbb{C}^t : \left|\mathbf{h}^H \mathbf{w}_i\right| \geq \left|\mathbf{h}^H \mathbf{w}_j\right|,\ \forall\ j \neq i\} \tag{3.12}$$

3.4 Capacity Loss with VQ-Based Feedback

In this section, we will derive analytical expressions for the performance of Q-EGT for the case of an i.i.d. Rayleigh flat-fading channel with zero mean unit variance complex Gaussian distributed entries. Specifically, we want to analytically characterize the capacity loss incurred due to the quantization of the EGT beamforming vector by a finite number of bits.

Recall from the previous section that $\mathcal{Q}(\mathbf{h}) = \mathbf{w}_i \ \forall \ \mathbf{h} \in \bar{\mathcal{R}}_i$. Define $\alpha \triangleq \sqrt{\frac{\|\mathbf{h}\|_1^2 P_s}{1+\|\mathbf{h}\|_1^2 P_s}}$, and $\xi_i \triangleq 1 - \frac{\left|\mathbf{h}^H \mathbf{w}_i\right|^2}{\|\mathbf{h}\|_1^2}$. Then, the expectation of (3.5), which is the capacity loss, can be expressed as

$$C_L = \sum_{i=1}^{N} \text{Prob}\left(\mathbf{h} \in \bar{\mathcal{R}}_i\right) \mathbf{E}_{\mathbf{h} \in \bar{\mathcal{R}}_i} \left\{ -\log\left(1 - \alpha^2 \xi_i\right) \right\}, \tag{3.13}$$

Deriving a closed-form expression of the above capacity loss is difficult for several reasons. First, the conditional density of ξ_i given $\mathbf{h} \in \bar{\mathcal{R}}_i$, from the optimum encoding in (3.11), is difficult to obtain since the quantization cells defined by (3.12) for $1 \leq i \leq N$ have complicated boundaries defined by neighboring code vectors and could have all different shapes. This geometrical complexity makes analytical expressions for the density of ξ_i intractable. Another factor adding to the intractability of the problem is the dependence of the random variables α and ξ_i, which is difficult to characterize. Using suitable approximations, we now show how to obtain closed-form expressions that closely approximate the capacity loss.

Let us initially consider the high-SNR case, i.e., P_s is large, and hence $\alpha \simeq 1$ holds in (3.13). Under the *high-resolution* approximation, i.e., when B is large, ξ_i is close to 0, and we can use the first-order Taylor series expansion $-\log(1-x) \simeq x$ to get

$$C_L \simeq \sum_{i=1}^{N} \text{Prob}\left(\mathbf{h} \in \bar{\mathcal{R}}_i\right) \mathbf{E}_{\mathbf{h} \in \bar{\mathcal{R}}_i} \left\{ \xi_i \right\} \tag{3.14}$$

To evaluate the capacity loss, we still need to find the conditional distribution $f_{\xi_i}(x|\mathbf{h} \in \bar{\mathcal{R}}_i)$ of $\xi_i = 1 - \frac{\left|\mathbf{h}^H \mathbf{w}_i\right|^2}{\|\mathbf{h}\|_1^2}$ given $\mathbf{h} \in \bar{\mathcal{R}}_i = \{\mathbf{h} \in \mathbb{C}^t : \left|\mathbf{h}^H \mathbf{w}_i\right| \geq \left|\mathbf{h}^H \mathbf{w}_j\right| \ \forall \ j \neq i\}$. To do this, we consider the following approximation to the quantization cell.

$$\begin{aligned} \bar{\mathcal{R}}_i &= \{\mathbf{h} \in \mathbb{C}^t : \xi_i \leq \xi_j \ \forall \ j \neq i\} \\ &\simeq \hat{\mathcal{R}}_i \triangleq \left\{\mathbf{h} \in \mathbb{C}^t : \xi_i \leq \delta\right\}; \ \forall \ 1 \leq i \leq N, \end{aligned} \tag{3.15}$$

where $\delta > 0$ is a constant chosen such that $\text{Prob}(\mathbf{h} \in \hat{\mathcal{R}}_i) = 1/N$. Although the approximate quantization cells $\hat{\mathcal{R}}_i$ do not form a set of non-intersecting regions that cover the $\mathbb{C}^t$ space (unlike $\bar{\mathcal{R}}_i$), the shape and dimensions of $\hat{\mathcal{R}}_i$ will asymptotically

approach those of $\bar{\mathcal{R}}_i$ as B increases. Geometrically, (15) corresponds to saying that when N is large, the quantization cell $\bar{\mathcal{R}}_i$ is approximately a "ball" centered at $\hat{\mathbf{w}}_i$, with the "distance" from the center measured as $1 - \frac{\left|\mathbf{h}^H \mathbf{w}_i\right|^2}{\|\mathbf{h}\|_1^2}$. That is, we expect that the most of the channel instantiations that lie in the quantization region $\bar{\mathcal{R}}_i$ will satisfy $\xi_i \leq \delta$ for a fixed δ independent of i. This is similar to [23], where the authors approximate the Voronoi regions by spherical caps centered at the code points to obtain the outage probability of quantized MRT systems. Thus, we now need to evaluate $f_{\xi_i}(x|\mathbf{h} \in \hat{\mathcal{R}}_i) = f_{\xi_i}(x|\xi_i \leq \delta)$, which is given by

$$f_{\xi_i}\left(x|\mathbf{h} \in \hat{\mathcal{R}}_i\right) = \frac{f_{\xi_i}(x)1_{[0,\delta)}(x)}{\text{Prob}(\mathbf{h} \in \hat{\mathcal{R}}_i)} = \frac{f_{\xi_i}(x)1_{[0,\delta)}(x)}{\int_{y=0}^{\delta} f_{\xi_i}(y)dy}, \tag{3.16}$$

for $1 \leq i \leq N$, where $1_A(x)$ is the indicator function with value 1 if $x \in A$ and 0 otherwise, and $f_{\xi_i}(x)$ is the unconditional distribution of ξ_i for a *fixed* beamforming vector $\mathbf{w}_i \in \mathcal{S}_E$. Due to the left invariance property of $\mathbf{h}$, $f_{\xi_i}(x)$ is independent of $\mathbf{w}_i$, i.e., we can use $f_{\xi_i}(x) = f_{\xi_0}(x)$, where $\xi_0 \triangleq 1 - \frac{\left|\mathbf{h}^H \mathbf{w}_0\right|^2}{\|\mathbf{h}\|_1^2}$, and $\mathbf{w}_0 \triangleq \underline{1} = [1, 1, \ldots, 1]^T$, without loss of generality.

Since $\text{Prob}(\mathbf{h} \in \hat{\mathcal{R}}_i) = 1/N$ independent of i, this implies that δ solves

$$\text{Prob}(\mathbf{h} \in \hat{\mathcal{R}}_i) = \text{Prob}(\xi_i \leq \delta) = \int_0^{\delta} f_{\xi_0}(x)dx = \frac{1}{N}. \tag{3.17}$$

Thus, we can find $f_{\xi_i}(x|\mathbf{h} \in \bar{\mathcal{R}}_i)$, the conditional distribution required to compute the capacity loss in (3.14), approximately as $f_{\xi_i}(x|\mathbf{h} \in \hat{\mathcal{R}}_i)$, which is a function independent of i given by

$$f_{\xi_i}\left(x|\mathbf{h} \in \hat{\mathcal{R}}_i\right) \;\; = \;\; 2^B f_{\xi_0}(x)1_{[0,\delta)}(x), \quad 1 \leq i \leq N. \tag{3.18}$$

Since the regions $\hat{\mathcal{R}}_i$ have identical shape and equal probability $1/N$, we can focus on any one quantization cell in determining the capacity loss (3.14), as follows

$$C_L \simeq \mathbf{E}_{\mathbf{h} \in \hat{\mathcal{R}}_1}\{\xi_1\} = \int_0^{\delta} x f_{\xi_1}(x|\mathbf{h} \in \hat{\mathcal{R}}_1)dx, \tag{3.19}$$

where $f_{\xi_1}(x|\mathbf{h} \in \hat{\mathcal{R}}_1)$, the probability density function (PDF) of ξ_1 given $\mathbf{h} \in \hat{\mathcal{R}}_1$, is given by (3.18), which in turn requires us to find $f_{\xi_0}(x)$. We postpone determining $f_{\xi_0}(x)$ in (3.18) till a later subsection; we first show that the above approximations lead to a lower bound on the capacity loss.

3.4.1 Performance Bound Using the Quantization Cell Approximation

The following lemma and theorem first establish a relationship between $f_{\xi_i}(x|\mathbf{h} \in \bar{\mathcal{R}}_i)$ and $f_{\xi_i}(x|\mathbf{h} \in \hat{\mathcal{R}}_i)$ when the codebook satisfies $\text{Prob}(\mathbf{h} \in \bar{\mathcal{R}}_i) = 1/N$ and then state that the approximate capacity loss is a lower bound on the actual capacity loss. We omit the proofs as they are similar to that of an analogous lemma and theorem for the quantized MRT case [55].

Lemma 7. *For $0 \leq x \leq \delta$, we have $f_{\xi_i}(x|\mathbf{h} \in \bar{\mathcal{R}}_i) \leq f_{\xi_i}(x|\mathbf{h} \in \hat{\mathcal{R}}_i)$.*

Theorem 2. *Consider the capacity loss in* (3.13) *when the codebook satisfies Prob$(\mathbf{h} \in \bar{\mathcal{R}}_i) = 1/N$:*

$$C_L = \mathbf{E}_{\mathbf{h} \in \bar{\mathcal{R}}_i} \left\{ -\log\left(1 - \alpha^2 \xi_i\right) \right\}. \tag{3.20}$$

The capacity loss $\hat{C}_L$ obtained by applying the quantization cell approximation $\bar{\mathcal{R}}_i \simeq \hat{\mathcal{R}}_i$ to the above equation is a lower bound *on the actual capacity loss C_L. That is, if we denote*

$$\hat{C}_L = \mathbf{E}_{\mathbf{h} \in \hat{\mathcal{R}}_i} \left\{ -\log\left(1 - \alpha^2 \xi_i\right) \right\}, \tag{3.21}$$

then we have

$$\hat{C}_L \leq C_L. \tag{3.22}$$

3.4.2 Distribution of ξ_0

The only step remaining in obtaining a closed-form expression for the capacity loss of (3.19) is to determine the density function $f_{\xi_0}(x)$ in (3.18), of the random variable $\xi_0 \triangleq 1 - \frac{\left|\mathbf{h}^H \mathbf{w}_0\right|^2}{\|\mathbf{h}\|_1^2}$ for any fixed vector $\mathbf{w}_0$, which we address next. We have been able to obtain a closed-form expression for the distribution only for the $t = 2$ case. For general t, we consider a natural extension of the expression for $t = 2$, and show through simulation that it works extremely well for small ξ_0 (which is the region of interest).

The $t = 2$ Case

For the special case of 2 transmit antennas, the cumulative distribution function (CDF) of ξ_0 has a surprisingly simple form, given by

$$F_{\xi_0}(x) = \sqrt{\frac{x}{2-x}};\ 0 \leq x \leq 1. \tag{3.23}$$

The derivation of this result is provided in Appendix 3.9.3.

General t Case

For a general t, when N is reasonably large, we can look for approximate expressions for the CDF of ξ_0 that closely match the actual CDF when ξ_0 is close to 0. It is found that the probability distribution function is well approximated by

$$F_{\xi_0}(x) \simeq \left(\frac{x}{2-x}\right)^{\frac{t-1}{2}};\ 0 \leq x \leq 1. \tag{3.24}$$

Fig. 3.1 shows the plot of the above expression for the CDF of ξ_0 along with results obtained through computer simulation, which shows that the approximation error is in fact small, especially for small ξ_0. We will therefore use the above expression for the CDF of ξ_0 in deriving the capacity loss expression.

3.5 Evaluating the Capacity Loss With VQ

We are now in a position to compute the capacity loss performance using the above approximations. Recall that the capacity loss is given by (3.13). Under the *high resolution approximation*, we have

$$C_L \simeq \sum_{i=1}^{N} \text{Prob}\left(\mathbf{h} \in \bar{\mathcal{R}}_i\right) \mathbf{E}_{\mathbf{h}\in\bar{\mathcal{R}}_i}\left\{\alpha^2 \xi_i\right\}. \tag{3.25}$$

Here, we do not make the high-SNR assumption $\alpha^2 \simeq 1$ in order to get a more general expression for the capacity loss, however, we will make the approximation that α^2 and ξ_i are uncorrelated. Although not strictly true, in practice it turns out that the error due to this approximation is small, as will

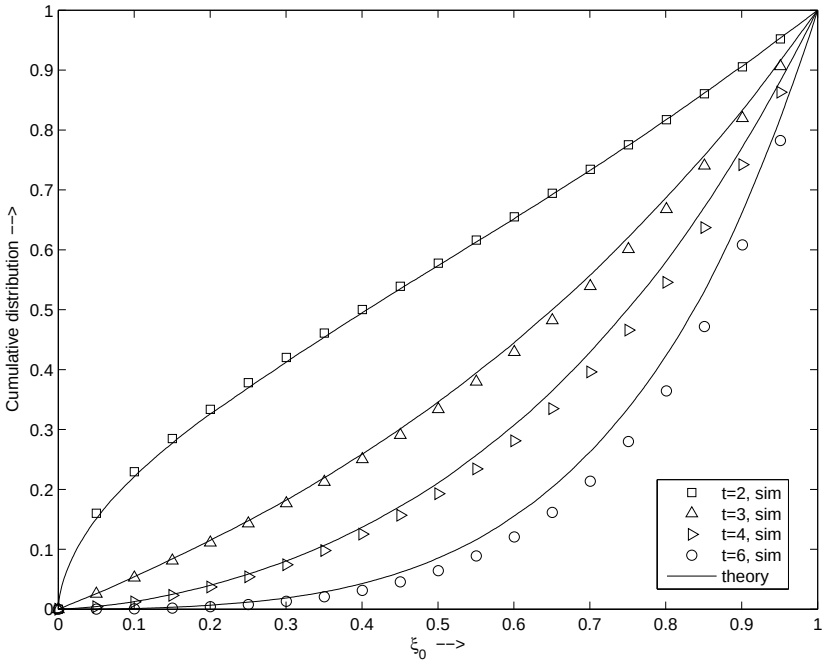


Figure 3.1: Cumulative distribution of ξ_0 for different values of t. Here, 'theory' refers to equation (3.24)

be shown through simulations. Recall that the quantization region $\bar{\mathcal{R}}_i$ does not depend on $\|\mathbf{h}\|_1^2$, hence conditioning on the region does not change its mean, i.e., we have $\mathbf{E}_{\mathbf{h}\in\bar{\mathcal{R}}_i}\{\alpha^2\} = \mathbf{E}\{\alpha^2\}$. Substituting into the above equation and using the quantization cell approximation in (3.15), we have

$$C_L \simeq \mathbf{E}\left\{\alpha^2\right\}\mathbf{E}_{\mathbf{h}\in\hat{\mathcal{R}}_1}\{\xi_1\}. \tag{3.26}$$

Thus, we need to find $\mathbf{E}\{\alpha^2\}$ and $\mathbf{E}_{\mathbf{h}\in\hat{\mathcal{R}}_1}\{\xi_1\}$, which we address next.

3.5.1 Evaluating the Expectation of α^2

In order to find the expectation of $\alpha^2 \triangleq \frac{P_s\|\mathbf{h}\|_1^2}{1+P_s\|\mathbf{h}\|_1^2}$, we need to know the distribution of $\|\mathbf{h}\|_1^2$. Unfortunately, this distribution is not known in closed-form, except for the $t = 2$ transmit antenna case [56]. In this chapter, we employ the moment matching method to approximate the distribution of $\|\mathbf{h}\|_1^2$ by a Gamma distribution, drawing our intuition from the fact that $\|\mathbf{h}\|_2^2$ is Gamma distributed

with $2t$ degrees of freedom. This technique has the advantage of being both simple and sufficiently accurate for our purposes, as will be demonstrated through numerical examples. By direct computation, since $|h_i|$ are i.i.d. Rayleigh distributed,

$$\begin{aligned}\mathbf{E}\left\{\|\mathbf{h}\|_1^2\right\} &= \psi(t) \triangleq t\left(1+\frac{\pi}{4}(t-1)\right),\\ \mathbf{E}\left\{\|\mathbf{h}\|_1^4\right\} &= \phi(t) \triangleq 2t+\frac{3(\pi+2)}{2}t(t-1)+\frac{3\pi}{2}t(t-1)(t-2)\\ &\quad +\frac{\pi^2}{16}t(t-1)(t-2)(t-3).\end{aligned}$$

Matching the moments, we get the PDF of $\|\mathbf{h}\|_1^2$ as

$$f_{\|\mathbf{h}\|_1^2}(x) \simeq \frac{x^{\nu-1}\exp(-x/\beta)}{\beta^\nu\Gamma(\nu)},\ x \geq 0, \tag{3.27}$$

where the scaling and location parameters ν and β are given by

$$\begin{aligned}\nu &\triangleq \frac{\psi^2(t)}{\phi(t)-\psi^2(t)},\\ \beta &\triangleq \frac{\phi(t)}{\psi(t)}-\psi(t).\end{aligned}$$

We can now obtain the expectation of $\alpha^2 \triangleq \frac{P_s\|\mathbf{h}\|_1^2}{1+P_s\|\mathbf{h}\|_1^2}$ as

$$\mathbf{E}\left\{\alpha^2\right\} \simeq \frac{1}{\beta^\nu\Gamma(\nu)}\int_0^\infty \frac{x^\nu\exp(-x/\beta)}{\frac{1}{P_s}+x}dx, \tag{3.28}$$

which can be evaluated using several formulas from [57] as

$$\mathbf{E}\left\{\alpha^2\right\} \simeq \frac{\nu}{(\beta P_s)^\nu}\left[\Gamma(-\nu)\ {}_1F_1\left(1+\nu;1+\nu;\frac{1}{\beta P_s}\right)+\frac{(\beta P_s)^\nu}{\nu}\ {}_1F_1\left(1;1-\nu;\frac{1}{\beta P_s}\right)\right], \tag{3.29}$$

where ${}_1F_1(\cdot;\cdot;\cdot)$ is the confluent hypergeometric function defined in [57]. Note that $\Gamma(-\nu)$ is well-defined because ν is not an integer. Fig. 3.2 shows the plot of the above expression for the mean of α^2, as well as the results obtained through computer simulation. It is clear that the theory agrees well with the simulations. Note also that (3.27) is not the only possible approximation for the PDF of $\|\mathbf{h}\|_1^2$, other approximations have been used in the context of finding the performance of coherent equal gain combining receivers in multipath fading channels. For example, Nakagami [58] approximated the sum of t χ-distributed random variables by

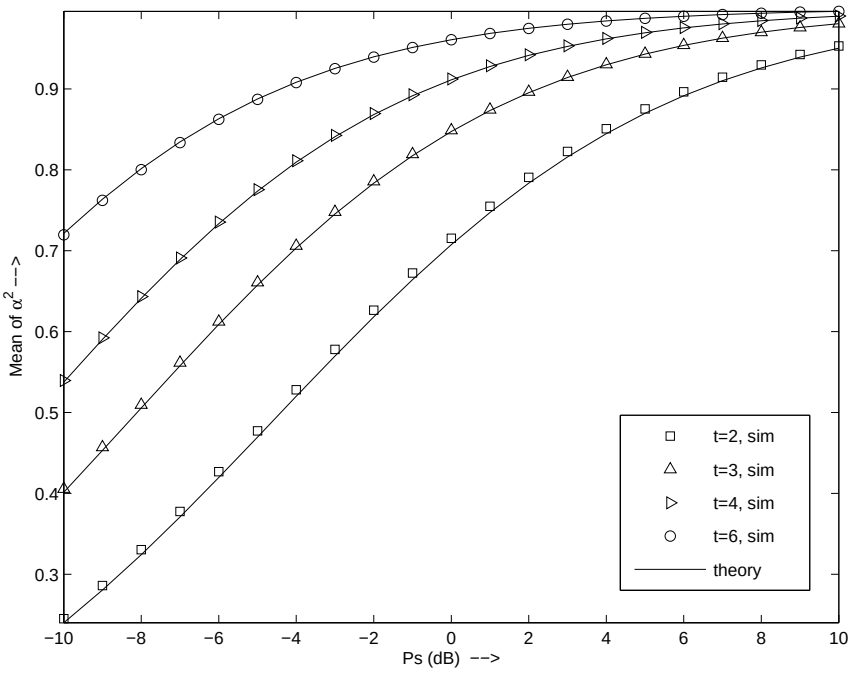


Figure 3.2: Expectation of α^2 as a function of P_s, for different values of t. Here, 'theory' refers to equation (3.29)

another χ random variable to obtain an expression for the PDF of $\|\mathbf{h}\|_1$. Using the transformation $y = x^2$, we obtain the PDF of $\|\mathbf{h}\|_1^2$ as the Gamma distribution

$$f_{\|\mathbf{h}\|_1^2}(x) \simeq \frac{x^{\nu-1}\exp(-x/\beta)}{\beta^{\nu}\Gamma(\nu)},\ x \geq 0, \tag{3.30}$$

where the scaling and location parameters ν and β are now given by

$$\nu \triangleq t, \qquad \beta \triangleq 1 + (t-1)\frac{\pi}{4}.$$

Closed-form expressions for the performance of quantized feedback can be derived using (3.30) also, and the expressions are in fact slightly simpler because $\nu = t$ is an integer, but in practice we have found that the performance obtained from approximation (3.27) is marginally better than that obtained by (3.30) and other similar approximations, and hence, we present the theoretical expressions obtained using the approximate density given by (3.27) only.

3.5.2 Evaluating the Expectation of ξ_1

We now evaluate the second term of (3.26), i.e., $\mathbf{E}_{\mathbf{h}\in\hat{\mathcal{R}}_1}\{\xi_1\}$. Note that $\hat{C}_L \triangleq \mathbf{E}_{\mathbf{h}\in\hat{\mathcal{R}}_1}\{\xi_1\}$ is the capacity loss (3.26) when $\alpha^2 \simeq 1$, i.e., under the *high-SNR, high-resolution* approximation. To specify $f_{\xi_1}(x)$ in (3.18), we first need to determine the value of δ from (3.17) by setting $F_{\xi_0}(\delta) = 1/N$. Substituting for $F_{\xi_0}(\delta)$ from (3.24) and rearranging, we obtain

$$\delta = \frac{2}{1+N^{\frac{2}{t-1}}}. \tag{3.31}$$

We can now compute $\hat{C}_L$ as

$$\hat{C}_L = \int_0^\delta x f_{\xi_1}(x)dx = \delta - N\int_0^\delta \left(\frac{x}{2-x}\right)^{\frac{t-1}{2}} dx. \tag{3.32}$$

Note that, for $0 \le x \le \delta$,

$$\left(\frac{x}{2}\right)^{\frac{t-1}{2}} \le \left(\frac{x}{2-x}\right)^{\frac{t-1}{2}} \le \left(\frac{x}{2-\delta}\right)^{\frac{t-1}{2}}. \tag{3.33}$$

Using the two limits, $\hat{C}_L$ is bounded by

$$\delta - \frac{N}{(2-\delta)^{\frac{t-1}{2}}}\frac{\delta^{\frac{t+1}{2}}}{(t+1)/2} \le \hat{C}_L \le \delta - \frac{N}{2^{\frac{t-1}{2}}}\frac{\delta^{\frac{t+1}{2}}}{(t+1)/2} \tag{3.34}$$

$$2\left(\frac{t-1}{t+1}\right)\left[\frac{1}{1+N^{\frac{2}{t-1}}}\right] \le \hat{C}_L \le \left(\frac{2}{1+N^{\frac{2}{t-1}}}\right)\left[1-\frac{2}{t+1}\left(1+N^{-\frac{2}{t-1}}\right)^{-\frac{t-1}{2}}\right].$$

In particular, when N is large, both the upper and lower bounds are approximately equal and

$$\hat{C}_L \simeq 2\left(\frac{t-1}{t+1}\right)\left[\frac{1}{1+N^{2/(t-1)}}\right] \simeq 2\left(\frac{t-1}{t+1}\right)2^{-\frac{2B}{t-1}}. \tag{3.35}$$

3.5.3 Summary and Discussion

Substituting (3.35) in (3.26), the capacity loss for Q-EGT under the high-resolution approximation is given by

$$C_L \simeq \mathbf{E}\left\{\alpha^2\right\} 2\left(\frac{t-1}{t+1}\right)2^{-\frac{2B}{t-1}}, \tag{3.36}$$

where $\mathbf{E}\{\alpha^2\}$ is given by (3.29), and is $\simeq 1$ in the high-SNR regime. In Section 3.8, we plot the capacity loss versus the number of feedback bits B in Fig. 3.5 and versus the P_s in Fig. 3.6 with vector quantization for an i.i.d. Rayleigh flat fading channel, which shows that the approximation error in the above expression is in fact small.

In [27], the expression for the capacity loss with quantized MRT (Q-MRT) under the high-SNR, high-resolution approximation was obtained as

$$C_L \simeq \left(\frac{t-1}{t}\right) 2^{-\frac{B}{t-1}}. \tag{3.37}$$

Comparing with (3.35), we see that Q-EGT requires roughly half the number of bits to achieve the same capacity loss as Q-MRT. This observation is intuitively satisfying, since the EGT beamforming vector has $t-1$ real parameters, as opposed to the MRT beamforming vector which has $2(t-1)$ real parameters. Also, it should be noted that the C_L in (3.35) is the capacity loss w.r.t *perfect EGT* while the C_L in (3.37) is the capacity loss w.r.t *perfect MRT*. Hence, the above equations should not be construed to mean that Q-EGT might offer a larger capacity than Q-MRT. Finally, there exist certain combinations of N and t for which there is negligible performance loss in imposing the per-antenna power constraint over the less restrictive total power constraint [39], for i.i.d. fading channels.

3.6 Outage Probability with VQ-Based Feedback

The analytical tools developed thus far in this chapter are in fact quite general, although the generalized Lloyd algorithm generates a codebook that specifically maximizes the MSwIP. Using a procedure similar to the one described above, analytical expressions for the outage probability can also be derived. Here, we outline the procedure and present only the final result, as the actual details are straightforward. The outage probability is given by

$$P_{\text{out}}(R, P_s) = \text{Prob}\left(\gamma \left|\mathbf{w}^H \hat{\mathbf{w}}\right|^2 \leq \tau\right), \tag{3.38}$$

where $\gamma \triangleq \|\mathbf{h}\|_1^2$, $\mathbf{w} \triangleq \mathbf{h}/\|\mathbf{h}\|_1$, $\tau \triangleq (2^R-1)/P_s$, and R is the target data rate. Now, assuming that $\left|\mathbf{w}^H\hat{\mathbf{w}}\right|^2$ and γ are approximately independent when B is large, and using the development in the previous section, we get

$$\begin{aligned} P_{\text{out}}(R,P_s) &= \mathbf{E}_\gamma\left\{\text{Prob}\left(\left|\mathbf{w}^H\hat{\mathbf{w}}\right|^2 \le \frac{\tau}{\gamma}\middle|\,\gamma\right)\right\} \\ &\simeq \mathbf{E}_\gamma\left\{1-F_{\xi_i}(1-\tau/\gamma)\right\} \triangleq \hat{P}_{\text{out}}(R,P_s), \end{aligned} \tag{3.39}$$

where $\mathbf{w} \triangleq \mathbf{h}/\|\mathbf{h}\|_1$ and $\tau \triangleq (2^R-1)/P_s$, and R is the target data rate, as before. Also, $F_{\xi_i}(x)$ is the CDF of ξ_i, which is obtained from (3.18) and (3.24) as

$$F_{\xi_i}(x) = \begin{cases} 0 & x<0, \\ N\left(\frac{x}{2-x}\right)^\kappa & 0\le x\le\delta, \\ 1 & x>\delta, \end{cases} \tag{3.40}$$

where $\delta = \frac{2}{1+N^{\frac{1}{\kappa}}}$, and $\kappa \triangleq \frac{t-1}{2}$. After substituting and simplifying, we get the final expression for the outage probability as

$$P_{\text{out}} \approx F_\gamma\left(\frac{\tau}{1-\delta}\right) - N\left(F_\gamma\left(\frac{\tau}{1-\delta}\right)-F_\gamma(\tau)\right)\left(1+\frac{2(1-\delta)}{\delta}\log\left(\frac{2-2\delta}{2-\delta}\right)\right), \tag{3.41}$$

where $F_\gamma(x)$ is the CDF of γ. Exact expressions for the outage probability can be obtained by employing the characteristic function approach derived in [56]. In the sequel, for simplicity, we will use an empirical distribution for γ, obtained through Monte Carlo simulations. In Sec. 3.8, we plot the outage probability versus the total transmit power for $B = 2, 3, 4$ with $t = 3$ transmit antennas in Fig. 3.8, and the results show that the theoretical results agree well with the simulations.

3.7 Scalar Quantization of Parameters

Scalar quantization is a low complexity alternative to VQ, when the computational resources required to run the Lloyd algorithm and to do the quantization are scarce. Here, we present the performance of SQ for the case of i.i.d. Rayleigh fading channels. Let $r_i \triangleq |h_i|$ and $\phi_i \triangleq \angle h_i - \angle h_1$, for $1 \le i \le t$. Also, let

$\theta_1 \triangleq \angle h_1$, $\underline{\phi} \triangleq [\phi_2, \ldots, \phi_t]$, $\mathbf{r} \triangleq [r_1, \ldots r_t]^T$, and $s_{\mathbf{r}} \triangleq \sum_{i=1}^{t} r_i = \|\mathbf{h}\|_1$. Thus, $\mathbf{h}$ can be rewritten as

$$\mathbf{h} = \exp(j\theta_1)\mathrm{diag}([1\, \exp(j\underline{\phi})])\mathbf{r}. \tag{3.42}$$

Note that $\phi_1 = 0$, and as we have already observed earlier, we only need to quantize the $t-1$ phase angle differences given by the entries of $\underline{\phi}$. With SQ, the $t-1$ phase angles are quantized independently of each other to get $\underline{\hat{\phi}} \triangleq [\hat{\phi}_2, \ldots, \hat{\phi}_t]$, an approach followed in [37], although the focus there was on the $t = 2$ case. In this section, we will consider the case of i.i.d. Rayleigh fading channels, and derive an upper bound for the expected MSwIP (approximate capacity loss) performance of SQ. When the channel $\mathbf{h}$ is i.i.d. Rayleigh fading, it is clear that the phase angles $\underline{\phi}$ are i.i.d. uniformly distributed on $[-\pi, \pi)$, and independent of the gains $\mathbf{r}$. Also, the gains $\mathbf{r}$ are i.i.d. χ-distributed. The capacity loss (3.25) can be written as

$$C_L \simeq \mathbf{E}\left\{\alpha^2\left(1 - \frac{\left|\hat{\mathbf{w}}^H\mathbf{h}\right|^2}{\|\mathbf{h}\|_1^2}\right)\right\}, \tag{3.43}$$

where, $\alpha^2 = \frac{P_s\|\mathbf{h}\|_1^2}{1+P_s\|\mathbf{h}\|_1^2}$ as before, and $\hat{\mathbf{w}}$ is the quantized beamforming vector corresponding to $\mathbf{h}$. Note that the right hand side is the expected MSwIP with Q-EGT. Now, $\hat{\mathbf{w}}$ is given by $\hat{\mathbf{w}} = [1\, \exp(j\underline{\hat{\phi}})]^T$. The above capacity loss can be expressed as $C_L = \mathbf{E}_{\mathbf{r}}\{C_{L,\mathbf{r}}\}$, where the expectation is taken over the joint distribution of $\mathbf{r}$; and $C_{L,\mathbf{r}}$, the capacity loss conditioned on the gains $\mathbf{r}$, is

$$\begin{aligned} C_{L,\mathbf{r}} &\simeq \mathbf{E}_{\underline{\phi}}\left\{\alpha^2\left(1 - \frac{\left|[1\, \exp(-j\underline{\hat{\phi}})]\exp(j\theta_1)\mathrm{diag}([1\, \exp(j\underline{\phi})])\mathbf{r}\right|^2}{s_{\mathbf{r}}^2}\right)\right\} \\ &= \mathbf{E}_{\underline{\phi}}\left\{\alpha^2\left(1 - \frac{\left|[1\, \exp(j(\underline{\phi} - \underline{\hat{\phi}}))]\mathbf{r}\right|^2}{s_{\mathbf{r}}^2}\right)\right\} \\ &= \alpha^2\left(1 - \frac{1}{s_{\mathbf{r}}^2}\mathbf{r}^T A_{\underline{\psi}}\mathbf{r}\right), \end{aligned} \tag{3.44}$$

where $\mathbf{E}_{\underline{\phi}}\{\}$ is the expectation over the joint distribution of phase angles $\underline{\phi}$. We have also defined $\underline{\psi} \triangleq \underline{\phi} - \underline{\hat{\phi}}$, and $A_{\underline{\psi}}$ as

$$A_{\underline{\psi}} \triangleq \mathbf{E}_{\underline{\phi}} \left\{ \begin{bmatrix} 1 \\ \exp(j\underline{\psi}) \end{bmatrix} [1 \exp(-j\underline{\psi})] \right\}. \tag{3.45}$$

Then, we have,

$$\begin{aligned} C_L &= \mathbf{E}_{\mathbf{r}}\{C_{L,\mathbf{r}}\} = \mathbf{E}_{\mathbf{r}}\left\{\alpha^2\right\} - \mathbf{E}_{\mathbf{r}}\left\{\frac{\alpha^2}{s_{\mathbf{r}}^2}\mathbf{r}^T A_{\underline{\psi}}\mathbf{r}\right\} \\ &= \mathbf{E}_{\mathbf{r}}\left\{\alpha^2\right\} - \mathbf{E}_{\mathbf{r}}\left\{\mathrm{tr}\left(A_{\underline{\psi}}A_{\mathbf{r}}\right)\right\}, \end{aligned} \tag{3.46}$$

where $A_{\mathbf{r}} \triangleq \mathbf{E}_{\mathbf{r}}\left\{\alpha^2\mathbf{r}\mathbf{r}^T/s_{\mathbf{r}}^2\right\}$. Note that $A_{\mathbf{r}}$ is a real symmetric positive definite matrix with $a_1 \triangleq \mathbf{E}_{\mathbf{r}}\left\{\alpha^2 r_i^2/s_{\mathbf{r}}^2\right\}$ along the diagonal, and $a_2 \triangleq \mathbf{E}_{\mathbf{r}}\left\{\alpha^2 r_i r_j/s_{\mathbf{r}}^2\right\}$, $i \neq j$ as its off-diagonal entries. Note that a_1 and a_2 do not depend on the specific indices i and j, but only on whether $i = j$ or $i \neq j$, since $\mathbf{r}$ has i.i.d entries. Substituting for $A_{\underline{\psi}}$ from (3.45) and simplifying, we have

$$C_L = \mathbf{E}_{\mathbf{r}}\left\{\alpha^2\right\} - \mathbf{E}_{\underline{\phi}}\left\{Q(\underline{\psi})\right\}, \tag{3.47}$$

where $Q(\underline{\psi})$ is defined as

$$Q(\underline{\psi}) \triangleq \mathbf{E}_{\underline{\phi}} \left\{ [1 \exp(-j\underline{\psi})] A_{\mathbf{r}} \begin{bmatrix} 1 \\ \exp(j\underline{\psi}) \end{bmatrix} \right\}. \tag{3.48}$$

Now, if the total number of feedback bits B is large, we can assume that the individual phase angle errors $\psi_i = \phi_i - \hat{\phi}_i$ are small, and therefore, $Q(\underline{\psi})$ can be approximated by its second-order Taylor series expansion

$$Q(\underline{\psi}) \simeq Q(\underline{0}) + \underline{\psi}^T \nabla_{\underline{\psi}} Q(\underline{0}) + \frac{1}{2}\underline{\psi}^T \nabla_{\underline{\psi}}^2 Q(\underline{0})\underline{\psi}, \tag{3.49}$$

where, $Q(\underline{0})$ is given by

$$Q(\underline{0}) = \underline{1}^T A_{\mathbf{r}} \underline{1} = \mathbf{E}_{\mathbf{r}}\left\{\frac{\alpha^2}{s_{\mathbf{r}}^2}\underline{1}^T\mathbf{r}\mathbf{r}^T\underline{1}\right\} = \mathbf{E}_{\mathbf{r}}\left\{\alpha^2\right\}. \tag{3.50}$$

Also, from the Appendix 3.9.2, it is readily verified the Gradient and Hessian of $Q(\underline{\psi})$ at $\underline{\psi} = \underline{0}$ are given by $\nabla_{\underline{\psi}} Q(\underline{0}) = \underline{0}$, and $\nabla_{\underline{\psi}}^2 Q(\underline{0}) = 2a_2\left(-tI_{t-1} + \underline{1}\underline{1}^T\right)$,

with $a_2 \triangleq \mathbf{E}_{\mathbf{r}}\left\{\alpha^2 r_i r_j / s_{\mathbf{r}}^2\right\}$, $i \neq j$, as before. Substituting (3.49) in (3.47), and simplifying a little, we get

$$C_L \simeq a_2 \mathbf{E}_{\underline{\phi}}\left\{\underline{\psi}^T \left(t I_{t-1} - \underline{1}\underline{1}^T\right) \underline{\psi}\right\}, \tag{3.51}$$

The approximate capacity loss given above is upper bounded by the case where a uniform quantizer in conjunction with an equal bit allocation ($B_i = B/(t-1)$) is used to quantize each angle ϕ_i. This choice is reasonable, since the phase angles are i.i.d. uniformly distributed, the capacity loss is a monotonically increasing convex function of the phase angle error, and it is isotropic, i.e., it is invariant to exchanging (say) index i and j. Hence, we can expect the upper bound to in fact be very close to the actual performance. Under the assumption of uniform independent quantization, we have the classical result from source coding [59] $\mathbf{E}_{\underline{\phi}}\left\{\psi_i \psi_j\right\} = \Delta^2 \delta(i-j)/12$, where $\Delta \triangleq (2\pi)2^{-B_i} = (2\pi)2^{-B/(t-1)}$ is the quantization bin-width, and δ_k is the Kronecker delta function. Substituting, we get,

$$C_L \leq a_2 (t-1)^2 \mathbf{E}_{\underline{\phi}}\left\{\psi_i^2\right\} = \frac{\pi^2 (t-1)^2 a_2}{3} 2^{-\frac{2B}{t-1}}. \tag{3.52}$$

Now, the factor $a_2 \triangleq \mathbf{E}_{\mathbf{r}}\left\{\alpha^2 r_i r_j / s_{\mathbf{r}}^2\right\}$, $i \neq j$ can be easily computed through simulation since r_i are i.i.d. χ-distributed. We reiterate that the above upper bound is tight, since the uniform bit allocation and uniform quantization is in fact very likely to be the optimum scalar quantizer when B is a multiple of $(t-1)$. In Sec. 3.8, we plot the capacity loss versus the number of feedback bits B in Fig. 3.5 with scalar quantization for an i.i.d. Rayleigh flat fading channel, which shows that the above expression is in fact accurate.

3.7.1 Discussion

We can now immediately compare SQ (3.52) with VQ (3.36) in the high SNR regime ($\alpha^2 \simeq 1$), reproduced here for convenience:

$$\begin{aligned} C_L^{VQ} &\approx c_1 2^{-\frac{2B}{t-1}} \\ C_L^{SQ} &\approx c_2 2^{-\frac{2B}{t-1}} \end{aligned} \tag{3.53}$$

where $c_1 \triangleq 2\left(\frac{t-1}{t+1}\right)$, $c_2 \triangleq \frac{\pi^2(t-1)^2 a_2}{3}$, and $a_2 \triangleq \mathbf{E_r}\left\{\frac{r_1 r_2}{\left(\sum_{i=1}^{t} r_i\right)^2}\right\}$. We see that both SQ and VQ achieve *the same convergence rate* as B increases, and only differ in the coefficient term. This agrees with classical results from source coding literature for r-th power distortion functions (see, for example, [60]). In table 3.1, we have compared SQ and VQ for different values of t. The second and third columns are the coefficients in the capacity loss expression in (3.53). The fourth column is the excess number of bits necessary to achieve the same capacity loss with SQ as with VQ, which is given by $\Delta B \triangleq \frac{t-1}{2}\log_2(c_2/c_1)$, and the last column is the excess bits per dimension. Note that the SQ apparently requires 0.0419 bits more than VQ to achieve the same capacity loss performance in the $t = 2$ case. When $t = 2$, there is only one parameter, $\phi_2 \triangleq \angle h_2 - \angle h_1$, that needs to be quantized, and ϕ_2 is uniformly distributed. Therefore, we would expect that VQ would default to SQ in this case, and they should perform the same. The discrepancy arises from the quantization cell approximation of (3.15), as a result of which, the capacity loss computed for the VQ case is in fact a lower bound, as given by Theorem 2. Another interesting observation is that the gap between the SQ and VQ appears to converge a constant of 0.357 bits per dimension. This can been seen by taking the limit of ΔB as t becomes large. It is easy to show that $\Delta B/(t-1)$ approximately converges to $\log_2\left(\frac{\pi^2}{6}\right)/2 \approx 0.359$.

3.8 Numerical Results

In this section, we present simulation results to illustrate the performance of codebooks designed using the MSwIP criterion, and to validate the theoretical expressions obtained in the previous sections. For the simulations, we assumed a Rayleigh flat-fading channel with t transmit antennas and 1 receive antenna. 10,000 random channel instantiations were used in the averaging.

Fig. 3.3 shows the experimental results for capacity loss obtained by using the Lloyd algorithm to generate the quantized beamforming vectors. We use

Table 3.1: Comparison of SQ and VQ methods for equal gain transmission.

t	$c_1 \triangleq 2\left(\frac{t-1}{t+1}\right)$	$c_2 \triangleq \frac{\pi^2(t-1)^2 a_2}{3}$	$\Delta B \triangleq \frac{t-1}{2}\log_2(c_2/c_1)$	$\Delta B/(t-1)$
2	0.6667	0.7065	0.0419	0.0419
3	1	1.3265	0.4076	0.2038
4	1.2	1.7281	0.7892	0.2631
5	1.3333	1.9854	1.1488	0.2872
6	1.4286	2.1807	1.525	0.3051
8	1.5556	2.4339	2.260	0.3229
16	1.7647	2.8298	5.109	0.3406
32	1.8788	3.0672	10.96	0.3535
64	1.9385	3.1781	22.467	0.3566

the correlated Rayleigh fading channel model in [61] with the antenna elements separated by $d/\lambda = 0.5$. Also plotted are the performance of the simpler high-SNR and low-SNR codebooks, and as expected, the codebooks perform nearly optimally in their respective SNR regimes. Therefore, a two-codebook strategy appears to be reasonable for practical implementation, and one would choose between the high- and low-SNR codebooks, depending on the SNR operating point.

For the remainder of this section, we assume that the channel $\mathbf{h}$ is i.i.d. Rayleigh distributed. Fig. 3.4 shows the capacity loss performance of Q-EGT with $t = 2, 3, 4$ as a function of B, with both SQ as well as VQ. Also plotted is the performance of the codebook obtained by using the Grassmannian beamforming criterion proposed by Love et. al. in [8]. For generating the latter codebook, a computer search over 10,000 random code vectors was used, as suggested by the authors. The results indicate that while the Grassmannian codebook performs reasonably well at low feedback rates, there are significant losses in performance at higher feedback rates. This is because the Grassmannian beamforming criterion is in fact a min-max (worst-case) design criterion, whereas the MSwIP criterion directly attempts to minimize the average capacity loss. Also, note that curves corresponding to SQ and VQ are parallel, in agreement with the result of (3.53).

In Fig. 3.5, we validate the high-SNR, high-resolution approximation

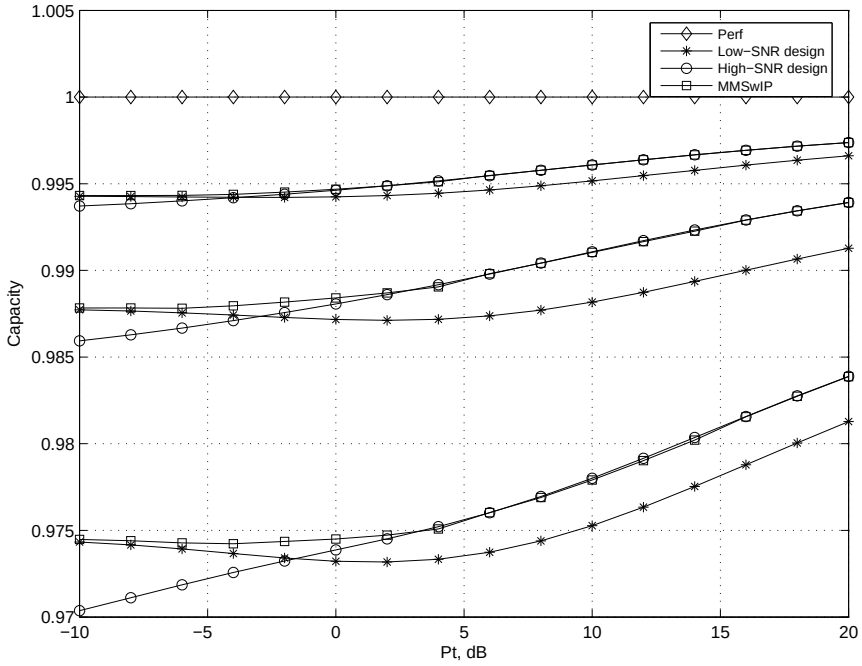


Figure 3.3: Ergodic capacity of the correlated MISO channel with Q-EGT for different quantizer design methods ($t = 3$ and $B = 1, 2, 3$ from the bottom). The capacities are normalized to the capacity of the perfect feedback system.

results for the capacity loss as given by (3.53). The curves for SQ and VQ coincide when $t = 2$, since there is only one parameter ($\phi_2 = \angle h_2 - \angle h_1$) that needs to be quantized. The theoretical expressions obtained closely match the results obtained from simulation. Fig. 3.6 plots the capacity loss of a $t = 3$ transmit antenna system versus the total transmit power (or SNR, since the noise power is unity). This illustrates that the combined error due to approximating the densities of α^2 and ξ_1 as well as assuming that they are uncorrelated is in fact small.

In Fig. 3.7, we compare the performance of the capacity achieved with four systems: a quantized MRT system, a quantized EGT system, a system that employs the identity transmit covariance matrix (which requires no feedback), and a system that employs a fixed beamforming vector (which also requires no feedback), for a fixed transmit power of $P_s = 15$dB. In the latter two cases, the

capacity expressions are given by [3]

$$\begin{aligned} C_{\text{no\_fb}} &= \mathbf{E}\left\{\log_2\left(I_t + P_s\mathbf{h}\mathbf{h}^H\right)\right\} = \mathbf{E}\left\{\log_2\left(1 + P_s\|\mathbf{h}\|_2^2\right)\right\}, \\ C_{\text{fix}} &= \mathbf{E}\left\{\log_2\left(1 + P_s\left|\mathbf{h}^H\underline{1}\right|^2\right)\right\}. \end{aligned}$$

The figure shows that the capacity of Q-MRT is always greater than or equal to that of Q-EGT for the same number of feedback bits, which is as expected since the Q-EGT is a more constrained system than the Q-MRT system. However, note that Q-MRT and Q-EGT perform almost the same in two cases ($t = 3, B = 3$) and ($t = 4, B = 3$). This is because these are close to ($t = 3, N = 7$) and ($t = 4, N = 7$), for which, it is known [39] that, with the minimum value for the maximum cross-correlation amplitude between any two code vectors as performance metric, the codebook designed under a total power constraint also satisfies a per-antenna power constraint. For these combinations of N and t, it is likely that the Q-MRT and Q-EGT have the same performance in terms of the capacity as well, when the channel is i.i.d. Rayleigh fading. Also, while the capacity of the system that employs the identity covariance matrix is higher than that obtained using a fixed beamforming vector, since the MISO channel is spatially single dimensional, even with one or two bits of feedback, the Q-MRT/Q-EGT system outperforms the system with the identity covariance matrix, when the receiver has perfect CSI and the feedback link is noiseless and has no delay. Analysis of the optimality of beamforming under quantized feedback can be found in [62], [63].

Fig. 3.8 shows the outage probability with $t = 3$ antennas at the transmitter as a function of the transmitted power P_s. Both the theoretical result of (3.41) and the simulation result obtained by using the MSwIP codebook are plotted. It is clear from the graph that the theoretical expressions agree well with the experimental results.

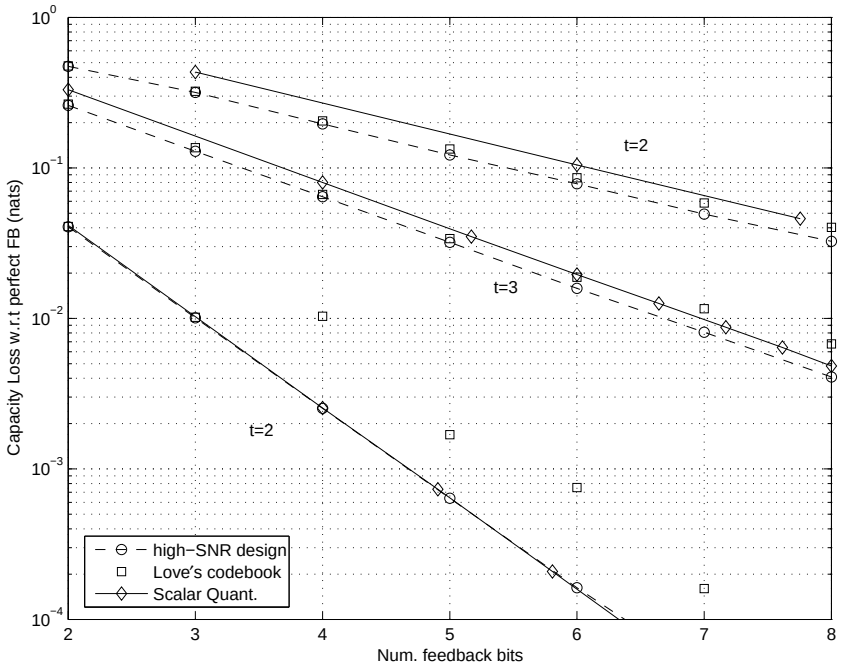


Figure 3.4: Capacity loss performance of Q-EGT, $t = 2, 3, 4$, $P_s = 10$ dB, and with SQ, VQ and Grassmannian beamforming.

3.9 Conclusion

We have investigated the problem associated with quantizing the per-antenna power constrained beamforming vector for MISO systems with finite-rate feedback. We have proposed a new design criterion, namely, maximizing the mean-square weighted inner-product (MSwIP), and developed a Lloyd-type VQ design algorithm, which can be used to design the codebook for flat-fading channels with any distribution. For practical implementation, we have proposed two sub-optimal *low-SNR* and *high-SNR* design algorithms as well. On the analytical side, we considered the i.i.d. Rayleigh fading case and derived theoretical expressions for the capacity loss and outage probability with both VQ as well as SQ. We see that quantized EGT requires half the number of parameters to be conveyed to the transmitter, and requires roughly half the number of feedback bits to achieve the same capacity loss (in bits), when compared to MRT. We also contrast the

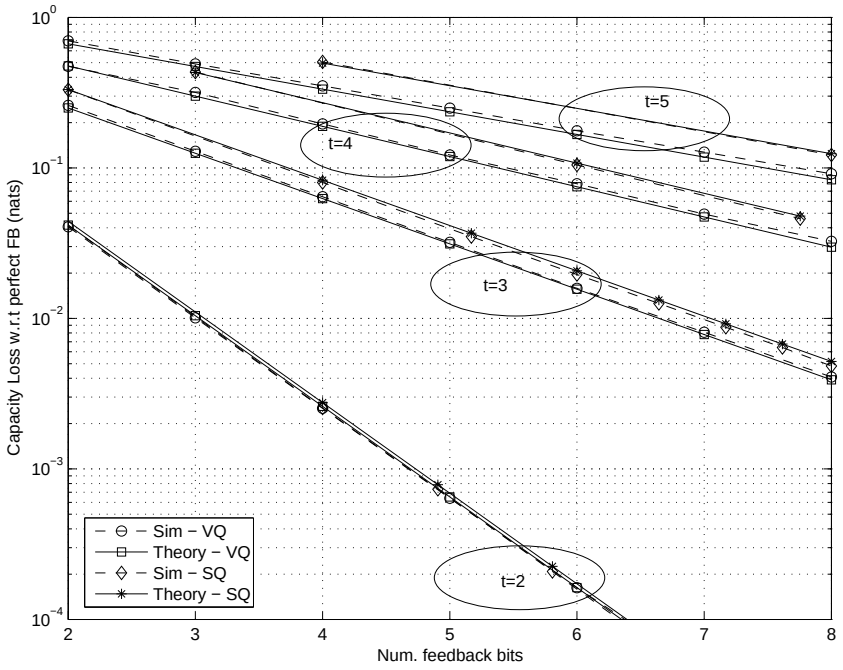


Figure 3.5: Capacity loss performance of Q-EGT, $t = 2, 3, 4, 5$. Here, 'theory' refers to equation (3.53).

performance of SQ and VQ, and see that both offer the same rate of convergence to the capacity with perfect feedback as B increases. However, there exists a constant gap between the two, with the VQ scheme requiring fewer bits to achieve a given level of capacity loss. For large number of transmit antennas, this gap is observed to approach a constant of 0.357 bits per dimension regardless of the number of transmit antennas. The accuracy of the theoretical expressions obtained are illustrated through Monte-Carlo simulations.

Appendix

3.9.1 Equivalence of Two Optimization Problems

Lemma 8. *The optimization problem* (3.3) *with* $\mathcal{S} = \mathcal{S}_P$ *is equivalent to the equal gain transmission (EGT) problem, recapitulated here for convenience:*

$$\mathbf{w}_o = \arg\max_{\mathbf{w}\in\mathcal{S}_E} \left|\mathbf{h}^H\mathbf{w}\right|^2, \tag{3.54}$$

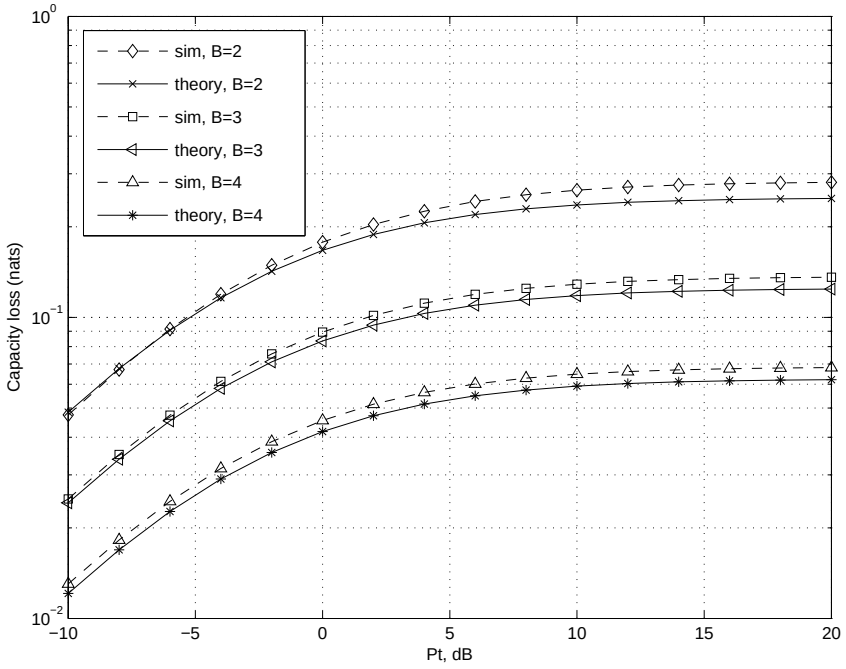


Figure 3.6: Capacity loss performance Q-EGT versus total transmit power, $t = 3$, $P_s = 10$dB, with vector quantization. Here, 'theory' refers to equation (3.36).

where $\mathcal{S}_E \triangleq \{\mathbf{w} : \mathbf{w} = [1, \exp(j\theta_2), \exp(j\theta_3), \ldots, \exp(j\theta_t)]^T\}$. *The problems are equivalent in the sense that they have the same solution.*

Proof. We wish to maximize $\left|\mathbf{h}^H\mathbf{w}\right|^2$, a convex function, over the convex set $\|\mathbf{w}\|_\infty \leq 1$. From [64] (theorem 3, page 181), we have that the maximum occurs at an *extreme* point. It is easy to show from the definition of an extreme point, that any $\mathbf{w} \in S_P$ is *not* an extreme point *unless* it satisfies $\mathbf{w} = [\exp(j\theta_1), \exp(j\theta_2), \ldots, \exp(j\theta_t)]^T$. However, as already observed, an overall phase (say, θ_1) is immaterial to the above maximization problem, and we can set $\theta_1 = 0$ without loss of generality, which proves the desired result. □

Although we have stated the above result for $r = 1$, it is valid for $r > 1$ as well, i.e., if $H \in \mathbb{C}^{r\times t}$, and we wish to solve the optimization problem $\max_{\mathbf{w}\in\mathcal{S}} \|H\mathbf{w}\|_2^2$, then the solution with $\mathcal{S} = \mathcal{S}_P$ is the same as that with $\mathcal{S} = \mathcal{S}_E$. Likewise, if $A \in \mathbb{C}^{t\times t}$ is a hermitian symmetric positive semi-definite matrix and

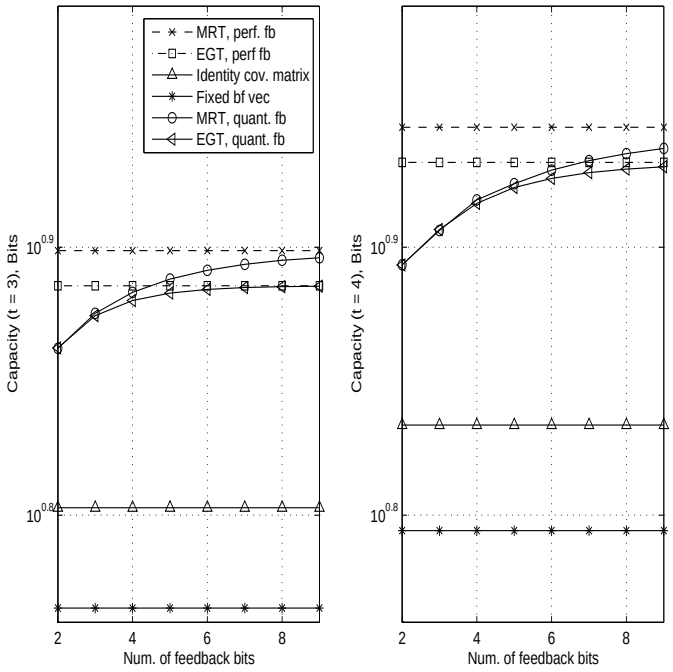


Figure 3.7: Capacity performance MRT, EGT (with perfect feedback), identity covariance matrix (no feedback), Q-MRT and Q-EGT versus the number of feedback bits B, for $t = 3$ (left) and $t = 4$ (right), and $P_s = 10$dB. Notice that even with 2 bits of feedback, Q-EGT/Q-MRT perform better than the identity covariance case, which requires no feedback.

we wish to solve $\max_{\mathbf{w}\in\mathcal{S}} \mathbf{w}^H A\mathbf{w}$, then the solution with $\mathcal{S} = \mathcal{S}_P$ is the same as that with $\mathcal{S} = \mathcal{S}_E$.

3.9.2 Gradient and Hessian of $Q(\mathbf{w}) \triangleq \mathbf{w}^H A\mathbf{w}$

In order to perform the Newton step to find the $\mathbf{w}$ that solves the constrained optimization problem given by (3.9), we need the gradient and Hessian of $Q(\mathbf{w}) \triangleq \mathbf{w}^H A\mathbf{w}$, where A is an arbitrary Hermitian symmetric matrix. The constrained optimization problem is best solved as an unconstrained optimization over the $\underline{\vartheta}$ space, where $\underline{\vartheta} \triangleq [\theta_2, \theta_3, \ldots, \theta_t]^T$. Write $\mathbf{w} = z_1(\underline{\vartheta}) \triangleq$

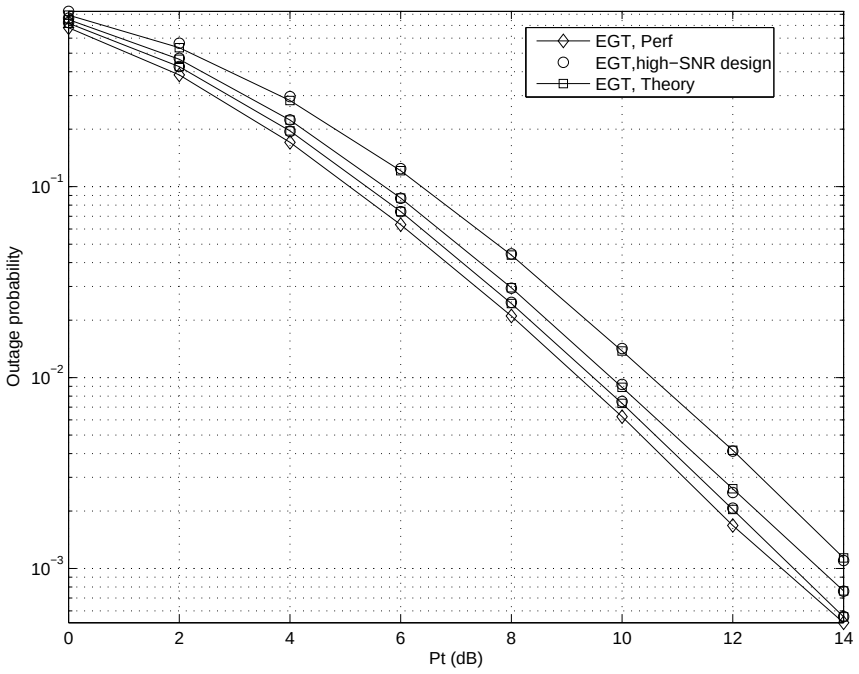


Figure 3.8: Outage probability of the MISO channel with quantized EGT ($t = 3; R = 2$ bits per channel use; $B = 2, 3, 4$ from the top. The theoretical curve refers to that obtained from (3.41).

$[1, \exp j\theta_2, \exp j\theta_3, \ldots, \exp j\theta_t]^T$. Also, let $z_2(\underline{\vartheta}) \triangleq z_1^*(\underline{\vartheta})$. Now,

$$\begin{aligned} Q(\mathbf{w}) &= \mathbf{w}^H A \mathbf{w} \\ \Rightarrow Q(\underline{\vartheta}) &= z_2(\underline{\vartheta})^T A z_1(\underline{\vartheta}) \end{aligned} \tag{3.55}$$

It is straightforward to obtain gradient as

$$\nabla_{\underline{\vartheta}} Q = (\nabla z_1(\underline{\vartheta}))\, A^T z_2(\underline{\vartheta}) + (\nabla z_2(\underline{\vartheta}))\, A z_1(\underline{\vartheta}), \tag{3.56}$$

where, $\nabla z_1(\underline{\vartheta})$ and $\nabla z_2(\underline{\vartheta})$ are given by

$$\nabla z_1(\underline{\vartheta}) = j \begin{bmatrix} 0 & \exp j\theta_2 & \cdots \\ 0 & 0 & \ddots \\ 0 & \cdots & \exp j\theta_t \end{bmatrix} \tag{3.57}$$

$$\nabla z_2(\underline{\vartheta}) = (\nabla z_1(\underline{\vartheta}))^* \tag{3.58}$$

Also, the Hessian is obtained as

$$\begin{aligned}\nabla^2_{\underline{\vartheta}} Q &= (\nabla z_2(\underline{\vartheta}))\, A\, (\nabla z_1(\underline{\vartheta}))^T + j\,(\nabla z_1(\underline{\vartheta}))\,\mathrm{diag}\left(A^T z_2(\underline{\vartheta})\right) \\ &+ (\nabla z_1(\underline{\vartheta}))\, A^T\, (\nabla z_2(\underline{\vartheta}))^T - j\,(\nabla z_2(\underline{\vartheta}))\,\mathrm{diag}\left(A z_1(\underline{\vartheta})\right) \qquad (3.59)\end{aligned}$$

3.9.3 Distribution of ξ_0 for 2 Transmit Antennas

First, note that we can consider $\mathbf{w}_0 = [1,1]^T$ without loss of generality due to the left-rotational invariance of the density of $\mathbf{h}$. Next, if we write the vector $\mathbf{h}$ with i.i.d entries as $\mathbf{h} = \exp(j\phi) \times [\alpha_1, \alpha_2 \exp(j\theta)]^T$, then α_1, α_2, ϕ and θ are independent, and α_i^2 are exponentially distributed, and ϕ and θ are uniformly distributed on $[0, 2\pi)$. Note that ξ_0 is independent of the overall phase ϕ, and therefore, without loss of generality, we can assume that $\mathbf{h}$ is of the form $[\alpha_1, \alpha_2 \exp(j\theta)]^T$. Then, we can write ξ_0 as

$$\xi_0 = \left\{ 1 - \frac{\left|\mathbf{h}^H \underline{1}\right|^2}{\|\mathbf{h}\|_1^2} \right\} = \frac{2\alpha_1\alpha_2\,(1-\cos(\theta))}{\alpha_1^2 + \alpha_2^2 + 2\alpha_1\alpha_2} \qquad (3.60)$$

Thus, the CDF $F_{\xi_0}(x) = \mathrm{Prob}(\xi_0 \le x)$ is given by

$$F_{\xi_0}(x) = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{Prob}(\xi_0 \le x|\theta) d\theta, \qquad (3.61)$$

where $\mathrm{Prob}(\xi_0 \le x|\theta)$ is the CDF of ξ_0 conditioned on θ, given by

$$\begin{aligned}\mathrm{Prob}(\xi_0 \le x|\theta) &= \mathrm{Prob}\left(\frac{\alpha_1^2}{\alpha_2^2} + \frac{\alpha_2^2}{\alpha_1^2} \ge z\right), \text{ where} \\ z &\triangleq 4\left(\frac{1-\cos\theta}{x} - 1\right)^2 - 2.\end{aligned}$$

Using the fact that $\alpha_i^2,\ i = \{1,2\}$ are exponentially distributed with PDF $f_{\alpha_i^2}(x) = \exp(-x)$, we can evaluate the above probability in closed-form as

$$\begin{aligned}\mathrm{Prob}(\xi_0 \le x|\theta) &= \begin{cases} 1 & z < 2 \\ 2 \cdot \mathrm{Prob}\left(\alpha_1^2 \le y_1 \alpha_2^2\right) & z \ge 2 \end{cases} \\ &= \begin{cases} 1 & z < 2 \\ \frac{2y_1}{1+y_1} & z \ge 2, \end{cases} \qquad (3.62)\end{aligned}$$

where y_1 is the smaller root that satisfies $y_1 + 1/y_1 = z$. Note that the condition $z \leq 2$ implies $\cos(\theta) \geq 1 - 2x$, and thus $F_{\xi_0}(x)$ can be simplified as

$$\begin{aligned} F_{\xi_0}(x) &= \frac{1}{\pi}\int_0^{\cos^{-1}(1-2x)} d\theta + \frac{1}{\pi}\int_{\cos^{-1}(1-2x)}^{\pi} \frac{2y_1}{1+y_1} d\theta \\ &= 1 - \frac{1}{\pi}\int_{\cos^{-1}(1-2x)}^{\pi} \frac{1-y_1}{1+y_1} d\theta \end{aligned} \tag{3.63}$$

Since

$$\frac{1-y_1}{1+y_1} = \left(\frac{z-2}{z+2}\right)^{\frac{1}{2}} = \frac{\sqrt{\left(\frac{1-\cos\theta}{x}\right)\left(\frac{1-\cos\theta}{x} - 2\right)}}{\frac{1-\cos\theta}{x} - 1}, \tag{3.64}$$

we can make the substitution $v \triangleq \frac{-(1+\cos\theta)}{2(1-x)}$ in (3.63) to get

$$F_{\xi_0}(x) = 1 - \frac{2(1-x)}{\pi}\int_0^1 \frac{\sqrt{1-v}}{((2-x) - 2(1-x)v)\sqrt{v}} dv \tag{3.65}$$

And again substitute $\sin^2\theta = v$ to get

$$F_{\xi_0}(x) = 1 - \frac{4(1-x)}{\pi}\int_0^{\pi/2} \frac{\cos^2\theta}{(2-x) - 2(1-x)\sin^2\theta} d\theta \tag{3.66}$$

Using formula 3.615.1 from [57],

$$\int_0^{\pi/2} \frac{\cos 2nx dx}{1 - a^2\sin^2 x} = \frac{(-1)^n\pi}{2\sqrt{1-a^2}}\left(\frac{1-\sqrt{1-a^2}}{a}\right)^{2n}, \quad \left[a^2 < 1\right], \tag{3.67}$$

we get

$$\begin{aligned} &\int_0^{\pi/2} \frac{\cos^2\theta}{(2-x) - 2(1-x)\sin^2\theta} d\theta = \\ &\frac{1}{2(2-x)}\left[\int_0^{\frac{\pi}{2}} \frac{d\theta}{1 - \frac{2(1-x)}{2-x}\sin^2\theta} + \int_0^{\frac{\pi}{2}} \frac{\cos 2\theta d\theta}{1 - \frac{2(1-x)}{2-x}\sin^2\theta}\right] \\ &= \frac{\pi}{2\sqrt{x(2-x)}}\left[1 - \left(\frac{1-\sqrt{1-\frac{2(1-x)}{2-x}}}{\sqrt{\frac{2(1-x)}{2-x}}}\right)^2\right] \\ &= \frac{\pi\left(\sqrt{x(2-x)} - x\right)}{4(1-x)\sqrt{x(2-x)}} \end{aligned} \tag{3.68}$$

Substituting into (3.66), we finally get $F_{\xi_0}(x)$ as

$$\begin{aligned} F_{\xi_0}(x) &= 1 - \frac{1}{\sqrt{x(2-x)}}\left(\sqrt{x(2-x)} - x\right) \\ &= \sqrt{\frac{x}{2-x}};\ 0 \leq x \leq 1. \end{aligned} \tag{3.69}$$

Differentiating, we get the PDF of ξ_0 as

$$f_{\xi_0}(x) = \frac{1}{(2-x)\sqrt{x\,(2-x)}};\ 0 \leq x \leq 1. \tag{3.70}$$

3.9.4 Distribution of the parameter

Recall that $\underline{\alpha} = |\mathbf{v}|$, where $\mathbf{v} \triangleq \mathbf{h}/\|\mathbf{h}\|_2$. Let $\mathbf{x} \triangleq \mathrm{Re}\,(\mathbf{v})$, $\mathbf{y} \triangleq \mathrm{Imag}\,(\mathbf{v})$ and $\underline{\theta} \triangleq \angle\mathbf{v}$. Then, we have $\mathbf{x} = \underline{\alpha}\cos\underline{\theta}$ and $\mathbf{y} = \underline{\alpha}\sin\underline{\theta}$. The Jacobian of the transformation is

$$J = \frac{\partial\,(\mathbf{x},\mathbf{y})}{\partial\,(\underline{\alpha},\underline{\theta})} = \begin{bmatrix} \cos\theta_1 & 0 & \dots & -\alpha_1\sin\theta_1 & 0 & \dots \\ 0 & & \ddots & & \ddots & \\ \dots & & \cos\theta_t & 0 & \dots & -\alpha_t\sin\theta_t \\ \sin\theta_1 & 0 & \dots & \alpha_1\cos\theta_1 & 0 & \dots \\ 0 & & \ddots & & \ddots & \\ \dots & & \sin\theta_t & 0 & \dots & \alpha_t\cos\theta_t \end{bmatrix} \tag{3.71}$$

It is easy to show that the determinant of the Jacobian is

$$\det[J] = \alpha_1\alpha_2\dots\alpha_t \tag{3.72}$$

Thus, we have

$$f_{\underline{\alpha},\underline{\theta}}(\underline{\alpha},\underline{\theta}) = \alpha_1\dots\alpha_t f_{\mathbf{x},\mathbf{y}}(\underline{\alpha}\cos\underline{\theta},\underline{\alpha}\sin\underline{\theta}) \tag{3.73}$$

But we know that $f_{\mathbf{x},\mathbf{y}}(\cdot)$ is a constant since $\mathbf{v}$ is uniformly distributed on the unit circle. The value of the constant is the inverse of the surface area of the $2(t-1)$ dimensional unit sphere, given by $2\pi\kappa_{2t-2}$. Thus, we wan to evaluate

$$\begin{aligned} f_{\underline{\alpha}}(\underline{\alpha}) &= \int f_{\underline{\alpha},\underline{\theta}}(\underline{\alpha},\underline{\theta})d\underline{\theta} \\ &= (2\pi\kappa_{2t-2})^{-1}\,\alpha_1\dots\alpha_t\int d\underline{\theta} \\ &= (2\pi\kappa_{2t-2})^{-1}\,\alpha_1\dots\alpha_t\,(2\pi)^t \end{aligned} \tag{3.74}$$

And note that $f_{\underline{\alpha}}(\underline{\alpha}) = 0$ whenever $\underline{\alpha}^T\underline{\alpha} \neq 1$.

Acknowledgement

This chapter, in part, is a reprint of a paper which has been accepted for publication in the *IEEE Transactions on Signal Processing* as C. R. Murthy and B. D. Rao, "Quantization methods for equal gain transmission with finite rate feedback".

4 High-Rate VQ for Noisy Channels With Random Index Assignment: Part 1: Theory

4.1 Introduction

It is well known that the performance source quantization can be very sensitive to errors introduced when the codepoint index is transmitted over a noisy channel. For example, speech is typically compressed using a highly efficient vector quantization (VQ) scheme prior to transmission over a noisy channel, and the resulting indices could be very sensitive to errors in the channel over which they are transmitted. Hence, the performance of VQ when the index is sent over a noisy channel is pertinent to practical communication systems.

The effect of channel errors on VQ can be modeled as an *index error*, that is, the index i corresponding to the current source instantiation is received as a possibly different index j. Classical source coding has devoted much effort to the problem of source compression for noisy channels in the past couple of decades, and two dominant approaches have emerged. The first is channel-optimized VQ, i.e., to replace the distortion measure used for optimizing the quantizer with the expected distortion over the noisy channel [65] - [68]. The second approach involves *index assignment*, i.e., to design the quantizer without considering channel errors and then map codewords to indices in such a way that codewords resulting

in small inter-codepoint distortions are mapped to index pairs that correspond to channel symbols with large transition probability and vice versa [69] - [72]. Other recent works on source quantization for noisy channels include [73] - [75]. Most of the works in the literature employ Mean Squared Error (MSE) as the distortion function. However, more general distortion measures are often of interest and considered here (such as Log Spectral Distortion (LSD) in wideband speech spectrum quantization) and the performance of source compression under channel errors is examined. Furthermore, in this work, it is assumed that the index assignment is *random*. Random indexing results in a so-called *simplex error channel*, i.e., the probability of receiving an index j when an index i is sent only depends on whether or not i and j are different. Thus, in this chapter, based on classical results from the source coding literature [59] - [78], new results are derived analyzing the effect of errors on the performance of source coding with arbitrary distortion measures for simplex error channels.

Clearly, it would be convenient if the overall distortion could be decomposed as the sum the distortion due to the source encoder and the distortion induced by channel errors. It is shown in Sec. 4.4 that while this decomposition is possible when the distortion is measured as the MSE, in general there is an interdependence between the source and channel errors. The interdependence is characterized for the case of discrete symmetric channels with random index assignment. The rest of this chapter is organized as follows. In Section 4.2, the system model is described, and in 4.3, the noisy channel model and some of its properties are indicated. In Section 4.4, the expected distortion of source coding for noisy channels with random index assignment is derived. Section 4.5 presents several extensions to the analysis, including optimization of the point density to minimize the expected distortion. Simulation results to verify the accuracy of the analysis are presented in Section 4.6, and concluding remarks are offered in Section 4.7.

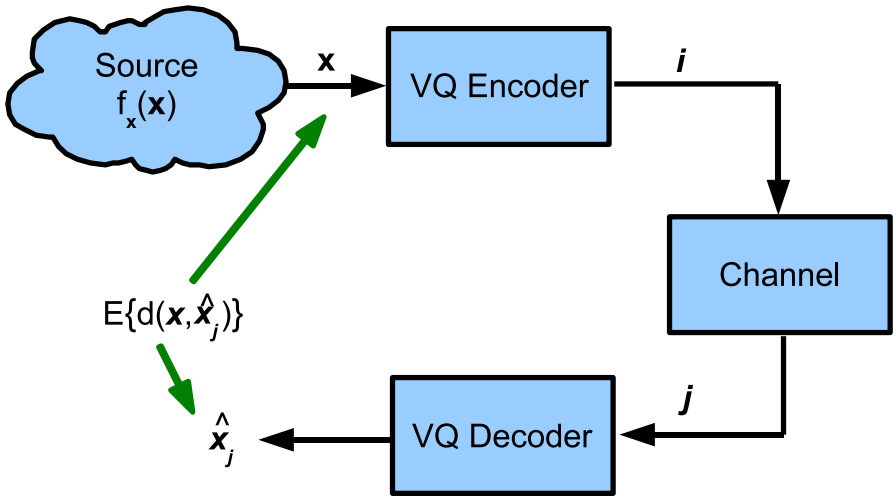


Figure 4.1: Block diagram of the vector quantizer and the noisy channel

4.2 Preliminaries

The block diagram of the system under consideration is shown in Fig. 4.1. Let the n-dimensional real vector $\mathbf{x} \in \mathcal{D}_{\mathbf{x}} \subset \mathbb{R}^n$ be the random source with a continuous pdf $f_{\mathbf{x}}(\mathbf{x})$, where $\mathcal{D}_{\mathbf{x}}$ is the domain of $\mathbf{x}$. For convenience, many of the results derived in this chapter will assume that $\mathcal{D}_{\mathbf{x}}$ is compact, although the results are generalizable with some effort. The vector quantization encoder and decoder are described by N *partition regions* $\bar{\mathcal{R}}_i, 1 \leq i \leq N$ that tile $\mathcal{D}_{\mathbf{x}}$, and an associated *codebook* of representation or reconstruction vectors $\mathcal{C} = \{\hat{\mathbf{x}}_i, 1 \leq i \leq N\}$, respectively. That is, whenever $\mathbf{x} \in \bar{\mathcal{R}}_i$, the encoder outputs the index i, and whenever the decoder receives index j, it outputs the vector $\hat{\mathbf{x}}_j$. Let $d(\mathbf{x}, \hat{\mathbf{x}})$ represent the distortion incurred in representing a source instantiation $\mathbf{x}$ by $\hat{\mathbf{x}}$. The distortion function is assumed to be non-negative, twice continuously differentiable and bounded, and with $d(\mathbf{x}, \mathbf{x}) = 0, \forall \mathbf{x}$.

Given the code book, two encoder structures are well-known: first, when the transmitter does not account for possible channel errors, the partition regions of the *transmit-unoptimized* quantizer are given by the well-known *nearest neighbor*

condition (NNC):

$$\bar{\mathcal{R}}_i = \{\mathbf{x} : Q(\mathbf{x}) = \hat{\mathbf{x}}_i\} = \{\mathbf{x} : d(\mathbf{x}, \hat{\mathbf{x}}_i) < d(\mathbf{x}, \hat{\mathbf{x}}_j) \; \forall \; j \neq i\} \,. \tag{4.1}$$

Note that, by definition, the code book at the transmitter satisfies $\hat{\mathbf{x}}_i \in \bar{\mathcal{R}}_i$. Also, the encoder operation can now be stated simply as $Q(\mathbf{x}) = \hat{\mathbf{x}}_i$ whenever $\mathbf{x} \in \bar{\mathcal{R}}_i$. The vectors in the code book $\hat{\mathbf{x}}_i$ can still be chosen to minimize the distortion while accounting for the channel characteristics, therefore, use of the NNC does not preclude optimizing the code book for noisy channels.

The probability of $\mathbf{x}$ lying in the region $\bar{\mathcal{R}}_i$ is $P_i = \int_{\bar{\mathcal{R}}_i} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}$, and the *cell diameter* is defined as

$$\delta_N \triangleq \max_{1 \leq i \leq N} \text{diam} \left(\bar{\mathcal{R}}_i \cap \mathcal{D}_{\mathbf{x}} \right) , \tag{4.2}$$

where diam is the Euclidean distance between two points that are furthest away in $\bar{\mathcal{R}}_i$. A sequence of quantizers is said to have *diminishing* cell diameters if $\lim_{N \to \infty} \delta_N = 0$. One of the consequences of having diminishing cell diameters and a continuous density $f_{\mathbf{x}}(\mathbf{x})$ is that $P_{\max}(N) \triangleq \max_{1 \leq i \leq N} P_i \to 0$ as $N \to \infty$.

Second, given $\mathbf{x}$, the *transmit-optimized* quantizer chooses the index i that minimizes the expected distortion *after* accounting for possible channel errors, denoted $d_c(\mathbf{x}, \hat{\mathbf{y}}_i)$:

$$d_c(\mathbf{x}, \hat{\mathbf{y}}_i) \triangleq \sum_{k=1}^{N} d(\mathbf{x}, \hat{\mathbf{y}}_k) P_{k|i}. \tag{4.3}$$

Therefore, the quantization region, now denoted $\bar{\mathcal{S}}_i$, is given by the *weighted nearest-neighbor* condition:

$$\bar{\mathcal{S}}_i \;\; = \;\; \{\mathbf{x} : d_c(\mathbf{x}, \hat{\mathbf{y}}_i) < d_c(\mathbf{x}, \hat{\mathbf{y}}_j), \forall \; j \neq i\} \tag{4.4}$$

where $P_{j|i}$ is the *index transition probability*, i.e., the conditional probability that the transmitted index i is received as index j. In this chapter, it will be assumed that the quantization regions are defined by either (4.1) or (4.4), i.e., *the encoder is optimized* either for an error-free channel or for the noisy channel.

Similarly, given the quantization regions at the encoder ($\bar{\mathcal{R}}_i$ or $\bar{\mathcal{S}}_i$), there exist two possibilities for designing the codebook at the receiver. The first option, called the *centroid condition*, is stated as

$$\hat{\mathbf{y}}_j = \arg\min_{\hat{\mathbf{y}}} \mathbf{E}\left\{d(\mathbf{x},\hat{\mathbf{y}})|Q(\mathbf{x})=\hat{\mathbf{x}}_j\right\}, \tag{4.5}$$

where the expectation is taken over the set of $\mathbf{x}$ that are quantized to index j. Note that the quantizer function$Q(\cdot)$ could either be the transmit-unoptimized quantizer described by (4.1) or the transmit-optimized quantizer described by (4.4). As the receiver does not account for the channel characteristics, this approach is called the *receive-unoptimized*. In contrast, the *weighted centroid condition*, which used in the *receive-optimized* approach, is described by

$$\hat{\mathbf{y}}_j = \arg\min_{\hat{\mathbf{y}}} \sum_{i=1}^{N} P_{j|i}\mathbf{E}\left\{d(\mathbf{x},\hat{\mathbf{y}})|Q(\mathbf{x})=\hat{\mathbf{x}}_i\right\}, \tag{4.6}$$

where the right hand side of the above expression is the expectation of the distortion over all possible quantized indices at the transmitter i, given that the received index is j. Note that the generalized Lloyd algorithm can be used to generate a codebook of *channel optimized* vectors in the case of MSE distortion, as described in [66]. In this chapter, it will be assumed that in the channel optimized case, the codebook is generated using the techniques available in the literature, and the focus will be on the theoretical analysis of such (locally) optimum codebooks. Note that a *sub-optimal* codebook, such as a product codebook or a structured codebook, does not, in general, satisfy either (4.5) or (4.6). To allow for the development of a more general result, this work does not assume that the centroid or weighted centroid condition is satisfied, i.e., the codebook may be sub-optimal. Next, the simple channel model considered in this chapter is described, and then the performance of VQ when the index is transmitted over such noisy channels is analyzed.

4.3 Discrete Symmetric Channels

The channel model considered in this chapter is the Discrete Symmetric Channel (DSC), whose *transition probability matrix*, which is an $N \times N$ matrix with (i, j)-th element being the transition probability $P_{j|i}$ that index j is received given that index i was sent, has the property that every row is a permutation of the first row, and every column is the permutation of the first column [79]. The DSC has the additional property that $P_{i|i}$ is independent of i, a fact that will be used in the derivations to follow. It is also reasonable to assume that $P_{i|i} \geq P_{j|i}$, $\forall j \neq i$, i.e., when index i is transmitted, the most likely index to be received is i itself, which implies $P_{i|i} \geq 1/N$. Finally, the average of the probabilities of all possible index errors is denoted as

$$\epsilon(N) \triangleq \frac{1}{N(N-1)} \sum_{\substack{i,j=1,\\ j\neq i}}^{N} P_{j|i}. \tag{4.7}$$

$\epsilon(N)$ could depend on N, and satisfies $0 \leq \epsilon(N) \leq 1/N$, since $P_{i|i} \geq 1/N$.

Note that with a DSC, the performance generally depends on the assignment of indices to code points. The number of possible index assignments is $N!$, and problem of finding the optimum index assignment is known to be NP complete [80], hence, random search-based techniques are employed in practice. In this chapter, however,the index assignment problem is circumvented by assuming that it is chosen randomly and uniformly from the $N!$ possible assignments, independent of the encoder. When a *random* index assignment is employed, the equivalent channel has a transition probability given by

$$P_{j|i} = \begin{cases} \epsilon(N), & j \neq i \\ 1-(N-1)\epsilon(N), & j = i \end{cases}, \tag{4.8}$$

i.e., it is the transition probability obtained after averaging over all possible permutations. The above transition probability could also arise, for example, when the indices are transmitted over the channel using any orthogonal modulation scheme.

For convenience, a channel with the index transition probability given above will be referred to as a *simplex error channel (SEC)*.

There are several reasons why a random index assignment is pertinent. Theoretically, it makes the analysis tractable, and closed-form expressions for the expected distortion for source coding for noisy channels can be derived. Moreover, the random index assignment approach provides an analytic upper bound on the performance of the best possible index assignment, and a lower bound on the performance of the worst index assignment, as pointed out in [68]. Also, in many practical systems, the quantizer output index is encoded using a fairly powerful channel code before its transmission over a noisy channel. This makes the computation of the pairwise index transition probability intractable; however, one can experimentally (or theoretically) often find the probability of an index error. Then, with random indexing, the channel can be modeled as a simplex error channel with the given probability of correct index reception, and the results presented in this chapter apply. However, this should not be construed to mean that little gains can be obtained from index assignment, but rather that if the number of quantization levels is large, finding the best (or even a good) index assignment can be computationally expensive. In particular, for applications such as wideband speech spectrum compression where the number of code-points is huge, the index assignment as well as the pairwise index transition probabilities could be computationally infeasible. In such cases, it is pertinent to consider a random index assignment for practical implementations, where the randomness could be achieved by employing different index assignments over time in a pre-specified pattern.

The discrete symmetric channel with random index assignment has the following interesting properties.

Property 1: When random index assignment is employed, the transmit-optimized quantizer of (4.4) is equivalent to employing the same codebook at the transmitter and the receiver. Let $\hat{\mathbf{y}}_i, 1 \leq i \leq N$ be the set of reconstruction vectors at the receiver. Then, the Voronoi regions corresponding to the transmit-optimized

quantizer is given by

$$\begin{aligned}
\bar{\mathcal{S}}_i &= \left\{ \mathbf{x} : \sum_{k=1}^{N} d(\mathbf{x}, \hat{\mathbf{y}}_k) P_{k|i} < \sum_{l=1}^{N} d(\mathbf{x}, \hat{\mathbf{y}}_l) P_{l|j}, \forall\, j \neq i \right\} \\
&= \left\{ \mathbf{x} : (1-(N-1)\epsilon(N))\, d(\mathbf{x}, \hat{\mathbf{y}}_i) + \epsilon(N) d(\mathbf{x}, \hat{\mathbf{y}}_j) + \epsilon(N) \sum_{\substack{k=1 \\ k\neq i,j}}^{N} d(\mathbf{x}, \hat{\mathbf{y}}_k) \right. \\
&\qquad \left. < (1-(N-1)\epsilon(N))\, d(\mathbf{x}, \hat{\mathbf{y}}_j) + \epsilon(N) d(\mathbf{x}, \hat{\mathbf{y}}_i) + \epsilon(N) \sum_{\substack{l=1 \\ l\neq i,j}}^{N} d(\mathbf{x}, \hat{\mathbf{y}}_l), \forall\, j \neq i \right\} \\
&= \left\{ \mathbf{x} : (1-N\epsilon(N))\, d(\mathbf{x}, \hat{\mathbf{y}}_i) < (1-N\epsilon(N))\, d(\mathbf{x}, \hat{\mathbf{y}}_j), \forall\, j \neq i \right\} \\
&= \left\{ \mathbf{x} : d(\mathbf{x}, \hat{\mathbf{y}}_i) < d(\mathbf{x}, \hat{\mathbf{y}}_j), \forall\, j \neq i \right\},
\end{aligned}$$

where the last equality follows because $\epsilon(N) < \frac{1}{N}$. Note that since the index assignment is random and independent of the encoding operation, the transition probability of the equivalent channel obtained by averaging over all possible permutations is used for $P_{j|i}$ above. Thus, the transmitter employs the same code-points $\{\hat{\mathbf{y}}_i, 1 \leq i \leq N\}$ as the receiver. Following the steps backwards, it is clear that if the same code-book is employed at the transmitter and the receiver, no further optimization is possible at the transmitter, i.e., the resulting quantizer is transmit-optimized. Due to this property, in the sequel, it is assumed that the encoder and the decoder share a *common codebook* $\{\hat{\mathbf{x}}_i,\ 1 \leq i \leq N\}$.

Property 2: With random index assignment, both the transmit unoptimized quantizer of (4.1) as well as the transmit optimized quantizer of (4.4) are guaranteed to be *regular* [1] [68]. The regularity of the transmit-unoptimized quantizer follows from its definition, while that of the transmit-optimized quantizer is a direct consequence of the above property.

Note that there exists an upper bound on the distortion for noisy channels that is independent of the channel behavior and N, as follows. If $\bar{\mathbf{x}}_d$ is defined as

[1] a quantizer is said to be *regular* if each encoding cell $\bar{\mathcal{S}}_i$ is convex and contains the code-vector $\hat{\mathbf{y}}_i = Q(\bar{\mathcal{S}}_i)$

the centroid of $\mathbf{x}$ under $d(\mathbf{x}, \mathbf{y})$, i.e.,

$$\bar{\mathbf{x}}_d \triangleq \arg\min_{\hat{\mathbf{x}}} \int_{\mathbf{x}} d(\mathbf{x}, \hat{\mathbf{x}}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}, \tag{4.9}$$

then, for the optimum quantizer, the expected distortion is upper bounded by

$$E_d \leq \int_{\mathbf{x}} d(\mathbf{x}, \bar{\mathbf{x}}_d) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}. \tag{4.10}$$

This is because the distortion can always be upper-bounded by the case where the codebook consists of the single vector $\bar{\mathbf{x}}_d$. In this case, channel errors have no effect on the performance, since the decoder always outputs $\bar{\mathbf{x}}_d$ irrespective of which index is received, and the expected distortion is simply given by (4.10).

In the above, one could also include the situation where the receiver outputs (say) index 0 to declare an erasure; in this case, it is equivalent to having one additional vector at the receiver compared to that at the transmitter. The analysis of such a system can also be carried out along the lines presented in this chapter, with straightforward modifications. Note that under this set-up, $P_{i|i} = 1 - (N-1)\epsilon(N)$ is the probability of correct reception. This implicitly assumes that as N is increased, more (or less) energy is used to transmit the symbol in order to maintain the probability of correct reception given above. For example, one simple model is obtained by assuming that as N is increased, the per-index transmit power is increased to maintain a constant probability of correct index reception, that is, $\epsilon(N) = \rho/(N-1)$. In this case, $P_{i|i} = 1-\rho$ is independent of N. Another example is when the index is mapped to a $L \triangleq \log_2(N)$ bit symbol, and each bit is transmitted over a Binary Symmetric Channel (BSC) with cross-over probability q. In this case, the probability of correct reception $P_{i|i} = (1-q)^L$, and thus $\epsilon(N) = \left(1 - (1-q)^L\right) / (N-1)$.

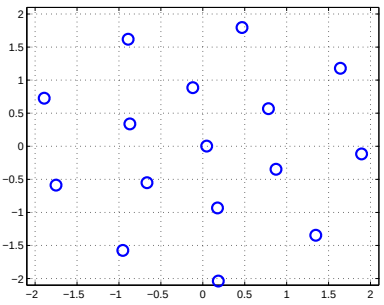
4.4 High-Rate Performance of Vector Quantization

In this section, the high-rate performance of vector quantization for the case of discrete symmetric channels with random index assignment is derived. The

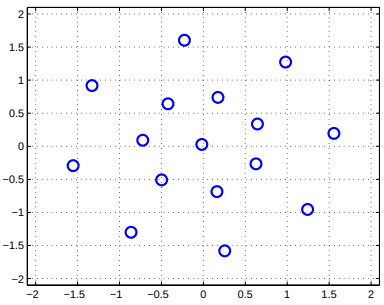
development in this section is as follows. First, the *expected distortion* performance is characterized, where the expectation is taken over the source statistics, the channel statistics as well as the random index assignment. To do this, a particular structure is assumed for the codebook, which leads to an upper bound on the performance with an optimum unstructured codebook. Specifically, it is assumed that some fraction of the codepoints are merged into the centroid $\bar{\mathbf{x}}_d$, while the remaining codepoints are distinct. Next, since the index assignment is random, the expected distortion is a random variable depending on the index assignment. When the fraction of merged codepoints goes to zero, an *upper bound* on the *variance* of the distortion is derived. From the upper bound, it is seen that when the number of dimensions $n \geq 4$, the variance of the distortion goes to zero *faster* than the expected distortion as N gets large. Practically, this implies that a vast majority of the index assignments are equally good (or equally bad), hence, random search based techniques would be computationally inefficient in finding the best index assignment as N gets large.

4.4.1 Codebook Structure

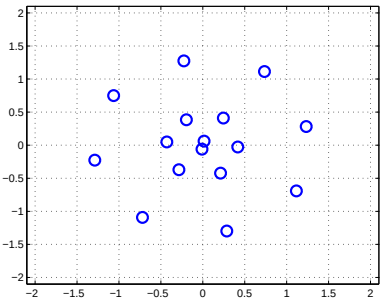
Fig. 4.2 shows the channel-optimized codepoints obtained using the generalized Lloyd algorithm described in [66], for 16-level quantization of a 2-dimensional i.i.d. Gaussian source. The Fig. 4.2(a) shows the conventional codebook, i.e., one that is optimized for an error-free channel. In Fig. 4.2(b), the channel is a BSC with bit cross-over probability $q = 0.02$, and random index assignment. Since the encoder is unaware of the particular index assignment, the codebook is designed for the equivalent SEC described in the previous section. Notice that the codepoints are now closer to the origin (i.e., the source centroid), as expected. Fig. 4.2(c) corresponds to $q = 0.05$, and here two codepoints are starting to *merge* at the origin. Loosely speaking, the generalized Lloyd algorithm attempts to perform *joint* source-channel coding by mapping multiple indices to the same vector (at the centroid). Finally, 4.2(d) shows the codepoints with $q = 0.1$, and it



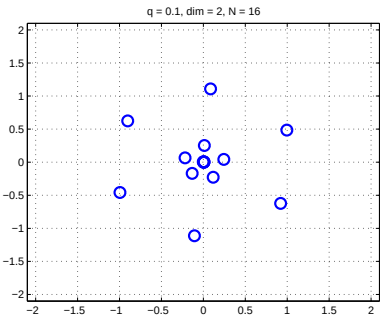
(a) VQ codebook, noiseless channel



(b) COVQ codebook, BSC, $q = 0.02$



(c) COVQ codebook, BSC, $q = 0.05$



(d) COVQ codebook, BSC, $q = 0.1$

Figure 4.2: VQ codepoints for $N = 16$ level quantization of a $n = 2$ dimensional i.i.d. zero mean unit variance Gaussian source. The codebooks were generated using the channel-optimized version of the generalized Lloyd algorithm.

is clear that 5 out of the 16 codepoints are merged at the centroid. Thus, as the channel deteriorates, more and more codepoints get merged at the origin, until, for a completely degenerate channel (i.e., one for which $P_{j|i} = 1/N$ for all i and j), the codebook contains just one distinct codepoint at $\bar{\mathbf{x}}_d$.

The above example motivates the particular structure assumed for the Channel Optimized Vector Quantizer (COVQ) codebook in the sequel. In classical source coding, when the channel is noiseless, it is known that the distribution of codepoints often approximates a *continuous point density*. However, when the channel has errors, the codepoints of the optimum codebook initially shrink closer towards the centroid of the source distribution, and eventually, some of the codepoints collapse together and the point density becomes *singular*. A singular point density can be thought of as the sum of a continuous point density and one or more singular points. In this section, an upper bound on the expected distortion is obtained by assuming that the optimum point density consists of a continuous part, and a delta function at the centroid $\bar{\mathbf{x}}_d$ given by (4.9). Then, of the total N code points, αN *distinct* points are drawn from a *continuous* density (it is assumed that αN is a large integer for large enough N), while the remaining $(1-\alpha)N$ points are at the *centroid* $\bar{\mathbf{x}}_d$, where $0 \leq \alpha \leq \frac{N-1}{N}$. Note that α can itself be a function of N and the channel transition probability, which then corresponds to tuning the quantizer to the channel at each specific N. Also note that when $\alpha = (N-1)/N$, the point density is continuous (for large N) as no codepoints are merged, whereas when $\alpha = 0$, there is only one code point at the centroid, and therefore the expected distortion is given by (4.10). Under this assumption, the code book can be equivalently represented as having $\alpha N+1$ points $\{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{\alpha N}, \bar{\mathbf{x}}_d\}$. The equivalent index transition probability matrix $P_{j|i}$ can then be expressed more compactly by

the $(N\alpha+1)\times(N\alpha+1)$ matrix

$$P=\begin{bmatrix} 1-(N-1)\epsilon(N) & \epsilon(N) & \dots & \epsilon(N) & N(1-\alpha)\epsilon(N) \\ \epsilon(N) & 1-(N-1)\epsilon(N) & \ddots & & N(1-\alpha)\epsilon(N) \\ \vdots & & & & \vdots \\ \epsilon(N) & \dots & & \epsilon(N) & 1-\alpha N\epsilon(N) \end{bmatrix}. \tag{4.11}$$

Notice that the above equivalent index transition probability is no longer simplex, because the point $\bar{\mathbf{x}}_d$ actually consists of $N(1-\alpha)$ codepoints, and when the source $\mathbf{x}$ lies in the quantization region for the point $\bar{\mathbf{x}}_d$ (denoted $\bar{\mathcal{R}}_{\bar{\mathbf{x}}_d}$), one of the $N(1-\alpha)$ indices corresponding to $\bar{\mathbf{x}}_d$ is randomly picked and sent across the channel. Since there are $N(1-\alpha)$ indices that can be received without making additional error due to the channel, with random index assignment, the probability that when one of the indices corresponding to the point $\bar{\mathbf{x}}_d$ is sent is received as any one of the indices corresponding to the same point $\bar{\mathbf{x}}_d$ is $1-\alpha N\epsilon(N)$. This is larger than the probability of correct index reception given by $1-(N-1)\epsilon(N)$ for all the other indices that have distinct corresponding codepoints.

4.4.2 Assumptions and Approximations

Recall that in the system model employed here, the transmitter and the receiver share a common codebook and the quantization regions are given by (4.1). With random index assignment, this implies that the quantization regions are also transmit optimized. The assumptions and approximations necessary for the development are as follows:

1. From Bennett [59]

   Assumption 1. *The number of distinct codepoints $\alpha N+1$ is large, such that the volumes $V(\bar{\mathcal{R}}_i)$ of the bounded cells $\bar{\mathcal{R}}_i$ are very small.*

   Assumption 2. *The source density function $f_{\mathbf{x}}(\mathbf{x})$ is smooth, so that Riemann sums approach Riemann integrals, and the mean value theorem of calculus applies.*

Assumption 3. *The total overload distortion is negligible. For example, all the probability is on a bounded set.*

2. From Lloyd [77]

   Assumption 4. *The specific point density (to be defined later) approaches a smooth point density function $\lambda(\mathbf{x})$ and N increases.*

3. From Gardner and Rao [78]

   Assumption 5. *The quantization cell $\bar{\mathcal{R}}_i$ is well approximated by a corresponding n-dimensional hyper-ellipsoid with the same volume as the quantization cell, whose shape is determined by the local sensitivity (Hessian) of the distortion function at $\hat{\mathbf{x}}_i$.*

For simplicity of presentation, the detailed derivation is relegated to the Appendices and the high-rate result is simply stated here.

4.4.3 Expected Distortion

The expected distortion is obtained by taking a triple expectation over the source distribution, the channel transition probabilities and the index assignments, as follows

$$E_d = \frac{1}{N!} \sum_{\pi} \sum_{i=1}^{N} \int_{\mathbf{x} \in \bar{\mathcal{R}}_i} \sum_{j=1}^{N} P_{\pi(j)|\pi(i)} d(\mathbf{x}, \hat{\mathbf{x}}_j) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}. \tag{4.12}$$

As pointed out earlier, the random index assignment converts any discrete memoryless channel into an equivalent SEC given by (4.8). Therefore, the expected distortion after averaging over the random index assignment is

$$E_d = \sum_{i=1}^{N} \int_{\mathbf{x} \in \bar{\mathcal{R}}_i} \sum_{j=1}^{N} P_{j|i} d(\mathbf{x}, \hat{\mathbf{x}}_j) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}, \tag{4.13}$$

where, the transition probability $P_{j|i}$ is given by (4.8). Let $\varphi_{(\alpha,N)} \triangleq \alpha N + 1$, and note that $1 \leq \varphi_{(\alpha,N)} \leq N$. It is shown in the Appendix 4.8.1 that

$$E_d = \int_{\mathbf{x}} E_{d,\mathbf{x}} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \tag{4.14}$$

where, $E_{d,\mathbf{x}}$, the expected distortion conditioned on the source instantiation $\mathbf{x}$, is

$$\begin{aligned} E_{d,\mathbf{x}} \approx\ & \varphi_{(\alpha,N)}\epsilon(N)\int d(\mathbf{x},\mathbf{y})\lambda(\mathbf{y})d\mathbf{y} \\ & +(N-\varphi_{(\alpha,N)})\epsilon(N)d(\mathbf{x},\bar{\mathbf{x}}_d) \\ & +\frac{\varphi_{(\alpha,N)}\epsilon(N)}{2(n+2)}\left(\frac{|D(\mathbf{x},\mathbf{x})|}{\varphi^2_{(\alpha,N)}\lambda^2(\mathbf{x})\kappa_n^2}\right)^{\frac{1}{n}}\mathrm{tr}\left(D^{-1}(\mathbf{x},\mathbf{x})\int D(\mathbf{x},\mathbf{y})\lambda(\mathbf{y})d\mathbf{y}\right) \\ & +\frac{n(1-N\epsilon(N))}{2(n+2)}\left(\frac{|D(\mathbf{x},\mathbf{x})|}{\varphi^2_{(\alpha,N)}\lambda^2(\mathbf{x})\kappa_n^2}\right)^{\frac{1}{n}} \\ & +\frac{\left(N-\varphi_{(\alpha,N)}\right)\epsilon(N)}{2(n+2)}\left(\frac{|D(\mathbf{x},\mathbf{x})|}{\varphi^2_{(\alpha,N)}\lambda^2(\mathbf{x})\kappa_n^2}\right)^{\frac{1}{n}}\mathrm{tr}\left(D^{-1}(\mathbf{x},\mathbf{x})D(\mathbf{x},\bar{\mathbf{x}}_d)\right) \quad (4.15) \end{aligned}$$

In the above equation, κ_n is the volume of an n-dimensional unit sphere, $D(\hat{\mathbf{x}},\hat{\mathbf{y}})$ is an $n\times n$ dimensional matrix with j,k-th element defined by

$$D_{k,j}(\hat{\mathbf{x}},\hat{\mathbf{y}}) = \frac{\partial^2 d(\mathbf{x},\hat{\mathbf{y}})}{\partial x_j \partial x_k}\Big|_{\mathbf{x}=\hat{\mathbf{x}}}, \quad (4.16)$$

and $\lambda(\mathbf{x})$ is the so-called *fractional point density*, and is defined as follows. The *specific point density* [76] is given by

$$\lambda_N(\mathbf{x}) \triangleq \frac{1}{NV(\bar{\mathcal{R}}_i)}, \text{if } \mathbf{x}\in\bar{\mathcal{R}}_i, \text{ for } i=1,2,\ldots,N. \quad (4.17)$$

Then, for large N, under Assumption 4, $\lambda_N(\mathbf{x})$ approximates a continuous non-negative density function $\lambda(\mathbf{x})$ having a unit integral.

The expected distortion in (4.15) is the sum of five terms. The first term represents the expected distortion when the source is quantized to one of the first αN distinct points in the codebook and there is an error in the channel, but the received codepoint is also one of the first αN codepoints. The second term is the distortion when the source is quantized to one of the first αN codepoints and there is no error in the channel. The third term is the distortion when the source is quantized to one of the first αN points in the codebook, but due to channel error it is received as $\bar{\mathbf{x}}_d$. The fourth term is the distortion when the source is quantized to $\bar{\mathbf{x}}_d$ but due to channel errors the received index corresponds to one of the first

αN codepoints. The last term is the distortion when the source is quantized to $\bar{\mathbf{x}}_d$, and the received index (with possible errors) corresponds to $\bar{\mathbf{x}}_d$ as well.

Note that when $\epsilon(N) = 0$, all but the fourth term in (4.15) drop out, and it is clear that that $\varphi_{(\alpha,N)} = N$ minimizes the distortion, i.e., for an error-free channel, no codepoints are merged, which agrees with conventional wisdom. With $\epsilon(N) = 0$ and $\varphi_{(\alpha,N)} = N$, (4.15) reduces to the classical high-rate distortion result in [78]. Also, one caveat in using (4.15) is that it is derived under the assumption that $\varphi_{(\alpha,N)}$ is large (a good rule of thumb is 2 or 3 bits per dimension). As seen earlier, as $\varphi_{(\alpha,N)} \to 1$, the distortion should approach the source distortion given by (4.10). However, the above expression does not reduce to (4.10) when $\varphi_{(\alpha,N)} \to 1$, as the assumption of $\varphi_{(\alpha,N)}$ being large is violated.

4.4.4 Variance of the Distortion over Index Assignments

When the index assignment is random, taking the expectation of the distortion over the source and channel statistics keeping the index assignment fixed yields a random variable that depends on the index assignment. Here, for simplicity, it is assumed that the point density is *continuous*, i.e., $\varphi_{(\alpha,N)} = N$, and the the rate at which the variance of the average distortion decreases as the number of quantization levels N becomes large is analyzed. A continuous point density would be optimal, for example, at low channel error rates or when the codebook employed is one designed for an ideal (i.e., error-free) channel. Define $\hat{d}_{ij}$ as

$$\hat{d}_{ij} \triangleq \int_{\mathbf{x}\in\bar{\mathcal{R}}_i} d(\mathbf{x}, \hat{\mathbf{x}}_j) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}, \tag{4.18}$$

and note that $\hat{d}_{ij}$ is bounded since the distortion was assumed to be a bounded function. Then, conditioned on the index assignment, the expected distortion can be written as

$$\begin{aligned} E_{d|\pi} &= \sum_{i,j=1}^{N} P_{\pi(j)|\pi(i)} \hat{d}_{ij} \\ &= \sum_{i\neq j} P_{\pi(j)|\pi(i)} \hat{d}_{ij} + \sum_{i} P_{\pi(i)|\pi(i)} \hat{d}_{ii}. \end{aligned} \tag{4.19}$$

One of the properties of a discrete symmetric channel is that $P_{\pi(i)|\pi(i)} = (1-(N-1)\epsilon(N))$ independent of i, and therefore, the second term above does not depend on the index assignment. Therefore, one can focus on the first term, which is denoted by $\tilde{E}_{d|\pi}$. It is shown in the Appendix 4.8.4 that

$$\mathrm{Var}(\tilde{E}_{d|\pi}) = O(1/N) + O(P_{\max}(N)), \tag{4.20}$$

where $P_{\max}(N) \triangleq \max_{1\leq i\leq N} P_i$, and P_i is the probability that $\mathbf{x} \in \bar{\mathcal{R}}_i$.

In other words, the variance of the distortion decreases at least as fast as either $\frac{1}{N}$ or $P_{\max}(N)$. If one assumes that the cell volumes decrease linearly with N, it is reasonable to expect that $P_{\max}(N)$ decreases as $1/N$. Then $\mathrm{Var}(\tilde{E}_{d|\pi})$ is upper bounded by $O(1/N)$, and since the expected distortion with random index assignment is at best given by $O(N^{\frac{-2}{n}})$, it is clear that at least for $n \geq 4$, the standard deviation of the average distortion goes to zero faster than the mean, i.e., most of the index assignments are asymptotically bad (or good). This also implies that random search based techniques would be inefficient in finding the best index assignment for large N.

4.5 Special Cases: $\varphi_{(\alpha,N)} = N$ and the MSE Distortion

The remainder of this chapter will be primarily concerned with the behavior of the expected distortion with random index assignment. Therefore, without loss of generality, consideration is restricted to the simplex error channel with transition probability given by (4.8). The high-rate distortion formula given by (4.15) is hard to directly interpret and optimize, hence, the development in this section is in two parts. For the first part, it is assumed that $\varphi_{(\alpha,N)} = N$, i.e., that no codepoints are merged at the centroid. In this case, it will be seen that at high rates, the expected distortion can be approximately expressed as the sum of the source quantization distortion and the channel error-induced distortion. In the second part, the assumption on $\varphi_{(\alpha,N)}$ is relaxed, but the distortion measure is assumed to be the MSE distortion. Again, it will be seen that the overall distortion

splits in two terms. Also, the point density is optimized to minimize the expected distortion for a wide range of channel error probabilities.

4.5.1 The $\varphi_{(\alpha,N)} = N$ Case

In this subsection, it is assumed that the point density has no singularity, i.e., $\varphi_{(\alpha,N)} = N$, and the distortion function is arbitrary. Then, (4.15) becomes

$$\begin{aligned} E_{d,\mathbf{x}} &\approx N\epsilon(N)\left\{\int_{\mathbf{y}} d(\mathbf{x},\mathbf{y})\lambda(\mathbf{y})d\mathbf{y} + \frac{N^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}}{2(n+2)}\lambda^{\frac{-2}{n}}(\mathbf{x})\left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n}}\right. \\ &\quad \left.\cdot \operatorname{tr}\left(D^{-1}(\mathbf{x},\mathbf{x})\left[\int_{\mathbf{y}}\left(D(\mathbf{x},\mathbf{y}) - D(\mathbf{x},\mathbf{x})\right)\lambda(\mathbf{y})d\mathbf{y}\right]\right)\right\} \\ &+ \frac{nN^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}}{2(n+2)}\lambda^{\frac{-2}{n}}(\mathbf{x})\left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n}}, \end{aligned} \tag{4.21}$$

where the last term is now the asymptotic distortion in the absence of channel errors (i.e., when $\epsilon(N) = 0$).

Asymptotic Performance

As N gets large, it is clear that the second term in (4.21) is always dominated by the first term, due to the presence of the $N^{\frac{-2}{n}}$ term. Therefore, this term can always be neglected in comparison to the first term as N gets large, and the high-rate distortion can be expressed as

$$E_{d,\mathbf{x}} \approx N\epsilon(N)\int_{\mathbf{y}} d(\mathbf{x},\mathbf{y})\lambda(\mathbf{y})d\mathbf{y} + \frac{nN^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}}{2(n+2)}\lambda^{\frac{-2}{n}}(\mathbf{x})\left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n}}, \tag{4.22}$$

where the first term is now the high-rate distortion purely due to *channel errors* whereas the second term is the high-rate distortion purely due to the *quantization error*. Thus, *for symmetric error channels, the asymptotic distortion can be expressed as the sum of the channel-error induced distortion and the source-quantization induced distortion for the class of bounded, twice differentiable distortion measures.* It is important to note, however, that this result does not imply the independence of the two sources of error in $\mathbf{x}$, namely, the error vector introduced

by the quantization and that introduced by channel errors. While this result is known for MSE distortion [81], to the best of the authors' knowledge, this is the first time that such a result has been shown for more general distortion measures. The above expression reduces to an expression in [68] for the case of a BSC with random index assignment and the MSE distortion.

Another important inference that can be drawn from (4.22) pertains to the trade off between the source code rate and the channel code rate. As N gets large, the quantization error decreases according to $N^{\frac{-2}{n}}$. Now, the behavior of $N\epsilon(N)$ depends on that of the sequence of channel codes used to transmit the index for each N. If $N\epsilon(N)$ decreases slower than $N^{\frac{-2}{n}}$, it implies that the channel code is inadequate in the sense that the overall distortion will eventually decrease according to the slower $N\epsilon(N)$ term, i.e., the distortion caused due to channel errors dominates the performance. On the other hand, if $N\epsilon(N)$ decreases faster than $N^{\frac{-2}{n}}$, it implies that the channel code is needlessly conservative, i.e., the transmitter can potentially save power by employing a slightly lower rate code. Thus, a sequence of channel codes is said to be *balanced with the source code* if $N\epsilon(N)$ decreases at a rate proportional to $N^{\frac{-2}{n}}$ for large N, because in this case the overall distortion also decreases at the rate $N^{\frac{-2}{n}}$ as N gets large.

Sensitivity of Conventional Source Coding to Channel Errors

With conventional source coding (i.e., source coding with no channel errors), the point density is chosen to minimize the second term in (4.22) subject to the constraint that $\int_{\mathbf{x}\in\mathcal{D}_{\mathbf{x}}} \lambda(\mathbf{x})d\mathbf{x} = 1$. The optimum point density can be found by applying Hölder's inequality as [54]

$$\lambda_{\text{conv}}(\mathbf{x}) = \frac{\left(|D(\mathbf{x},\mathbf{x})|^{\frac{1}{n}} f_{\mathbf{x}}(\mathbf{x})\right)^{\frac{n}{n+2}}}{\int_{\mathbf{x}\in\mathcal{D}_{\mathbf{x}}} \left(|D(\mathbf{x},\mathbf{x})|^{\frac{1}{n}} f_{\mathbf{x}}(\mathbf{x})\right)^{\frac{n}{n+2}} d\mathbf{x}}. \tag{4.23}$$

Substituting the above $\lambda_{\text{conv}}(\mathbf{x})$ into (4.22), the expected distortion can be computed. The following two simple examples illustrate this for MSE distortion, i.e.,

$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2$. One property of the MSE distortion is that $D(\hat{\mathbf{x}},\hat{\mathbf{y}}) = 2I_n$ for any $(\hat{\mathbf{x}},\hat{\mathbf{y}})$; therefore, the second term in (4.21) equals zero and drops out.

Example 1: The source vector $\mathbf{x}$ is n-dimensional i.i.d. Gaussian distributed with zero mean and unit variance. Then, the mean and variance of $\mathbf{x}$ are given by $m_{\mathbf{x}} = \underline{\mathbf{0}}$ and $\sigma^2_{\mathbf{x}} = n$ respectively. Also, it can be verified that $\lambda_{\text{conv}}(\mathbf{x})$ is an n-dimensional i.i.d. Gaussian density with zero mean and variance $(n+2)/n$ per dimension. Then, the expected distortion can be shown to be

$$E_d \;\doteq\; 2(n+1)N\epsilon(N) + 2\pi N^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}\left(\frac{n+2}{n}\right)^{\frac{n}{2}} \tag{4.24}$$

Example 2: The source vector $\mathbf{x}$ is n-dimensional i.i.d. uniformly distributed with each entry uniformly distributed on $[0,1]$. Then, $m_{\mathbf{x}} = 0.5\ \underline{1}$ where $\underline{1}$ is a vector of ones, and $\sigma^2_{\mathbf{x}} = n/12$. It can be verified that $\lambda_{\text{conv}}(\mathbf{x})$ is n-dimensional with i.i.d. entries uniformly distributed on $[0,1]$. Then, the expected distortion is

$$E_d \doteq \frac{nN\epsilon(N)}{6} + \frac{nN^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}}{n+2} \tag{4.25}$$

Optimization of the Point Density

Now consider optimization of the point density $\lambda(\mathbf{x})$ in (4.22), i.e., the problem of determining the $\lambda(\mathbf{x})$ that minimizes the expected distortion. With a slight abuse of notation, label the first and second terms of (4.22) as $E_d^{(1)}$ and $E_d^{(3)}$ respectively. From (4.22), using standard variational calculus [82], if a continuous point density exists, it is given by

$$\lambda_{\text{opt}}(\mathbf{x}) = \left[\frac{N^{-\frac{2}{n}}\kappa_n^{-\frac{2}{n}}\,|D(\mathbf{x},\mathbf{x})|^{\frac{1}{n}}\,f_{\mathbf{x}}(\mathbf{x})}{(n+2)\left(N\epsilon(N)\int d(\mathbf{x},\mathbf{y})f_{\mathbf{x}}(\mathbf{y})d\mathbf{y} + \mu\right)}\right]^{\frac{n}{n+2}}, \tag{4.26}$$

with $\mu > 0$ being a normalization constant. This optimum point density is valid for large N and is dependent on N, which is different from the classical notion of the point density. The above integral is clearly a monotonically decreasing function of μ, therefore, a continuous point density will not exist if, with $\mu = 0$, $\lambda_{\text{opt}}(\mathbf{x})$

integrates to a value less than 1. That is, a continuous point density exists if and only if,

$$\int \left[\frac{N^{-\frac{2}{n}} \kappa_n^{-\frac{2}{n}} |D(\mathbf{x},\mathbf{x})|^{\frac{1}{n}} f_{\mathbf{x}}(\mathbf{x})}{(n+2)\left(N\epsilon(N)\int d(\mathbf{x},\mathbf{y}) f_{\mathbf{x}}(\mathbf{y}) d\mathbf{y}\right)} \right]^{\frac{n}{n+2}} d\mathbf{x} > 1. \tag{4.27}$$

If we define I_* as

$$I_* \triangleq \left[\int \left(\frac{\kappa_n^{-\frac{2}{n}} |D(\mathbf{x},\mathbf{x})|^{\frac{1}{n}} f_{\mathbf{x}}(\mathbf{x})}{(n+2)\left(\int d(\mathbf{x},\mathbf{y}) f_{\mathbf{x}}(\mathbf{y}) d\mathbf{y}\right)} \right)^{\frac{n}{n+2}} d\mathbf{x} \right]^{\frac{n+2}{n}} \tag{4.28}$$

then a continuous point density does exists, provided

$$N^{\frac{n+2}{n}} \epsilon(N) < I_*. \tag{4.29}$$

Thus, the existence of a continuous point density depends on the relative rates of at which $N\epsilon(N)$ and $N^{-2/n}$ decrease with N as N gets large.There are three possibilities, assuming that $N\epsilon(N)$ is well-behaved as N gets large:

1. If $N\epsilon(N) = o\left(N^{\frac{-2}{n}}\right)$, the error is dominated by the $E_d^{(3)}$ term, i.e., channel errors play an insignificant role in the asymptotic distortion. In this case, asymptotically, the optimum point density is given by $\lambda_{\text{conv}}(\mathbf{x})$, and the distortion is given by the second term of (4.22), i.e.,

$$E_d \approx \frac{n N^{\frac{-2}{n}} \kappa_n^{\frac{-2}{n}}}{n+2} \left(\int_{\mathbf{x}\in\mathcal{D}_{\mathbf{x}}} \left(|D(\mathbf{x},\mathbf{x})|^{\frac{1}{n}} f_{\mathbf{x}}(\mathbf{x}) \right)^{\frac{n}{n+2}} d\mathbf{x} \right)^{\frac{n+2}{n}}. \tag{4.30}$$

2. If $N^{\frac{-2}{n}} = o\left(N\epsilon(N)\right)$, the error is dominated by the channel errors (i.e., the $E_d^{(1)}$ term above). Then, as N gets large, the optimum point density approaches a delta function at the centroid of the source $\bar{\mathbf{x}}_d$ given by (4.9), and the distortion is given by

$$E_d \approx N\epsilon(N) \int_{\mathbf{x}} d(\mathbf{x},\bar{\mathbf{x}}_d) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}. \tag{4.31}$$

   Note that in this case, since channel errors dominate the performance, the optimum quantizer avoids incurring additional distortion due to channel errors by collapsing all the codepoints onto $\bar{\mathbf{x}}_d$.

For example, with MSE distortion, when the index is mapped to a $L \triangleq \log_2(N)$ bit word and transmitted over a BSC with cross-over probability q, $\epsilon(N) = \left(1-(1-q)^L\right)/(N-1)$, and hence $N\epsilon(N) \to 1$ as N increases. Thus, the asymptotic distortion $E_d \approx \sigma_{\mathbf{x}}^2$ for large N, i.e., BSC with random index assignment is asymptotically "bad", as the distortion approaches the source variance, a fact that is in agreement with findings in [83].

3. If the approximation $N\epsilon(N) \approx c_n N^{\frac{-2}{n}}$ is valid for large N with $c_n < I_*$ given by (4.28), the optimum point density of (4.26) reduces to

$$\lambda_{\text{opt}}(\mathbf{x}, c_n) = \frac{\kappa_n^{\frac{-2}{n+2}} \left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n+2}} f_{\mathbf{x}}^{\frac{n}{n+2}}(\mathbf{x})}{\left(c_n\,(n+2)\int d(\mathbf{y},\mathbf{x}) f_{\mathbf{x}}(\mathbf{y})d\mathbf{y} + \mu\right)^{\frac{n}{n+2}}}, \tag{4.32}$$

where the normalization constant μ is chosen such that $\int \lambda_{\text{opt}}(\mathbf{x}, c_n)d\mathbf{x} = 1$. Additionally, if c_n^* can be chosen such that $c_n^* = I_*$, then the optimum point density can be written more simply as

$$\lambda_{\text{opt}}^*(\mathbf{x}) = \left[\frac{\alpha\left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n}} f_{\mathbf{x}}(\mathbf{x})}{\int d(\mathbf{y},\mathbf{x}) f_{\mathbf{x}}(\mathbf{y})d\mathbf{y}}\right]^{\frac{n}{n+2}}, \tag{4.33}$$

where α is a normalization constant. Comparing with (4.23), the optimum point density is matched to a "virtual" source with density

$$f_{\mathbf{x}}^{\text{virt}}(\mathbf{x}) = \frac{\beta f_{\mathbf{x}}(\mathbf{x})}{\int d(\mathbf{y},\mathbf{x}) f_{\mathbf{x}}(\mathbf{y})d\mathbf{y}}, \tag{4.34}$$

where β is a normalization constant. This is intuitively satisfying because the optimum quantizer compresses the points closer to the centroid of the source under $d(\mathbf{x},\mathbf{y})$.

4.5.2 Mean-Squared Error Distortion

In this subsection, it is assumed that the distortion function is the MSE, and point density can potentially have a singularity, i.e., $\varphi_{(\alpha,N)} \leq N$. When $d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2$ is the MSE function, $D(\mathbf{x},\mathbf{y}) = 2I_n$ regardless of $\mathbf{x}$ and $\mathbf{y}$, and

therefore the expected distortion given by (4.15) simplifies to

$$\begin{aligned} E_d &\approx \varphi_{(\alpha,N)}\epsilon(N)\int\int \|\mathbf{x}-\mathbf{y}\|^2\lambda(\mathbf{y})f_{\mathbf{x}}(\mathbf{x})d\mathbf{y}d\mathbf{x} \\ &\quad +(N-\varphi_{(\alpha,N)})\epsilon(N)\int \|\mathbf{x}-\bar{\mathbf{x}}_d\|^2 f_{\mathbf{x}}(\mathbf{x})d\mathbf{x} \\ &\quad +\frac{n}{n+2}\left(\kappa_n\varphi_{(\alpha,N)}\right)^{-\frac{2}{n}}\int \lambda^{-\frac{2}{n}}(\mathbf{x})f_{\mathbf{x}}(\mathbf{x})d\mathbf{x}, \end{aligned} \tag{4.35}$$

which can in turn be simplified to

$$\begin{aligned} E_d &\approx N\epsilon(N)\sigma_{\mathbf{x}}^2+\varphi_{(\alpha,N)}\epsilon(N)\left(\sigma_{\mathbf{y}}^2+\|m_{\mathbf{x}}-m_{\mathbf{y}}\|^2\right) \\ &\quad +\frac{n}{n+2}\varphi_{(\alpha,N)}^{-\frac{2}{n}}\kappa_n^{-\frac{2}{n}}\int \lambda^{-\frac{2}{n}}(\mathbf{x})f_{\mathbf{x}}(\mathbf{x})d\mathbf{x}, \end{aligned} \tag{4.36}$$

where $\varphi_{(\alpha,N)} \triangleq \alpha N+1$ as before, and $m_{\mathbf{x}}=\int \mathbf{x}f_{\mathbf{x}}(\mathbf{x})d\mathbf{x}$ and $m_{\mathbf{y}}=\int \mathbf{y}\lambda(\mathbf{y})d\mathbf{y}$ are the means of the source and the code-point locations respectively, and $\sigma_{\mathbf{x}}^2$ and $\sigma_{\mathbf{y}}^2$ are the variances of the source and the code-point distributions, respectively. The fact that, with the MSE distortion, $\bar{\mathbf{x}}_d = m_{\mathbf{x}}$, has also been used to obtain the above expression. When $n=1$ (scalar quantization) and $\varphi_{(\alpha,N)}=N$, the above expression reduces to similar expressions in [73, 81].

4.5.3 Optimization of the Point Density

A First Look at the Problem

Without loss of generality, one can let $m_{\mathbf{x}}=\underline{\mathbf{0}}$ to express the expected distortion as

$$\begin{aligned} E_d &\approx N\epsilon(N)\sigma_{\mathbf{x}}^2+(\alpha N+1)\,\epsilon(N)\left(\sigma_{\mathbf{y}}^2+\|m_{\mathbf{y}}\|^2\right) \\ &\quad +\frac{n}{n+2}\left(\alpha N+1\right)^{-\frac{2}{n}}\kappa_n^{-\frac{2}{n}}\int \lambda^{-\frac{2}{n}}(\mathbf{x})f_{\mathbf{x}}(\mathbf{x})d\mathbf{x}. \end{aligned} \tag{4.37}$$

Notice that the first term is independent of both the point density and α, and thus the expected distortion with the codebook structure assumed here is lower-bounded by $N\epsilon(N)\sigma_{\mathbf{x}}^2$ for simplex error channels (or for discrete memoryless channels with random index assignment). The last term in the above expression is minimized by the conventional point density (4.23). However, it is possible to reduce the overall

distortion by choosing a point density that has a smaller second moment than the conventional point density, as that would lead to a reduction in the second term. The above expression also shows the effect of choosing different values of $0 \leq \alpha \leq (N-1)/N$. A small value of α implies that the overall point density is highly singular. In this case, the second term, which involves the second moment of the point density, is small, but the third term, that depends on a negative power of $\alpha N + 1$, is relatively large. On the other hand, if we choose α close to $(N-1)/N$, the last term is small; however, the second term is large. The choice of $\alpha = (N-1)/N$ is therefore likely to be optimal only in the case where the second term is small, i.e., when $\epsilon(N)$ is small. Clearly, when $\epsilon(N) = 0$, $\alpha = (N-1)/N$ is optimum, which corresponds to employing a point density with no singularity at the origin. As the channel gets worse, $\epsilon(N)$ gets larger, and therefore the second term starts dominating the performance. In this case, $\alpha = (N-1)/N$ is no longer optimal, and one must employ a smaller value of α to balance the second and the third terms.

Optimum Point Density with MSE Distortion

While the qualitative arguments in the preceding subsection provide an intuitive feel for the relative importance of the different terms, the following analysis shows how to optimize the point density as well as how to optimally choose the value of α depending on the channel condition. Note that $\sigma_{\mathbf{y}}^2 + \|m_{\mathbf{y}}\|^2 = \int \mathbf{y}^T\mathbf{y}\lambda(\mathbf{y})d\mathbf{y}$, and for simplicity of notation let $\varphi \triangleq \alpha N+1$ and $g_n \triangleq n\kappa_n^{-\frac{2}{n}}/(n+2)$, and write the high-rate distortion as

$$E_d \approx g_n \varphi^{-\frac{2}{n}} \int \lambda^{-\frac{2}{n}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} + \varphi\epsilon(N) \int \mathbf{y}^T\mathbf{y}\lambda(\mathbf{y})d\mathbf{y} + N\epsilon(N)\sigma_{\mathbf{x}}^2. \qquad (4.38)$$

Note that $1 \leq \varphi \leq N$, and that φ is actually an integer. However, for the purposes of optimization, when N is large, φ can be considered to be a continuous variable, and since the function E_d is a continuous function of φ, it is reasonable to expect that the optimum discrete value of φ would be one of the nearby integers. Thus,

taking the partial derivative with respect to φ and equating to zero,

$$-\frac{2}{n} g_n \varphi^{-\frac{n+2}{n}} \int \lambda^{-\frac{2}{n}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} + \epsilon(N)\sigma_{\mathbf{y}}^2 = 0 \tag{4.39}$$

which yields

$$\varphi_{\text{opt}} = \left[\frac{2g_n \int \lambda^{-\frac{2}{n}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}}{n\epsilon(N)\sigma_{\mathbf{y}}^2} \right]^{\frac{n}{n+2}} \tag{4.40}$$

It is easily verified the second partial derivative is positive, i.e., φ_{opt} given above is indeed a local minimizer. Note that the above equation is valid provided $1 \leq \varphi_{\text{opt}} \leq N$. Otherwise, the optimum value of φ is likely to be one of the end-points, either 1 or N. When φ_{opt} is given by (4.40), the expected distortion becomes

$$\begin{aligned} E_d \;\approx\; & \frac{n+2}{2} \left(\frac{2g_n}{n} \right)^{\frac{n}{n+2}} \epsilon^{\frac{2}{n+2}}(N) \left(\int \mathbf{y}^T \mathbf{y} \lambda(\mathbf{y}) d\mathbf{y} \right)^{\frac{2}{n+2}} \left(\int \lambda^{-\frac{2}{n}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \right)^{\frac{n}{n+2}} \\ & + N\epsilon(N)\sigma_{\mathbf{x}}^2 \end{aligned} \tag{4.41}$$

From the above expression, the point density that minimizes the overall distortion must find the right trade-off between minimizing the second moment of the point density (the first integral) and being matched to the source density $f_{\mathbf{x}}(\mathbf{x})$ (i.e., minimizing the second integral). The point density that minimizes the second moment is a delta-function, whereas the point density that minimizes the second integral is the conventional point density. Therefore, *provided φ_{opt} in* (4.40) *satisfies $1 < \varphi_{opt} < N$, the optimum point density is independent of N and the channel behavior $\epsilon(N)$ for large N.* As N increases, however, the value of φ and therefore α would change, allowing the overall point density to have a larger or smaller fraction of the total codepoints at the centroid. Now, finding the point density that minimizes (4.41) directly is hard, so an indirect method is adopted here. Using the calculus of variations, the optimum point density $\lambda(\mathbf{x})$ subject to the constraints (positive, and integrates to unity) can be shown to be given by

$$\lambda_{\text{opt}}(\mathbf{x}) = \left[\frac{2g_n f_{\mathbf{x}}(\mathbf{x})}{n\left(\varphi^{\frac{n+2}{n}} \epsilon(N) \mathbf{x}^T \mathbf{x} + \mu \varphi^{\frac{2}{n}} \right)} \right]^{\frac{n}{n+2}} \tag{4.42}$$

where the normalization constant μ is chosen such that $\lambda_{\text{opt}}(\mathbf{x})$ integrates to 1. Again, it is easily verified that since the second partial derivative is positive, the above $\lambda_{\text{opt}}(\mathbf{x})$ is indeed a local minimizer. Now, from the analysis above, provided $1 < \varphi_{\text{opt}} < N$, the optimum point density is a function independent of N and $\epsilon(N)$. This is possible only if the value of φ varies with N such that

$$\frac{n\varphi^{\frac{n+2}{n}}\epsilon(N)}{2g_n} = K, \tag{4.43}$$

where K is a constant independent of N. Then, the term $M \triangleq n\mu\varphi^{\frac{2}{n}}/2g_n$ can be considered to be a new normalization constant, the $\lambda_{\text{opt}}(\mathbf{x})$ is independent of N.

To summarize, provided $1 < \varphi_{\text{opt}} < N$, the minimum expected MSE is

$$E_d^{\min} = g_n\varphi^{-\frac{2}{n}}\left[\int \lambda_{\text{opt}}^{-\frac{2}{n}}(\mathbf{x})f_{\mathbf{x}}(\mathbf{x})d\mathbf{x} + K\int \mathbf{y}^T\mathbf{y}\lambda_{\text{opt}}(\mathbf{y})d\mathbf{y}\right] + N\epsilon(N)\sigma_{\mathbf{x}}^2. \tag{4.44}$$

The optimum point density $\lambda_{\text{opt}}(\mathbf{x})$ is given by

$$\lambda_{\text{opt}}(\mathbf{x}) = \left[\frac{f_{\mathbf{x}}(\mathbf{x})}{K\mathbf{x}^T\mathbf{x} + M}\right]^{\frac{n}{n+2}}, \tag{4.45}$$

with M being a normalization constant. Surprisingly, it will be shown that $M = 0$ yields the optimum density, and that the value of K is chosen independent of N and $\epsilon(N)$ such that the term inside the square braces of (4.44) is minimized. Although K is independent of N or $\epsilon(N)$, both N and $\epsilon(N)$ affect the actual value of the overall distortion, as they should.

Discussion

Consider what happens as N is held fixed, and $\epsilon(N)$ starts at 0 and increases to its maximum possible value of $1/N$. When $\epsilon(N) = 0$, the second term in (4.38) drops out, and therefore the conventional point density given by (4.23) minimizes the overall distortion.

Let φ_{conv} denote the value of φ obtained from (4.40) by substituting $\lambda(\mathbf{x}) = \lambda_{\text{conv}}(\mathbf{x})$. Note that φ_{opt} obtained from (4.40) with the optimum point density necessarily satisfies $\varphi_{\text{opt}} \geq \varphi_{\text{conv}}$, since the $\lambda_{\text{conv}}(\mathbf{x})$ minimizes the numerator in (4.40), while the $\lambda_{\text{opt}}(\mathbf{x})$ allows the numerator to increase slightly while

reducing the denominator in order to minimize the expected distortion. Therefore, for small values of $\epsilon(N)$, if $\varphi_{\text{conv}} \geq N$, it is clear that φ_{opt} will remain fixed at its upper limit of N, i.e., no codepoints are merged at the centroid. For this range of $\epsilon(N)$, $\lambda_{\text{opt}}(\mathbf{x})$ is directly given by (4.42) with $\varphi = N$, which does depend on $\epsilon(N)$ and N. Thus, one critical value of $\epsilon(N)$ is when $\varphi_{\text{conv}} = N$, i.e.,

$$\epsilon_{\text{crit},1} = N^{-\frac{n+2}{n}} \left[\frac{2g_n \int \lambda_{\text{conv}}^{-\frac{2}{n}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}}{n \int \mathbf{x}^T \mathbf{x} \lambda_{\text{conv}}(\mathbf{x}) d\mathbf{x}} \right], \tag{4.46}$$

and for $0 \leq \epsilon(N) \leq \epsilon_{\text{crit},1}$, $\varphi_{\text{opt}} = N$ is satisfied.

Another critical value for $\epsilon(N)$ is when the normalization constant $\mu = 0$ in (4.42) with $\varphi = N$. This implies

$$\epsilon_{\text{crit},2} = N^{-\frac{n+2}{n}} \left[\int \left(\frac{2g_n f_{\mathbf{x}}(\mathbf{x})}{n\, \mathbf{x}^T \mathbf{x}} \right)^{\frac{n}{n+2}} d\mathbf{x} \right]^{\frac{n+2}{n}}, \tag{4.47}$$

and the optimum point density is given by

$$\lambda_{\text{opt}}(\mathbf{x}) = \left(\frac{f_{\mathbf{x}}(\mathbf{x})}{K \mathbf{x}^T \mathbf{x}} \right)^{\frac{n}{n+2}}, \tag{4.48}$$

where K is the normalization constant:

$$K = \left[\int \left(\frac{f_{\mathbf{x}}(\mathbf{x})}{\mathbf{x}^T \mathbf{x}} \right)^{\frac{n}{n+2}} d\mathbf{x} \right]^{\frac{n+2}{n}}. \tag{4.49}$$

It can be verified that at this critical value of $\epsilon_{\text{crit},2}$, $\lambda(\mathbf{x})$ given by (4.48) and $\varphi = N$ satisfy (4.40) and (4.42), i.e., the local optimality conditions are all satisfied. Since the critical value of $\epsilon_{\text{crit},2}$ in (4.47) is of the order $N^{-(n+2)/n}$, for sufficiently large N it will be smaller than $1/N$, the upper bound on $\epsilon(N)$. For $\epsilon_{\text{crit},2} \leq 1/N$ to hold, it is necessary that

$$N \geq \left[\int \left(\frac{2g_n f_{\mathbf{x}}(\mathbf{x})}{n \mathbf{x}^T \mathbf{x}} \right)^{\frac{n}{n+2}} d\mathbf{x} \right]^{\frac{n+2}{2}} \tag{4.50}$$

Thus, if $\epsilon_{\text{crit},2} \leq \epsilon(N) \leq 1/N$, even with $\mu = 0$, the point density function given by (4.42) integrates to a value less than 1. Then, it is necessary to make $\varphi < N$ in order to for $\lambda_{\text{opt}}(\mathbf{x})$ to be a valid point density. However, as already

observed, when $1 < \varphi < N$, the optimum point density is independent of N and $\epsilon(N)$. This implies that for this range of $\epsilon(N)$, the optimum point density is given by (4.48), and the optimum value of φ is obtained from (4.43) as

$$\varphi_{\text{opt}} = \left[\frac{2Kg_n}{n\epsilon(N)}\right]^{\frac{n}{n+2}} \tag{4.51}$$

It can be verified that this value of φ_{opt} and $\lambda_{\text{opt}}(\mathbf{x})$ satisfy (4.40) as well, thereby showing that the point density and φ obtained here are indeed the local minimizers.

This leaves the case $\epsilon_{\text{crit},1} < \epsilon(N) < \epsilon_{\text{crit},2}$. Since $\varphi_{\text{opt}} = N$ for $\epsilon(N) = \epsilon_{\text{crit},1}$ as well as for $\epsilon(N) = \epsilon_{\text{crit},2}$, it is likely that $\varphi_{\text{opt}} = N$ in the range $\epsilon_{\text{crit},1} < \epsilon(N) < \epsilon_{\text{crit},2}$. The optimality conditions (4.40) and (4.42) can be jointly solved numerically by simply sweeping φ over the range $1 \leq \varphi \leq N$, and for each value of φ, computing the normalization constant μ in (4.42), and substituting $\lambda_{\text{opt}}(\mathbf{x})$ in the overall distortion expression in (4.38), and finding the φ that attains the minimum distortion. It has been found that for the cases considered in this chapter, the minimum distortion is always attained when $\varphi = N$. Note that both $\epsilon_{\text{crit},1}$ and $\epsilon_{\text{crit},2}$ are of the order $N^{-\frac{n+2}{n}}$, i.e., they correspond to the case where the source and channel code rates are balanced.

One caveat is that as $\epsilon(N)$ gets closer to $1/N$, the value of φ_{opt} given by (4.51) becomes small, i.e., the assumption that φ is large no longer holds, since the number of free codepoints is no longer large. Then, the expression in (4.38) which is based on the high-rate assumption becomes inaccurate, however, the distortion with channel optimized quantization is always upper-bounded by $\sigma_{\mathbf{x}}^2$, i.e., the distortion obtained by having only one code point at the source centroid.

4.6 Simulation Results

4.6.1 Sensitivity of Conventional Source Coding to Channel Errors

For simplicity, consider an n-dimensional source $\mathbf{x}$ that is i.i.d. and uniformly distributed on $[-0.5, 0.5)$, with MSE as the performance metric. The quan-

tized index is mapped to a B bit word and sent over a BSC with cross-over probability q, and a random index assignment is employed. To verify the sensitivity of the performance of VQ to channel errors, the conventionally optimized codebook is generated using the Lloyd algorithm [77] for different values of n and B in the absence of channel errors. For training the Lloyd algorithm, as well as for evaluating the performance, 50,000 independent instantiations of $\mathbf{x}$ were employed.

Fig. 4.3 shows the MSE distortion versus the number of quantized bits B for the uniformly distributed random vector with dimension $n = 1, 2, 3$ and 4. The theoretical distortion is given by (4.25), with $\epsilon(N) = \left(1 - (1-q)^L\right) / (N-1)$, and q, the bit transition probability, fixed at 10^{-3}. From the bottom curve ($n = 1$), notice that as B is increased, the overall distortion initially decreases, and later starts to increase again. Asymptotically, the distortion would approach the channel error-induced distortion of 1/6. Fig. 4.4 shows the MSE distortion versus q, with $B = 5$, or $N = 32$ quantization levels. As in the previous figure, the four sets of curves correspond to $n = 1, 2, 3$ and 4 from bottom to top. For small values of q, the distortion with $N = 32$ is dominated by the source-quantization distortion term (i.e., the second term), whereas, as q increases, the inter-codepoint distortion (i.e., the first term) gets larger and the distortion finally approaches $n(1-0.5^B)/6$ as $q \to 0$. Fig. 4.5 shows the $E_d^{(1)}$ term only versus B, where $E_d^{(1)}$ represents the inter-codepoint distortion caused by channel errors only, i.e.,

$$E_d^{(1)} \approx \frac{nN\left(1 - (1-q)^B\right)}{6\,(N-1)}. \tag{4.52}$$

Notice that as B increases, the $E_d^{(1)}$ term approaches the source distortion of $n/6$. However, the above equation gives us the rate at which it approaches $n/6$ for a given q. Thus, from Fig. 4.6, which shows the $E_d^{(1)}$ term versus q with B fixed at 5 bits, we see that for large $q \approx 0.5$, the distortion approaches $n/6$. For small q, when B is large, the above equation is well approximated by $E_d^{(1)} \approx nBq/6$, i.e., $E_d^{(1)}$ increases linearly with q, as is evident from the graph.

It is also interesting to observe the rate B at which the high-rate approx-

imations become accurate. The distortion without channel errors is the second term in (4.25), and a good rule-of-thumb in source coding is that the high-rate results apply when we have about 2-3 bits per dimension. This can be seen, for example, from Fig. 4.3, when there are no channel errors. With channel errors, however, both the first and second terms (respectively, the $E_d^{(1)}$ and the $E_d^{(3)}$ terms) need to converge to their theoretical values as B increases. From Fig. 4.5, a good rule-of-thumb for the convergence of the $E_d^{(1)}$ term is about 3-4 bits per dimension, slightly higher than the corresponding number for the $E_d^{(3)}$ term. Also, as B gets large, the $E_d^{(1)}$ term will dominate the channel behavior. Therefore, the simulation results show that there is a range of B, for which, high rate results apply, while at the same time, the $E_d^{(1)}$ term does not yet completely dominate the performance. Were this not the case, i.e., if the $E_d^{(1)}$ term was always an order of magnitude larger than the $E_d^{(3)}$ term (or vice-versa), joint analysis of the $E_d^{(1)}$ term and the $E_d^{(3)}$ terms would have been a moot point. The simulation results thus show that there is a range of B for which the complete analysis can be usefully applied.

4.6.2 Optimization of the Point Density

For this subsection, the source is assumed to be a 2-dimensional Gaussian distributed random vector with zero mean and unit variance per dimension. The channel is modeled as a BSC with bit transition probability q and random index assignment, as before. The generalized Lloyd algorithm described in [66] is used to generate a codebook of *channel optimized* vectors with MSE as the distortion metric. Figs. 4.7 and 4.8 plot the expected MSE distortion performance versus the number of quantized bits B and the bit transition probability q, respectively. The plots show the improvement in MSE performance that can be obtained by using an optimum codebook (compared with the curves obtained using the conventional codebook, labeled "unopt-CB"). Also, in Fig. 4.8, the two values of q corresponding to $\epsilon(N) = \epsilon_{\text{crit},1}$ and $\epsilon(N) = \epsilon_{\text{crit},2}$ are also plotted, to show that the simulation results agree with the theory over a wide range of values of q. Also, when $q > q_{\text{crit},2}$,

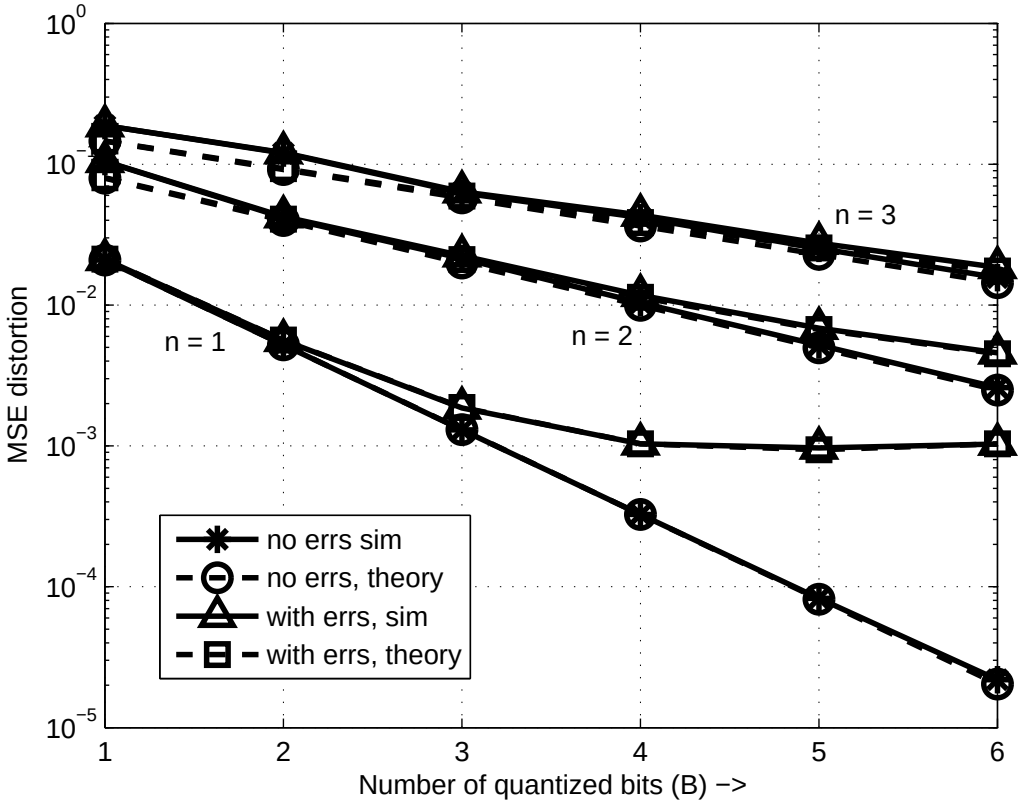


Figure 4.3: MSE distortion versus number of quantized bits B, for a uniformly distributed random vector and index sent over the BSC with bit transition probability $q = 10^{-3}$. The codebook is generated using the conventional Lloyd algorithm with 10,000 training vectors. The theoretical curves are generated using (4.25).

the optimum point density is singular, i.e., $\varphi < N$.

Table 4.1 compares the theoretical and simulation-based values of φ for different BSC bit transition probabilities q and number of codepoints N, which shows that the theoretical and experimental values of φ match closely. For the experimental results, the φ was computed as the difference between the total number of codepoints N and the number of codepoints whose Voronoi cells were empty in the Lloyd algorithm. Also shown in the table is the expected MSE distortion. It is interesting to note that when φ is close to N, i.e., when only a few codepoints are

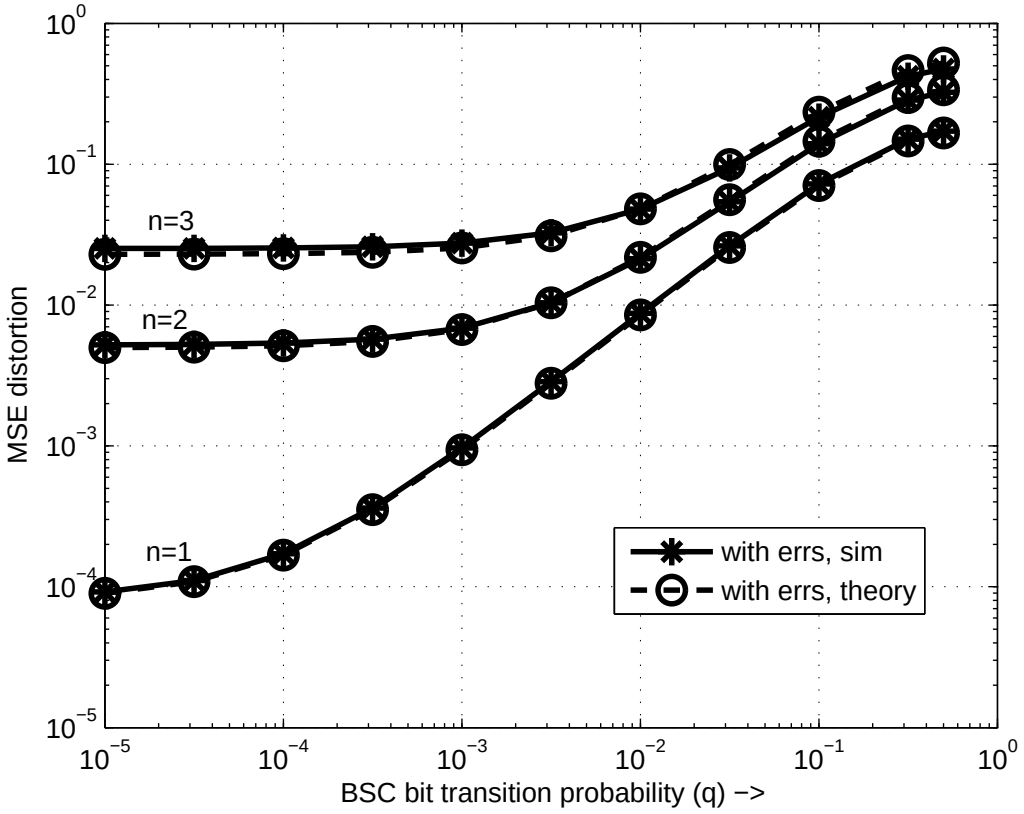
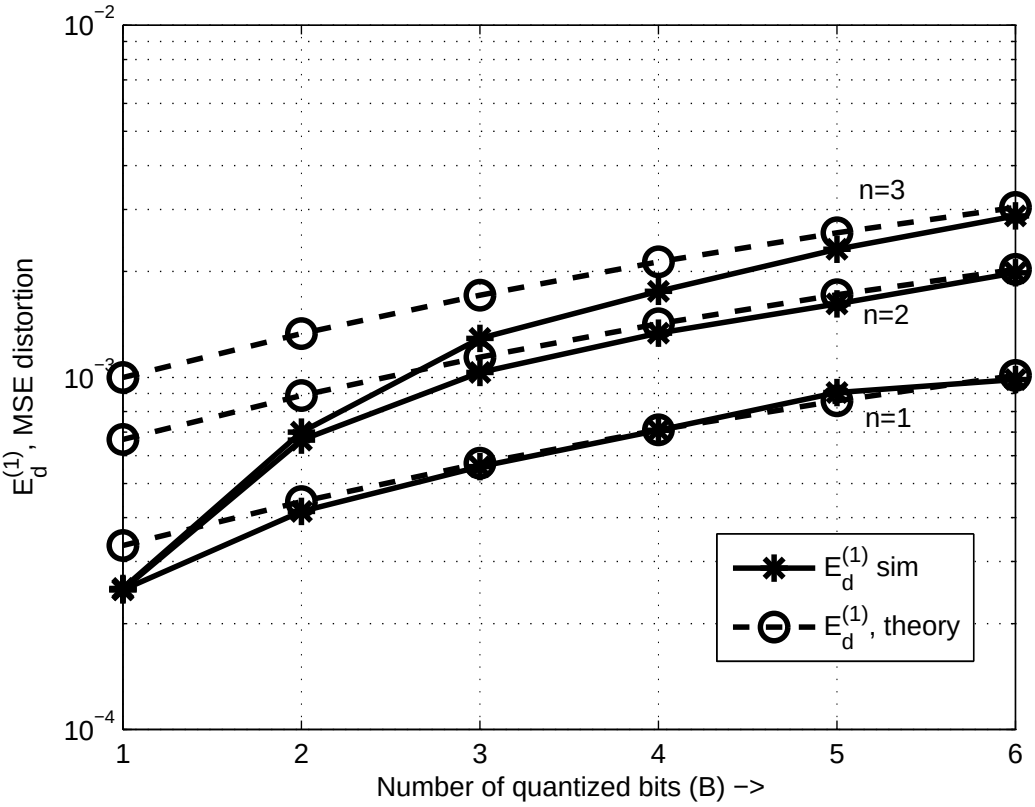


Figure 4.4: MSE distortion for a uniformly distributed random vector with the conventional point density, and the number of quantization bits B fixed at 5 bits. The quantized index is sent over a BSC with bit transition probability q (the x-axis). The theoretical curves are generated using (4.25).

merged, the performance improves when N is increased with q being kept fixed. For example, when $q = 0.01$, the performance improves as N is increased from 32 to 128. On the other hand, at a larger error rate, when a half or more of the code-points are merged at the centroid, little performance improvement can be achieved by increasing N. This is seen, for example, when $q = 0.05$, where the expected distortion is about 0.81 regardless of the value of N. From (4.47), the critical value of the bit transition probability $q_{\text{crit},2}$ beyond which $\varphi_{\text{opt}} < N$ can be shown to be approximately proportional to $2^{-B}/B$, and from the table, one can see that this



critical value is roughly at $q = 0.02$ when $N = 32$, at $q = 0.01$ when $N = 64$ and at a value slightly lower than 0.005 when $N = 128$, in agreement with the theory. Comparing the distortion when $N = 32$ and $q = 0.02$ with that when $N = 64$ and $q = 0.05$ (or when $N = 64$ and $q = 0.01$ with that when $N = 128$ and $q = 0.02$), both cases have the same number of effective codepoints, but expected distortion is larger with the larger N. From this, it is tempting to think that one could lower the distortion by lowering N when φ is of the order $N/2$ or smaller. However, this is only true if the q decreases adequately when the power is re-assigned to the fewer

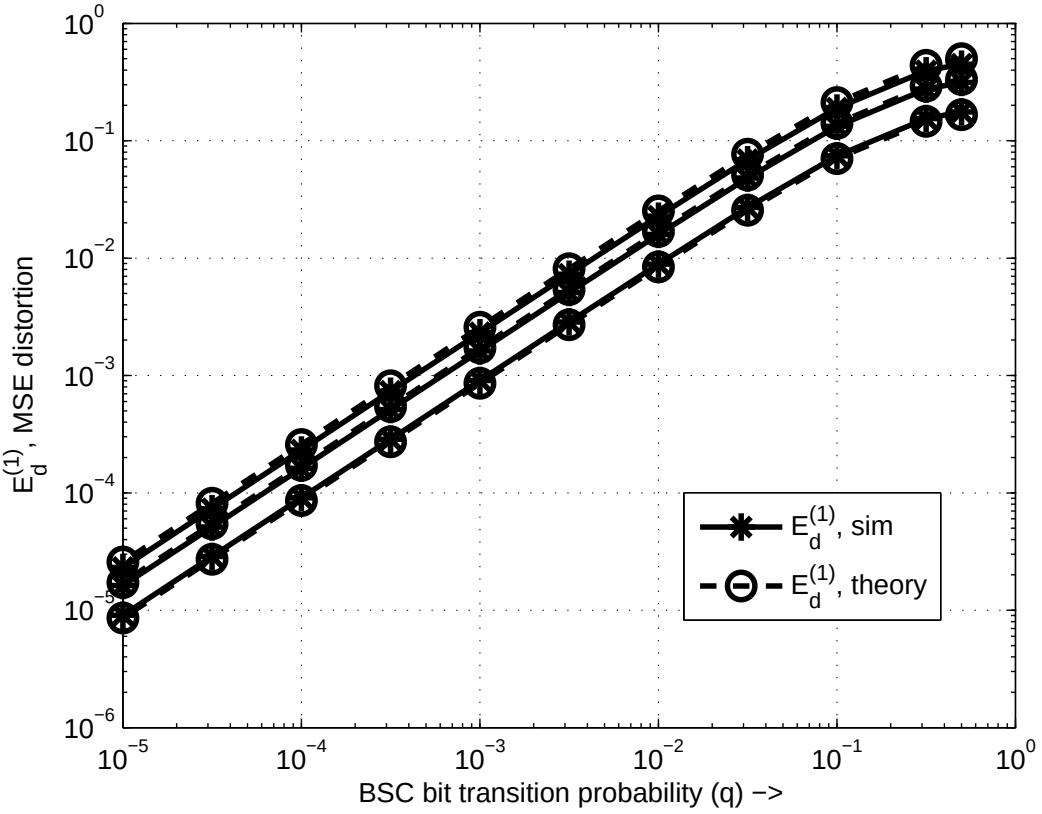


Figure 4.6: Inter-codepoint MSE distortion term $E_d^{(1)}$ for a uniformly distributed random vector, versus the BSC bit transition probability q. The number of quantization bits B is fixed at 5. The theoretical curves are generated using (4.52).

number of bits. For example, if $q = 0.05$ when $N = 64$ implies that by dropping one bit and going down to $N = 32$, the bit transition probability can improve to $q = 0.02$ or better, then it is true that one could improve performance by using the codebook with smaller N.

Finally, note that although the expressions for the expected distortion were derived using the general distortion function $d(\mathbf{x}, \mathbf{y})$, simulation results have been provided only for the MSE distortion case. This is mainly because the solution to the weighted centroid step of the generalized Lloyd algorithm is known in closed form for the MSE distortion (it is simply a weighted geometric centroid). More

Table 4.1: Experimental and Theoretical Values of φ for different N and q. The tuples correspond to $(\varphi_{\text{exp}}, \varphi_{\text{theory}})$, for a 2-dimensional standard Gaussian random vector. The number below the tuple is E_d. φ_{theory} is computed from (4.51).

$N\backslash q$	0	0.0020	0.0050	0.0100
32	$(32, 32)$ 0.1136	$(32, 32)$ 0.1587	$(32, 32)$ 0.2216	$(32, 32)$ 0.3150
64	$(64, 64)$ 0.0600	$(64, 64)$ 0.1153	$(64, 64)$ 0.1869	$(63, 57)$ 0.2867
128	$(128, 128)$ 0.0309	$(128, 128)$ 0.0941	$(114, 107)$ 0.1674	$(82, 76)$ 0.2670
$N\backslash q$	0.0200	0.0500	0.1000	0.1200
32	$(32, 31)$ 0.4753	$(24, 20)$ 0.8156	$(17, 15)$ 1.2092	$(16, 14)$ 1.3399
64	$(45, 41)$ 0.4472	$(31, 27)$ 0.8132	$(22, 20)$ 1.2320	$(20, 19)$ 1.3538
128	$(60, 52)$ 0.4337	$(37, 36)$ 0.8169	$(27, 27)$ 1.2547	$(22, 25)$ 1.3849

general distortion functions will be explored in future publications.

4.7 Conclusions

In this chapter, the source quantization problem when the quantized index is sent over a noisy discrete symmetric error channel before being reproduced at the receiver was considered. When random index assignment is employed, the channel transition probability becomes simplex, and in this case, a theoretical analysis of the asymptotic performance with channel errors and arbitrary distortion functions was presented. Further, when the distortion is measured as the mean-squared error, it was demonstrated that the distortion is given by the sum of the distortion due to inter-codepoint distortion (i.e., purely due to channel errors) and the representation error in quantizing the source using a finite number of bits. The rate of decay of the two terms as the number of quantization levels N increases can be different,

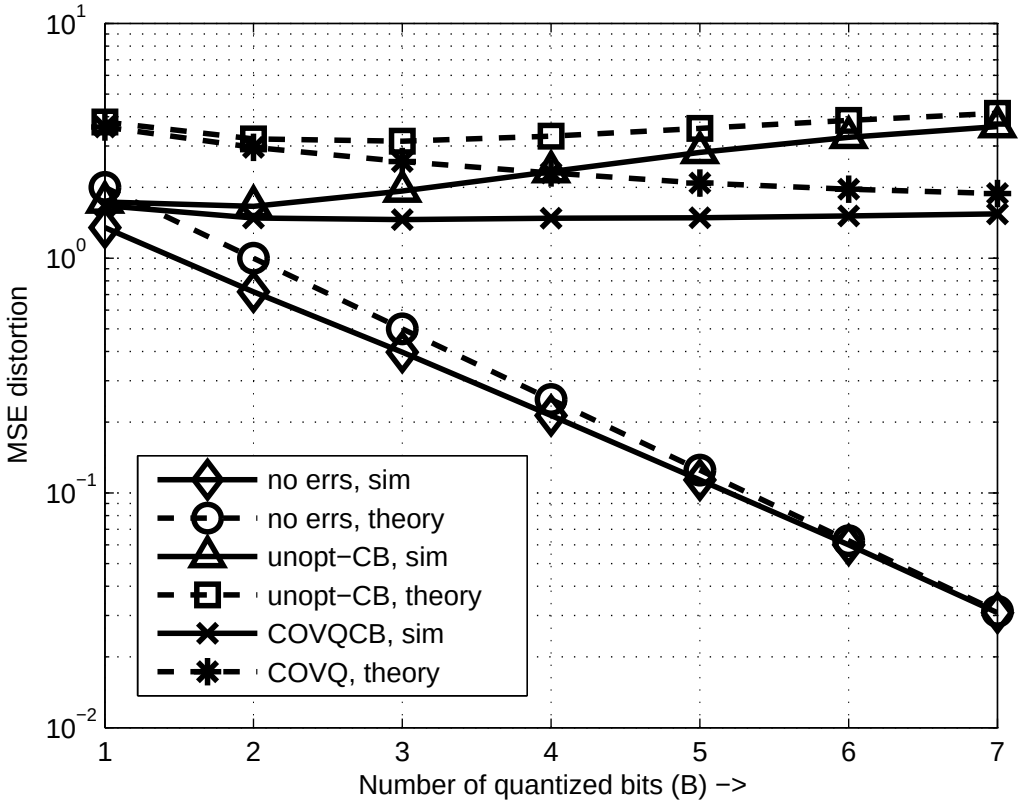


Figure 4.7: MSE distortion versus number of quantized bits B, for a 2-dimensional standard Gaussian random vector and index sent over the BSC with bit transition probability $q = 0.1$.

in which case, one of the two terms will dominate as N gets large. Also, a novel theoretical analysis of the optimum *singular* point density that minimizes the overall distortion was derived, and its MSE performance evaluated. The accuracy of the theoretical results were illustrated through Monte-Carlo simulations.

4.8 Appendix

4.8.1 Proof of (4.15)

For ease of presentation, the proof of (4.15) will be presented in two parts. First, it will be assumed that $\varphi = N$ and that the point density is non-singular. Next, analysis is extended to arbitrary φ , i.e., to a singular point density where $N - \varphi$ codepoints are merged at the centroid $\bar{\mathbf{x}}_d$.

4.8.2 The $\varphi = N$ Case

When $\varphi = N$, and no particular structure is assumed for the codebook, the expected distortion is obtained by averaging over the source statistics and the channel transition probabilities as

$$\begin{aligned} E_d &= \sum_{i=1}^{N} \int_{\mathbf{x}\in\bar{\mathcal{S}}_i} \sum_{j=1}^{N} P_{j|i} d(\mathbf{x}, \hat{\mathbf{x}}_j) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ &\overset{a}{\approx} \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1}^{N} P_{j|i} \int_{\mathbf{x}\in\bar{\mathcal{S}}_i} d(\mathbf{x}, \hat{\mathbf{x}}_j) d\mathbf{x}, \end{aligned} \tag{4.53}$$

where the approximation 'a' is valid when Bennett's Assumptions 1, 2 and 3 hold. Let $\mathbf{x} \in \bar{\mathcal{S}}_i$ be written as $\mathbf{x} = \hat{\mathbf{x}}_i + \mathbf{e}$, and note that for high-rate quantization, $\mathbf{e}$ is a "small" vector since $\bar{\mathcal{S}}_i$ is regular. From the Taylor series expansion,

$$d(\mathbf{x}, \hat{\mathbf{x}}_j) = d(\hat{\mathbf{x}}_i + \mathbf{e}, \hat{\mathbf{x}}_j) = d(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) + \mathbf{g}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j)\mathbf{e} + \frac{1}{2}\mathbf{e}^T D(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j)\mathbf{e} + O(\|\mathbf{e}\|^3), \tag{4.54}$$

where the gradient $\mathbf{g}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is an $1 \times n$ vector with j-th element

$$\mathbf{g}_j(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \triangleq \frac{\partial d(\mathbf{x}, \hat{\mathbf{y}})}{\partial x_j}\Big|_{\mathbf{x}=\hat{\mathbf{x}}}, \tag{4.55}$$

where x_j is the j-th element of $\mathbf{x}$, and $D(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is the n by n Hessian matrix defined in (4.16). Note that from the definition of the distortion function, $d(\hat{\mathbf{x}}, \hat{\mathbf{x}}) = 0$ for any $\hat{\mathbf{x}}$, and since $\hat{\mathbf{y}} = \arg\min_{\mathbf{y}} d(\mathbf{y}, \hat{\mathbf{x}})$, $\mathbf{g}(\hat{\mathbf{x}}, \hat{\mathbf{x}}) = \underline{\mathbf{0}}$ for any $\hat{\mathbf{x}}$ as well. This fact will be used later to simplify the distortion expressions. Also, when $\hat{\mathbf{x}}_i = \hat{\mathbf{x}}_j$ in (4.54),

it reduces to the Taylor expansion used in [78]. Substituting the Taylor expansion (4.54) in the expression for E_d in (4.53):

$$\begin{aligned} E_d &\approx \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1}^{N} P_{j|i} \\ &\quad \cdot \int_{\mathbf{e}\in\bar{\mathcal{E}}_i} \left(d(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) + \mathbf{g}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j)\mathbf{e} + \frac{1}{2}\mathbf{e}^T D(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j)\mathbf{e} \right) d\mathbf{e} \end{aligned} \tag{4.56}$$

where $\bar{\mathcal{E}}_i \triangleq \{\mathbf{e} : \mathbf{e} + \hat{\mathbf{x}}_i \in \bar{\mathcal{S}}_i\}$, i.e., the Voronoi region corresponding to the i-th code point, but shifted to the origin. This is valid because the Jacobian of the transformation $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}_i$ is the identity matrix.

Consider the first term in (4.56), i.e.,

$$\begin{aligned} E_d^{(1)} &\triangleq \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1}^{N} P_{j|i} \int_{\mathbf{e}\in\bar{\mathcal{E}}_i} d(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) d\mathbf{e} \\ &= \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1}^{N} P_{j|i}\, d(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) \int_{\mathbf{e}\in\bar{\mathcal{E}}_i} d\mathbf{e} \\ &= \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1}^{N} P_{j|i}\, d(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) V_{\bar{\mathcal{E}}_i} \end{aligned} \tag{4.57}$$

Substituting for $P_{j|i}$ from (4.8),

$$E_d^{(1)} = \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1}^{N} \epsilon(N)\, d(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) V_{\bar{\mathcal{E}}_i} \tag{4.58}$$

where the fact that $d(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i) = 0$ has been implicitly used to simplify the above expression by including the $j = i$ term in the inner summation. Now, for large N, under Assumption 2, the above expression approximates the Riemann integral

$$E_d^{(1)} = \int_{\mathbf{x}} f_{\mathbf{x}}(\mathbf{x}) \sum_{j=1}^{N} \epsilon(N)\, d(\mathbf{x}, \hat{\mathbf{x}}_j) d\mathbf{x}. \tag{4.59}$$

Next, recall that volume of the region $\bar{\mathcal{S}}_i$ (or $\bar{\mathcal{E}}_i$) is approximately given by

$$V(\bar{\mathcal{E}}_i) \approx \frac{1}{N\lambda(\hat{\mathbf{x}}_i)}, \tag{4.60}$$

where $\lambda(\mathbf{x})$ is the point density. The next proposition relates a summation of a function of code-point locations to the point density.

Proposition 1. *If the point generating density is $\lambda(\mathbf{x})$, for any continuous, bounded function $\beta(\hat{\mathbf{x}}_i)$ of the code-point location,*

$$S_\beta \triangleq \frac{1}{N}\sum_{i=1}^{N}\beta(\hat{\mathbf{x}}_i) \approx \int_{\mathbf{y}}\beta(\mathbf{y})\lambda(\mathbf{y})d\mathbf{y} \tag{4.61}$$

Proof. First, note that the summation in S_β is the same as taking the expectation of the random variable $\beta(\mathbf{x})$, where $\mathbf{x}$ is uniformly distributed over the codebook $\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_N\}$. From (4.17),

$$\begin{aligned} S_\beta &= \sum_{i=1}^{N}\beta(\hat{\mathbf{x}}_i)\lambda_N(\hat{\mathbf{x}}_i)V(\bar{\mathcal{S}}_i) \\ &= \sum_{i=1}^{N}\beta(\hat{\mathbf{x}}_i)\int_{\mathbf{x}\in\bar{\mathcal{S}}_i}\lambda_N(\mathbf{x})d\mathbf{x}. \end{aligned} \tag{4.62}$$

Also, note that $\int_{\mathbf{y}}\beta(\mathbf{y})\lambda(\mathbf{y})d\mathbf{y} = \sum_{i=1}^{N}\int_{\mathbf{y}\in\bar{\mathcal{S}}_i}\beta(\mathbf{y})\lambda(\mathbf{y})d\mathbf{y}$. The lemma is now established by bounding the absolute value of the difference between the expressions on either side of (4.61) by an expression that goes to 0 as $N \to \infty$. Assume that $|\beta(\mathbf{x})| \leq B$ for all $\mathbf{x} \in \mathcal{D}_{\mathbf{x}}$. Thus,

$$\begin{aligned} D_N &\triangleq \left|\sum_{i=1}^{N}\beta(\hat{\mathbf{x}}_i)\int_{\mathbf{x}\in\bar{\mathcal{S}}_i}\lambda_N(\mathbf{x})d\mathbf{x} - \sum_{i=1}^{N}\int_{\mathbf{y}\in\bar{\mathcal{S}}_i}\beta(\mathbf{y})\lambda(\mathbf{y})d\mathbf{y}\right| \\ &\leq \sum_{i=1}^{N}\int_{\mathbf{x}\in\bar{\mathcal{S}}_i}\left|\beta(\hat{\mathbf{x}}_i)\lambda_N(\mathbf{x}) - \beta(\mathbf{x})\lambda(\mathbf{x})\right|d\mathbf{x} \\ &\leq \sum_{i=1}^{N}\int_{\mathbf{x}\in\bar{\mathcal{S}}_i}\Big[\left|\beta(\hat{\mathbf{x}}_i)\left(\lambda_N(\mathbf{x}) - \lambda(\mathbf{x})\right)\right| + \left|\lambda(\mathbf{x})\left(\beta(\hat{\mathbf{x}}_i) - \beta(\mathbf{x})\right)\right|\Big]d\mathbf{x} \end{aligned}$$

Since $\beta(\mathbf{x})$ is bounded and continuous, one can substitute $\beta(\hat{\mathbf{x}}_i) \leq B$ in the first term, and $|\beta(\hat{\mathbf{x}}_i) - \beta(\mathbf{x})| \leq B\delta_N$ in the second term, where δ_N is the cell diameter defined in (4.2). Thus,

$$\begin{aligned} D_N &\leq B\sum_{i=1}^{N}\int_{\mathbf{x}\in\bar{\mathcal{S}}_i}\left|\lambda_N(\mathbf{x}) - \lambda(\mathbf{x})\right|d\mathbf{x} + B\delta_N\sum_{i=1}^{N}\int_{\mathbf{x}\in\bar{\mathcal{S}}_i}\lambda(\mathbf{x})d\mathbf{x} \\ &\leq B\int_{\mathcal{D}_{\mathbf{x}}}\left|\lambda_N(\mathbf{x}) - \lambda(\mathbf{x})\right|d\mathbf{x} + B\delta_N \end{aligned} \tag{4.63}$$

Thus, from the assumption of diminishing cell diameters, $\delta_N \to 0$, and by Scheffé's theorem [84], $\int_{\mathcal{D}_{\mathbf{x}}}\left|\lambda_N(\mathbf{x}) - \lambda(\mathbf{x})\right|d\mathbf{x} \to 0$; as $N \to \infty$. □

Substituting (4.61) in (4.59), the first term finally reduces to

$$E_d^{(1)} = N\epsilon(N) \int_{\mathbf{x}} \int_{\mathbf{y}} d(\mathbf{x},\mathbf{y}) f_{\mathbf{x}}(\mathbf{x}) \lambda(\mathbf{y}) d\mathbf{y} d\mathbf{x}. \tag{4.64}$$

Now, consider the second $(\mathbf{g}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j)\mathbf{e})$ term:

$$E_d^{(2)} = \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1}^{N} P_{j|i} \mathbf{g}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) \int_{\mathbf{e}\in\bar{\mathcal{E}}_i} \mathbf{e} d\mathbf{e} \tag{4.65}$$

Later, it will be shown that under the quantization cell Approximation 3, the above term goes to zero. However, to get a more concrete feel for the contribution of $E_d^{(2)}$ to the expected distortion, given a convex polytope $\bar{\mathcal{E}}$ in $\mathbb{R}^n$, define the *normalized mean vector* $\mathbf{w}_{\bar{\mathcal{E}}}$ as

$$\mathbf{w}_{\bar{\mathcal{E}}} \triangleq \frac{\int_{\bar{\mathcal{E}}} \mathbf{e} d\mathbf{e}}{\left[V(\bar{\mathcal{E}})\right]^{1+1/n}}, \tag{4.66}$$

and note that $\mathbf{w}_{\bar{\mathcal{E}}}$ depends only on the shape and orientation, but not the volume of $\bar{\mathcal{E}}$. That is, $\mathbf{w}_{\alpha\bar{\mathcal{E}}} = \mathbf{w}_{\bar{\mathcal{E}}}$ holds for $\alpha > 0$, where the polytope $\alpha\bar{\mathcal{E}} \triangleq \{\alpha\mathbf{e} : \mathbf{e} \in \bar{\mathcal{E}}\}$. Then, (4.65) can be written as

$$\begin{aligned} E_d^{(2)} &= \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1}^{N} P_{j|i} \mathbf{g}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) \mathbf{w}_{\bar{\mathcal{E}}}(\hat{\mathbf{x}}_i) \left[V(\bar{\mathcal{E}}_i)\right]^{1/n} V(\bar{\mathcal{E}}_i) \\ &= \epsilon(N) \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1}^{N} \mathbf{g}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) \mathbf{w}_{\bar{\mathcal{E}}}(\hat{\mathbf{x}}_i) \left[V(\bar{\mathcal{E}}_i)\right]^{1/n} V(\bar{\mathcal{E}}_i) \end{aligned} \tag{4.67}$$

where (4.67) is obtained by substituting for $P_{j|i}$ from (4.8) and recalling that $\mathbf{g}(\hat{\mathbf{x}}, \hat{\mathbf{x}}) = \underline{\mathbf{0}}$. Also, note that the argument $\hat{\mathbf{x}}_i$ has been included in writing $\mathbf{w}_{\bar{\mathcal{E}}}(\hat{\mathbf{x}}_i)$, since the normalized mean vector depends on the orientation of the Voronoi region corresponding to $\hat{\mathbf{x}}_i$. As N goes to infinity, this can be expected (or rather, hypothesized) to become a smooth function $\mathbf{w}_{\bar{\mathcal{E}}}(\mathbf{x})$. Substituting (4.60) in the $\left[V(\bar{\mathcal{E}}_i)\right]^{1/n}$ term of (4.67), and approximating the summations by integrals,

$$\begin{aligned} E_d^{(2)} &= N\epsilon(N) \int_{\mathbf{x}} f_{\mathbf{x}}(\mathbf{x}) \left(\int_{\mathbf{y}} \lambda(\mathbf{y}) \mathbf{g}(\mathbf{x},\mathbf{y}) d\mathbf{y} \right) \mathbf{w}_{\bar{\mathcal{E}}}(\mathbf{x}) N^{\frac{-1}{n}} \lambda^{\frac{-1}{n}}(\mathbf{x}) d\mathbf{x} \\ &= N^{\frac{-1}{n}} N\epsilon(N) \left(\int_{\mathbf{x}} \int_{\mathbf{y}} \lambda^{\frac{-1}{n}}(\mathbf{x}) \lambda(\mathbf{y}) \mathbf{g}(\mathbf{x},\mathbf{y}) \mathbf{w}_{\bar{\mathcal{E}}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{y} d\mathbf{x} \right), \end{aligned} \tag{4.68}$$

where (4.61) has been used to approximate the inner summation in (4.67) by the corresponding integral. The above equation is significant because it explicitly shows the dependence of $E_d^{(2)}$ on N. Thus, as N increases, the $E_d^{(2)}$ term decreases at a *higher rate* than $E_d^{(1)}$ which drops off at the rate $N\epsilon(N)$. Thus, the $E_d^{(2)}$ term can always be neglected compared to $E_d^{(1)}$ as N gets large. Moreover, it will be shown that under the quantization cell Approximation 3, the $E_d^{(2)}$ term in fact drops out because the approximated regions are symmetric about the origin (hence, $\mathbf{w}_{\bar{\mathcal{E}}}(\mathbf{x}) \approx \underline{\mathbf{0}}$). Therefore, $E_d^{(2)}$ will be simply dropped from the analysis. Finally, consider the third term in (4.56), i.e.,

$$E_d^{(3)} \triangleq \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1}^{N} P_{j|i} \int_{\mathbf{e}\in\bar{\mathcal{E}}_i} \frac{1}{2}\mathbf{e}^T D(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j)\mathbf{e}\, d\mathbf{e} \tag{4.69}$$

In order to evaluate the above integral, one needs to know the Voronoi region $\bar{\mathcal{E}}_i$, which is generally complicated. Hence, consider the following quantization cell approximation, which is standard in source coding literature [78,85]. At high rate, the quantization cells are well approximated by the n-dimensional hyper-ellipsoid,

$$\bar{\mathcal{E}}_i \approx \hat{\mathcal{E}}_i \triangleq \mathcal{T}(\underline{\mathbf{0}}, M(\hat{\mathbf{x}}_i), V_{\bar{\mathcal{E}}_i}) \tag{4.70}$$

where, $V_{\bar{\mathcal{E}}_i}$ is the volume of the i-th Voronoi region, $M(\hat{\mathbf{x}}_i)$ is as described below, and the hyper-ellipsoidal region $\mathcal{T}(\hat{\mathbf{x}}, M, v)$ is defined as

$$\mathcal{T}(\hat{\mathbf{x}}, M, v) \triangleq \left\{ \mathbf{x} \middle| \left(\frac{\kappa_n}{v^2\,|M|} \right)^{1/n} (\mathbf{x}-\hat{\mathbf{x}})^T M\,(\mathbf{x}-\hat{\mathbf{x}}) \leq 1 \right\} \tag{4.71}$$

and κ_n is the volume of an n-dimensional unit sphere. Note that this approximation is "strict", that is, the approximation error does not go to zero as N goes to infinity, because the approximated regions cannot form a Dirichlet partition. A good explanation of this approximation can be found in [85]. Also note that under the quantization cell approximation, $\mathbf{w}_{\bar{\mathcal{E}}}(\mathbf{x}) = \underline{\mathbf{0}}$, i.e., the $E_d^{(2)}$ term is zero.

The function $M(\hat{\mathbf{x}}_i)$ is the local sensitivity (Hessian) matrix that arises

due to the distortion function. The Voronoi region $\bar{\mathcal{S}}_i$ can be approximated as

$$\bar{\mathcal{S}}_i = \left\{ \mathbf{x} : d(\mathbf{x}, \hat{\mathbf{x}}_i) < \min_{j \neq i} d(\mathbf{x}, \hat{\mathbf{x}}_j) \right\} \tag{4.72}$$

$$\approx \left\{ \mathbf{x} : d(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i) + \mathbf{g}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i)\mathbf{e} + \frac{1}{2}\mathbf{e}^T D(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i)\mathbf{e} < \tau \right\}, \tag{4.73}$$

where τ is a threshold and $\mathbf{e} \triangleq \mathbf{x} - \hat{\mathbf{x}}_i$. As noted earlier, $d(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i) = 0$ and $\mathbf{g}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i) = \underline{\mathbf{0}}$, thus, the quantization region is approximated by (4.70) with $M(\hat{\mathbf{x}}_i) = D(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i)$. The error in making this approximation is investigated in [86], and it is shown that the error is under four percent for spaces with dimensions under 5, and becomes asymptotically tight as the number of dimensions increases. Note also that the quantization cell approximation holds when the encoder is optimum, i.e., it is defined using either (4.1) or (4.4) (which are equivalent for an SEC).

The following result from [78] is useful in evaluating the inner integral of (4.69) under the quantization cell approximation

$$\int_{\mathbf{e} \in \mathcal{T}} \mathbf{e}^T D(\hat{\mathbf{x}}_i, \mathbf{y})\mathbf{e} \, d\mathbf{e} = \frac{V_{\bar{\mathcal{E}}_i}}{n+2} \left(\frac{V_{\bar{\mathcal{E}}_i}^2 \, |D(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i)|}{\kappa_n^2} \right)^{1/n} \operatorname{tr}\left(D^{-1}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i) D(\hat{\mathbf{x}}_i, \mathbf{y}) \right). \tag{4.74}$$

It is also easy to show that

$$\int_{\mathbf{e} \in \mathcal{T}\left(\underline{\mathbf{0}}, D(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i), V_{\bar{\mathcal{E}}_i}\right)} \mathbf{e} \, d\mathbf{e} = \underline{\mathbf{0}}, \tag{4.75}$$

i.e., under the quantization cell approximation, the normalized mean vector $\mathbf{w}_{\bar{\mathcal{E}}}(\mathbf{x}) \approx \underline{\mathbf{0}}$ and thus the $E_d^{(2)}$ term drops out of the analysis.

Next, using (4.74) in (4.69),

$$E_d^{(3)} \approx \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1}^{N} P_{j|i} \frac{V_{\bar{\mathcal{E}}_i}}{2(n+2)} \left(\frac{V_{\bar{\mathcal{E}}_i}^2 \, |D(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i)|}{\kappa_n^2} \right)^{1/n} \operatorname{tr}\left(D^{-1}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i) D(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) \right). \tag{4.76}$$

Notice that unlike in the $E_d^{(1)}$ term, here it is necessary to separately consider the $i = j$ and $i \neq j$ terms. This is because, the summand of the $E_d^{(1)}$ term was zero when $i = j$, whereas, the summand of $E_d^{(3)}$ is non-zero. Also, note that when $j = i$,

$\mathrm{tr}\left(D^{-1}(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_i)D(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_i)\right)=n$. Thus,

$$\begin{aligned} E_d^{(3)} &\approx \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1,j\neq i}^{N} \epsilon(N)\frac{V_{\bar{\mathcal{E}}_i}}{2(n+2)}\left(\frac{V_{\bar{\mathcal{E}}_i}^2\left|D(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_i)\right|}{\kappa_n^2}\right)^{1/n} \\ &\quad \mathrm{tr}\left(D^{-1}(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_i)D(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_j)\right) \\ &+ \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i)(1-(N-1)\epsilon(N))\frac{nV_{\bar{\mathcal{E}}_i}}{2(n+2)}\left(\frac{V_{\bar{\mathcal{E}}_i}^2\left|D(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_i)\right|}{\kappa_n^2}\right)^{1/n}. \end{aligned} \quad (4.77)$$

This can be simplified further by including the $i=j$ term in the first summation and suitably modifying the second summation to get

$$\begin{aligned} E_d^{(3)} &\approx \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \sum_{j=1}^{N} \epsilon(N)\frac{V_{\bar{\mathcal{E}}_i}}{2(n+2)}\left(\frac{V_{\bar{\mathcal{E}}_i}^2\left|D(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_i)\right|}{\kappa_n^2}\right)^{1/n} \mathrm{tr}\left(D^{-1}(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_i)D(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_j)\right) \\ &+ \sum_{i=1}^{N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i)\left(1-N\epsilon(N)\right)\frac{nV_{\bar{\mathcal{E}}_i}}{2(n+2)}\left(\frac{V_{\bar{\mathcal{E}}_i}^2\left|D(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_i)\right|}{\kappa_n^2}\right)^{1/n}. \end{aligned} \quad (4.78)$$

This term is now converted into an integral by making the substitution $\lambda(\hat{\mathbf{x}}_i)\simeq 1/(NV_{\bar{\mathcal{E}}_i})$, and replacing $\hat{\mathbf{x}}_i$ and $V_{\bar{\mathcal{E}}_i}$ by $\mathbf{x}$ and $d\mathbf{x}$ respectively to obtain

$$\begin{aligned} E_d^{(3)} &\approx \frac{N^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}}{2(n+2)}N\epsilon(N)\int_{\mathbf{x}}\lambda^{\frac{-2}{n}}(\mathbf{x})\left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n}} \\ &\quad \cdot\mathrm{tr}\left(D^{-1}(\mathbf{x},\mathbf{x})\left[\frac{1}{N}\sum_{j=1}^{N}D(\mathbf{x},\hat{\mathbf{x}}_j)\right]\right)f_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x} \\ &+ \frac{nN^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}}{2(n+2)}\left(1-N\epsilon(N)\right)\int_{\mathbf{x}}\lambda^{\frac{-2}{n}}(\mathbf{x})\left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n}}f_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x} \end{aligned} \quad (4.79)$$

The summation over j can be simplified using (4.61) to get

$$\begin{aligned} E_d^{(3)} &\approx \frac{N^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}}{2(n+2)}N\epsilon(N)\int_{\mathbf{x}}\lambda^{\frac{-2}{n}}(\mathbf{x})\left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n}} \\ &\quad \cdot\mathrm{tr}\left(D^{-1}(\mathbf{x},\mathbf{x})\left[\int_{\mathbf{y}}D(\mathbf{x},\mathbf{y})\lambda(\mathbf{y})d\mathbf{y}\right]\right)f_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x} \\ &+ \frac{nN^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}}{2(n+2)}\left(1-N\epsilon(N)\right)\int_{\mathbf{x}}\lambda^{\frac{-2}{n}}(\mathbf{x})\left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n}}f_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x} \end{aligned} \quad (4.80)$$

In summary, the expected distortion of source quantization in the pres-

ence of channel errors is given by

$$\begin{aligned} E_d &\approx N\epsilon(N)\int_{\mathbf{x}}\int_{\mathbf{y}} d(\mathbf{x},\mathbf{y}) f_{\mathbf{x}}(\mathbf{x})\lambda(\mathbf{y})d\mathbf{y}d\mathbf{x} \\ &+ \frac{N^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}}{2(n+2)} N\epsilon(N)\int_{\mathbf{x}} \lambda^{\frac{-2}{n}}(\mathbf{x})\left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n}} \\ &\quad \cdot \mathrm{tr}\left(D^{-1}(\mathbf{x},\mathbf{x})\left[\int_{\mathbf{y}} D(\mathbf{x},\mathbf{y})\lambda(\mathbf{y})d\mathbf{y}\right]\right) f_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x} \\ &+ \frac{nN^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}}{2(n+2)}\left(1-N\epsilon(N)\right)\int_{\mathbf{x}} \lambda^{\frac{-2}{n}}(\mathbf{x})\left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n}} f_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x}. \end{aligned} \tag{4.81}$$

4.8.3 Extension to arbitrary φ

As pointed out earlier, for an arbitrary φ, the expected distortion is given as the sum of five terms representing mutually exclusive and exhaustive possible pairs of transmit and receive indices. Thus,

$$\begin{aligned} E_d &\approx \sum_{i=1}^{\alpha N}\int_{\mathbf{x}\in\bar{\mathcal{R}}_i}\sum_{j=1,j\neq i}^{\alpha N}\epsilon(N)d(\mathbf{x},\hat{\mathbf{x}}_j)f_{\mathbf{x}}(\mathbf{x})d\mathbf{x} \\ &+ \sum_{i=1}^{\alpha N}\int_{\mathbf{x}\in\bar{\mathcal{R}}_i}\left(1-(N-1)\,\epsilon(N)\right)d(\mathbf{x},\hat{\mathbf{x}}_i)f_{\mathbf{x}}(\mathbf{x})d\mathbf{x} \\ &+ \sum_{i=1}^{\alpha N}\int_{\mathbf{x}\in\bar{\mathcal{R}}_i} N(1-\alpha)\epsilon(N)d(\mathbf{x},\bar{\mathbf{x}}_d)f_{\mathbf{x}}(\mathbf{x})d\mathbf{x} \\ &+ \int_{\mathbf{x}\in\bar{\mathcal{R}}_{\bar{\mathbf{x}}_d}}\sum_{i=1}^{\alpha N}\epsilon(N)d(\mathbf{x},\hat{\mathbf{x}}_j)f_{\mathbf{x}}(\mathbf{x})d\mathbf{x} \\ &+ \int_{\mathbf{x}\in\bar{\mathcal{R}}_{\bar{\mathbf{x}}_d}}(1-\alpha N\epsilon(N))d(\mathbf{x},\bar{\mathbf{x}}_d)f_{\mathbf{x}}(\mathbf{x})d\mathbf{x} \end{aligned} \tag{4.82}$$

where $\bar{\mathcal{R}}_{\bar{\mathbf{x}}_d}$ is the Voronoi region corresponding to the centroid codepoint $\bar{\mathbf{x}}_d$, as before. Denote by $A_1,\ldots,A_5$, the five terms in the above summation, respectively. Define $\varphi_{(\alpha,N)} \triangleq \alpha N+1$. Using the approximations in the previous subsection,

$$\begin{aligned} A_1 &= \sum_{i=1}^{\alpha N}\sum_{j=1,j\neq i}^{\alpha N}\epsilon(N)d(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_j)f_{\mathbf{x}}(\hat{\mathbf{x}}_i)V_{\bar{\mathcal{E}}_i} \\ &+ \sum_{i=1}^{\alpha N}\sum_{j=1,j\neq i}^{\alpha N}\frac{\epsilon(N)f_{\mathbf{x}}(\hat{\mathbf{x}}_i)}{2}\int_{\mathbf{x}\in\hat{\mathcal{R}}_i}(\mathbf{x}-\hat{\mathbf{x}}_i)^T D(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_j)(\mathbf{x}-\hat{\mathbf{x}}_i) \end{aligned} \tag{4.83}$$

Then inner summation over j can be converted into the corresponding integral over the point density, however, since the centroid codepoint $\bar{\mathbf{x}}_d$ is not included in the summation, distortion $d(\hat{\mathbf{x}}_i, \bar{\mathbf{x}}_d)$ arising due to that codepoint needs to be subtracted, as follows

$$\begin{aligned} A_1 &\approx \epsilon(N) \sum_{i=1}^{\alpha N} \left[\varphi_{(\alpha,N)} \int d(\hat{\mathbf{x}}_i, \mathbf{y}) \lambda(\mathbf{y}) d\mathbf{y} - d(\hat{\mathbf{x}}_i, \bar{\mathbf{x}}_d) \right] f_{\mathbf{x}}(\hat{\mathbf{x}}_i) V_{\bar{\mathcal{E}}_i} \\ &+ \sum_{i=1}^{\alpha N} \sum_{j=1, j\neq i}^{\alpha N} \frac{\epsilon(N) f_{\mathbf{x}}(\hat{\mathbf{x}}_i) V_{\bar{\mathcal{E}}_i}}{2(n+2)} \left(\frac{V_{\bar{\mathcal{E}}_i}^2 \left| D(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i) \right|}{\kappa_n^2} \right)^{\frac{1}{n}} \operatorname{tr}\left(D^{-1}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i) D(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) \right) \end{aligned}$$

After considerable simplification, it can be shown that

$$\begin{aligned} A_1 &\approx \varphi_{(\alpha,N)} \epsilon(N) \int \int d(\mathbf{x}, \mathbf{y}) \lambda(\mathbf{y}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{y} d\mathbf{x} \\ &- \varphi_{(\alpha,N)} \epsilon(N) \int d(\bar{\mathbf{x}}_d, \mathbf{y}) \lambda(\mathbf{y}) d\mathbf{y} \cdot f_{\mathbf{x}}(\bar{\mathbf{x}}_d) V_{\bar{\mathcal{E}}_d} \\ &- \epsilon(N) \int d(\mathbf{x}, \bar{\mathbf{x}}_d) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} + \frac{\epsilon(N)}{2(n+2)} \int \left(\frac{|D(\mathbf{x}, \mathbf{x})|}{\varphi_{(\alpha,N)}^2 \lambda^2(\mathbf{x}) \kappa_n^2} \right)^{\frac{1}{n}} \\ &\quad \cdot \operatorname{tr}\left(D^{-1}(\mathbf{x}, \mathbf{x}) \left(\varphi_{(\alpha,N)} \int D(\mathbf{x}, \mathbf{y}) \lambda(\mathbf{y}) d\mathbf{y} \right) \right) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ &- \frac{n\epsilon(N)}{2(n+2)} \int \left(\frac{|D(\mathbf{x}, \mathbf{x})|}{\varphi_{(\alpha,N)}^2 \lambda^2(\mathbf{x}) \kappa_n^2} \right)^{\frac{1}{n}} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ &- \frac{\epsilon(N)}{2(n+2)} \int \left(\frac{|D(\mathbf{x}, \mathbf{x})|}{\varphi_{(\alpha,N)}^2 \lambda^2(\mathbf{x}) \kappa_n^2} \right)^{\frac{1}{n}} \operatorname{tr}\left(D^{-1}(\mathbf{x}, \mathbf{x}) D(\mathbf{x}, \bar{\mathbf{x}}_d) \right) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ &- \frac{\epsilon(N) f_{\mathbf{x}}(\bar{\mathbf{x}}_d) V_{\bar{\mathcal{E}}_d}}{2(n+2)} \left(\frac{V_{\bar{\mathcal{E}}_d}^2 \left| D(\bar{\mathbf{x}}_d, \bar{\mathbf{x}}_d) \right|}{\kappa_n^2} \right)^{\frac{1}{n}} \\ &\quad \cdot \operatorname{tr}\left(D^{-1}(\bar{\mathbf{x}}_d, \bar{\mathbf{x}}_d) \left(\varphi_{(\alpha,N)} \int D(\bar{\mathbf{x}}_d, \mathbf{y}) \lambda(\mathbf{y}) d\mathbf{y} \right) \right) \\ &+ \frac{n\epsilon(N) f_{\mathbf{x}}(\bar{\mathbf{x}}_d) V_{\bar{\mathcal{E}}_d}}{n+2} \left(\frac{V_{\bar{\mathcal{E}}_d}^2 \left| D(\bar{\mathbf{x}}_d, \bar{\mathbf{x}}_d) \right|}{\kappa_n^2} \right)^{\frac{1}{n}} \end{aligned} \tag{4.84}$$

Next, the A_2 term is given by

$$\begin{aligned} A_2 &= (1 - (N-1)\epsilon(N)) \sum_{i=1}^{\alpha N} \int_{\mathbf{x} \in \bar{\mathcal{R}}_i} d(\mathbf{x}, \hat{\mathbf{x}}_i) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ &\approx (1 - (N-1)\epsilon(N)) \sum_{i=1}^{\alpha N} f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \int_{\mathbf{x} \in \bar{\mathcal{R}}_i} \frac{1}{2} (\mathbf{x} - \hat{\mathbf{x}}_i)^T D(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i) (\mathbf{x} - \hat{\mathbf{x}}_i) d\mathbf{x} \end{aligned}$$

Again, it can be shown that A_2 reduces to

$$\begin{aligned} A_2 &\approx \frac{n(1-(N-1)\epsilon(N))}{2(n+2)} \left[\int \left(\frac{|D(\mathbf{x},\mathbf{x})|}{\varphi^2_{(\alpha,N)}\lambda^2(\mathbf{x})\kappa_n^2} \right)^{\frac{1}{n}} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \right. \\ &\quad \left. - f_{\mathbf{x}}(\bar{\mathbf{x}}_d) V_{\bar{\mathcal{E}}_d} \left(\frac{V^2_{\bar{\mathcal{E}}_d} |D(\bar{\mathbf{x}}_d, \bar{\mathbf{x}}_d)|}{\kappa_n^2} \right)^{\frac{1}{n}} \right] \end{aligned}$$

Next, the A_3 term:

$$\begin{aligned} A_3 &= \sum_{i=1}^{\alpha N} \int_{\mathbf{x}\in\hat{\mathcal{R}}_i} N(1-\alpha)\epsilon(N) d(\mathbf{x}, \bar{\mathbf{x}}_d) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ &\approx \sum_{i=1}^{\alpha N} N(1-\alpha)\epsilon(N) f_{\mathbf{x}}(\hat{\mathbf{x}}_i) \\ &\quad \cdot \int_{\mathbf{x}\in\hat{\mathcal{R}}_i} \left(d(\hat{\mathbf{x}}_i, \bar{\mathbf{x}}_d) + \frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}}_i)^T D(\hat{\mathbf{x}}_i, \bar{\mathbf{x}}_d)(\mathbf{x}-\hat{\mathbf{x}}_i) \right) d\mathbf{x}, \end{aligned} \tag{4.85}$$

which reduces to

$$\begin{aligned} A_3 &= N(1-\alpha)\epsilon(N) \int d(\mathbf{x}, \bar{\mathbf{x}}_d) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ &\quad + \frac{N(1-\alpha)\epsilon(N)}{2(n+2)} \int \left(\frac{|D(\mathbf{x},\mathbf{x})|}{\varphi^2_{(\alpha,N)}\lambda^2(\mathbf{x})\kappa_n^2} \right)^{\frac{1}{n}} \operatorname{tr}\left(D^{-1}(\mathbf{x},\mathbf{x}) D(\mathbf{x}, \bar{\mathbf{x}}_d) \right) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ &\quad - \frac{nN(1-\alpha)\epsilon(N)}{2(n+2)} f_{\mathbf{x}}(\bar{\mathbf{x}}_d) V_{\bar{\mathcal{E}}_d} \left(\frac{V^2_{\bar{\mathcal{E}}_d} |D(\bar{\mathbf{x}}_d, \bar{\mathbf{x}}_d)|}{\kappa_n^2} \right)^{\frac{1}{n}}. \end{aligned} \tag{4.86}$$

And the A_4 term:

$$\begin{aligned} A_4 &= \epsilon(N) \sum_{j=1}^{\alpha N} \int_{\mathbf{x}\in\hat{\mathcal{R}}_d} d(\mathbf{x}, \hat{\mathbf{x}}_j) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ &\approx \epsilon(N) f_{\mathbf{x}}(\bar{\mathbf{x}}_d) \sum_{j=1}^{\alpha N} \int_{\mathbf{x}\in\hat{\mathcal{R}}_d} d(\bar{\mathbf{x}}_d, \hat{\mathbf{x}}_j) + \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_d)^T D(\bar{\mathbf{x}}_d, \hat{\mathbf{x}}_j)(\mathbf{x}-\bar{\mathbf{x}}_d) d\mathbf{x} \\ &\approx \varphi_{(\alpha,N)}\epsilon(N) f_{\mathbf{x}}(\bar{\mathbf{x}}_d) V_{\bar{\mathcal{E}}_d} \left[\int d(\bar{\mathbf{x}}_d, \mathbf{y}) \lambda(\mathbf{y}) d\mathbf{y} \right. \\ &\quad \left. + \frac{1}{2(n+2)} \left(\frac{V^2_{\bar{\mathcal{E}}_d} |D(\bar{\mathbf{x}}_d, \bar{\mathbf{x}}_d)|}{\kappa_n^2} \right)^{\frac{1}{n}} \operatorname{tr}\left(D^{-1}(\bar{\mathbf{x}}_d, \bar{\mathbf{x}}_d) \int D(\bar{\mathbf{x}}_d, \mathbf{y}) \lambda(\mathbf{y}) d\mathbf{y} \right) \right] \\ &\quad - \epsilon(N) f_{\mathbf{x}}(\bar{\mathbf{x}}_d) V_{\bar{\mathcal{E}}_d} \frac{n}{2(n+2)} \left(\frac{V^2_{\bar{\mathcal{E}}_d} |D(\bar{\mathbf{x}}_d, \bar{\mathbf{x}}_d)|}{\kappa_n^2} \right)^{\frac{1}{n}} \end{aligned} \tag{4.87}$$

And finally, A_5 is given by

$$A_5 \approx (1-\alpha N\epsilon(N)) f_{\mathbf{x}}(\bar{\mathbf{x}}_d) V_{\bar{\mathcal{E}}_d} \frac{n}{2(n+2)} \left(\frac{V_{\bar{\mathcal{E}}_d}^2 \, |D(\bar{\mathbf{x}}_d, \bar{\mathbf{x}}_d)|}{\kappa_n^2} \right)^{\frac{1}{n}} \tag{4.88}$$

Collecting the terms A_1 through A_5 together and simplifying, all the $f_{\mathbf{x}}(\bar{\mathbf{x}}_d) V_{\bar{\mathcal{E}}_d}$ terms cancel out, yielding

$$\begin{aligned} E_d \;\approx\; & \varphi_{(\alpha,N)}\epsilon(N) \int\int d(\mathbf{x},\mathbf{y})\lambda(\mathbf{y}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{y} d\mathbf{x} \\ & + (N - \varphi_{(\alpha,N)})\epsilon(N) \int d(\mathbf{x}, \bar{\mathbf{x}}_d) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ & + \frac{\varphi_{(\alpha,N)}\epsilon(N)}{2(n+2)} \int \left(\frac{|D(\mathbf{x},\mathbf{x})|}{\varphi_{(\alpha,N)}^2 \lambda^2(\mathbf{x}) \kappa_n^2} \right)^{\frac{1}{n}} \\ & \cdot \mathrm{tr} \left(D^{-1}(\mathbf{x},\mathbf{x}) \int D(\mathbf{x},\mathbf{y}) \lambda(\mathbf{y}) d\mathbf{y} \right) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ & + \frac{n(1 - N\epsilon(N))}{2(n+2)} \int \left(\frac{|D(\mathbf{x},\mathbf{x})|}{\varphi_{(\alpha,N)}^2 \lambda^2(\mathbf{x}) \kappa_n^2} \right)^{\frac{1}{n}} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ & + \frac{\left(N - \varphi_{(\alpha,N)}\right) \epsilon(N)}{2(n+2)} \int \left(\frac{|D(\mathbf{x},\mathbf{x})|}{\varphi_{(\alpha,N)}^2 \lambda^2(\mathbf{x}) \kappa_n^2} \right)^{\frac{1}{n}} \\ & \cdot \mathrm{tr} \left(D^{-1}(\mathbf{x},\mathbf{x}) D(\mathbf{x}, \bar{\mathbf{x}}_d) \right) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \end{aligned} \tag{4.89}$$

4.8.4 Variance of the expected distortion

The derivation in this section follows along the lines of a related one in [68], hence, only an outline of the proof is provided here. The variance is given by $\mathrm{Var}(\tilde{E}_{d|\pi}) = \mathbf{E}\left\{\tilde{E}_{d|\pi}^2\right\} - \mathbf{E}\left\{\tilde{E}_{d|\pi}\right\}^2$. The expectation of $\tilde{E}_{d|\pi}$ over all possible index assignments is well approximated by the sum of the first two terms of (4.21), for large N. Therefore, only $\mathbf{E}\left\{\tilde{E}_{d|\pi}^2\right\}$ needs to be evaluated. Since

$$\tilde{E}_{d|\pi}^2 = \sum_{i \neq j} \sum_{k \neq l} P_{\pi(j)|\pi(i)} P_{\pi(l)|\pi(k)} \hat{d}_{ij} \hat{d}_{kl}, \tag{4.90}$$

the above double summation can be upper bounded by expressing it as the sum of seven mutually disjoint and exhaustive cases for the indices i, j, k and l, with corresponding terms $E_1, \ldots E_7$, as follows.

Case: i, j, k, l **distinct**

Using properties of the DSC, it can be shown that

$$\mathbf{E}_\pi \left\{ P_{\pi(j)|\pi(i)} P_{\pi(l)|\pi(k)} \right\} \leq \frac{N(N-1)}{(N-2)(N-3)} \epsilon^2(N), \tag{4.91}$$

which implies

$$\begin{aligned} E_1 &\triangleq \sum_{i,j,k,l \text{ distinct}} P_{\pi(j)|\pi(i)} P_{\pi(l)|\pi(k)} \hat{d}_{ij} \hat{d}_{kl}, \\ &\leq \frac{N(N-1)}{(N-2)(N-3)} \left(\epsilon(N) \sum_{i,j \text{ distinct}} \hat{d}_{ij} \right)^2, \\ &= \frac{N(N-1)}{(N-2)(N-3)} \mathbf{E} \left\{ \tilde{E}_{d|\pi} \right\}^2. \end{aligned} \tag{4.92}$$

Thus, $E_1 - \mathbf{E} \left\{ \tilde{E}_{d|\pi} \right\}^2 = O(1/N)$ as N gets large.

Case: $i = k$ **and** $j = l$

It can be shown that

$$\mathbf{E}_\pi \left\{ P_{\pi(j)|\pi(i)} P_{\pi(l)|\pi(k)} \right\} \leq (1 - (N-1)\,\epsilon(N))\,\epsilon(N), \tag{4.93}$$

which implies

$$E_2 \leq (N-1)\epsilon(N)\,(1 - (N-1)\,\epsilon(N))\,d_{\max}^2 P_{\max}(N) = O(P_{\max}(N)) \tag{4.94}$$

where $d_{\max} \triangleq \sup_{\mathbf{x},\mathbf{y} \in \mathcal{D}_\mathbf{x}} d(\mathbf{x}, \mathbf{y}) < \infty$ since the distortion is bounded, and $P_{\max}(N) \triangleq \max_{1 \leq i \leq N} P_i \to 0$ as $N \to \infty$ because of the diminishing cell diameters.

Case: $l = i$ **and** $j = k$

In this case also, it can be shown that

$$\mathbf{E}_\pi \left\{ P_{\pi(j)|\pi(i)} P_{\pi(l)|\pi(k)} \right\} \leq (1 - (N-1)\,\epsilon(N))\,\epsilon(N), \tag{4.95}$$

from which, after some simplification, we get

$$E_3 \leq \epsilon(N)\,(1 - (N-1)\,\epsilon(N))\,d_{\max}^2 = O(1/N). \tag{4.96}$$

Case: $i = k$ **and** $j \neq l$

In this case, it can be shown that

$$\mathbf{E}_\pi \left\{ P_{\pi(j)|\pi(i)} P_{\pi(l)|\pi(k)} \right\} \leq \frac{1}{(N-1)(N-2)}, \tag{4.97}$$

and therefore,

$$E_4 \leq d_{\max}^2 P_{\max}(N) = O(P_{\max}(N)). \tag{4.98}$$

Case: $j = l$ **and** $i \neq k$

Here, we have

$$\mathbf{E}_\pi \left\{ P_{\pi(j)|\pi(i)} P_{\pi(l)|\pi(k)} \right\} \leq \frac{(N-1)\epsilon^2(N)}{(N-2)}, \tag{4.99}$$

from which, it can be shown that

$$E_5 \leq (N-1)\epsilon^2(N) d_{\max}^2 = O(1/N). \tag{4.100}$$

Case: $j = k$ **and** $i \neq l$

Again, it can be shown that for discrete symmetric channels,

$$\mathbf{E}_\pi \left\{ P_{\pi(j)|\pi(i)} P_{\pi(l)|\pi(k)} \right\} \leq \frac{1}{(N-1)(N-2)}, \tag{4.101}$$

and thus

$$E_6 \leq \frac{d_{\max}^2}{N-1} = O(1/N). \tag{4.102}$$

Case: $i = l$ **and** $k \neq j$

Finally, it can be shown that in this case,

$$\mathbf{E}_\pi \left\{ P_{\pi(j)|\pi(i)} P_{\pi(l)|\pi(k)} \right\} \leq \frac{1}{(N-1)(N-2)}, \tag{4.103}$$

and thus

$$E_7 \leq \frac{d_{\max}^2}{N-1} = O(1/N). \tag{4.104}$$

Putting it all together, at high rates,

$$\mathrm{Var}(\tilde{E}_{d|\pi}) = O(1/N) + O(P_{\max}(N)). \tag{4.105}$$

Acknowledgement

This chapter, in part, has appeared in C. R. Murthy and B. D. Rao, "High-Rate Analysis of Source Coding for Symmetric Error Channels", *Data Compression Conference (DCC)*, Snowbird, UT, Mar. 2006, and C. R. Murthy, E. R. Duni and B. D. Rao, "High-rate analysis of vector quantization for noisy channels", *Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, Tolouse, France, May 2006.

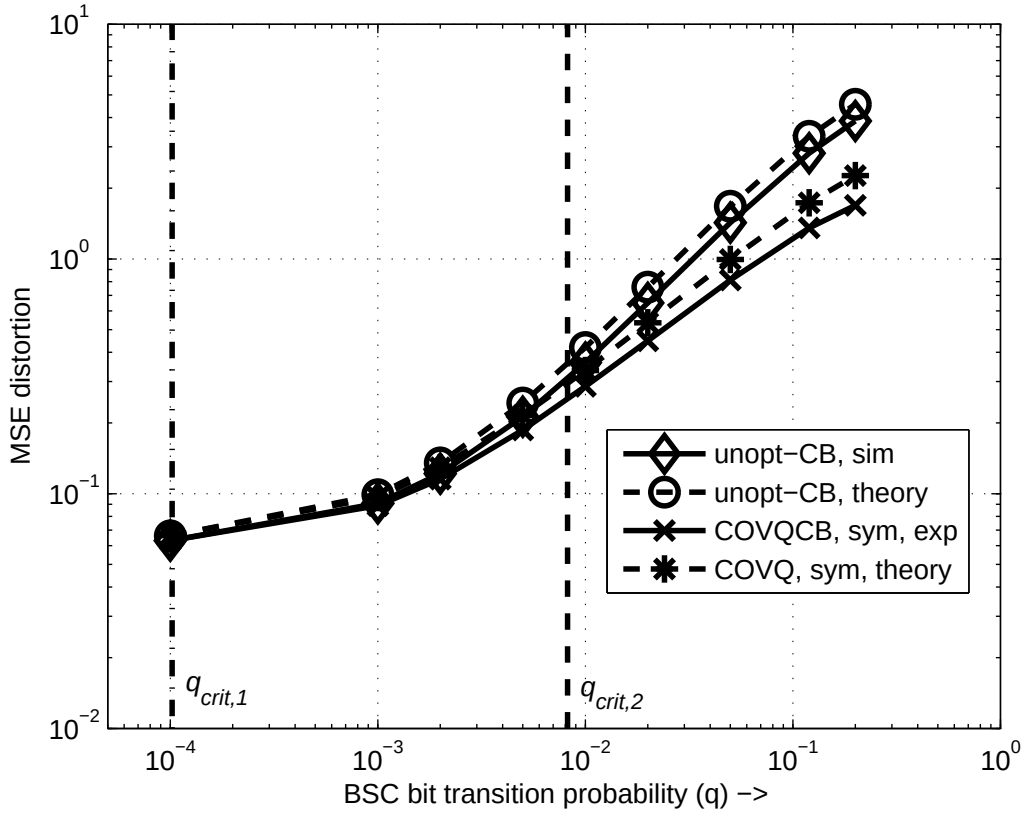


Figure 4.8: MSE distortion for a 2-dimensional standard Gaussian random vector with the conventional point density, and the number of quantization bits B fixed at 6 bits. The quantized index is sent over a BSC with bit transition probability q (the x-axis). The two vertical lines show the values of q corresponding to $\epsilon_{\text{crit},1}$ and $\epsilon_{\text{crit},2}$, the two critical values of $\epsilon(N)$, respectively.

5 High-Rate Vector Quantization for Noisy Channels With Random Index Assignment, Part 2: Applications

5.1 Introduction

In this chapter, the results presented in the Chap. 4 are used to derive the performance of high-rate vector quantization for Simplex Error Channels (SEC) in two specific applications. The first application is in feedback-based transmission on multiple input, single output (MISO) communication systems. In practice, per-antenna transmit power constraints are more meaningful than total transmit power constraints, as they impose less stringent fidelity requirements on the transmit RF power amplifiers. When the CSI is known perfectly at the transmitter, beamforming is the optimum method of transmission in MISO systems to maximize the channel capacity both under a per-antenna power constraint as well as under a total power constraint [7], [8]. However, due to feedback channel bit rate constraints, only a quantized version of the CSI can be made available to the transmitter. Quantized channel feedback is also under consideration in 3rd generation mobile and wireless LAN standards, for example in the closed-loop mode specification in 3GPP High Speed Downlink Packet Access (HSDPA) [4] and in the

eigenbeamforming mode specification in IEEE 802.11 [5] and IEEE 802.16 [6]. Recent overviews of work quantifying the performance of finite-rate feedback systems with beamforming at the transmitter can be found in [34] and [87]. All of the previous works assume that the feedback channel is noiseless. Errors in the feedback channel can adversely affect the accuracy of the CSI available at the transmitter, thereby lowering the performance of the communication system.

In this chapter, the sensitivity of quantized Equal Gain Transmission (EGT) [8] systems to errors in the feedback channel is analyzed. An EGT beamforming vector is given by $\mathbf{w} = [1, \exp(j\theta_2), \exp(j\theta_3), \ldots, \exp(j\theta_t)]^T$, where θ_i denotes the phase rotation applied at antenna element i. In Chapter 3, Vector Quantization (VQ) of the parameters was considered, and it was shown that the capacity loss with quantized EGT (Q-EGT) drops off with the number of feedback bits B approximately as $\frac{2(t-1)}{t+1}2^{-\frac{2B}{t-1}}$ bits, when the feedback channel is noiseless. A VQ based approach was employed to design the codebook for per-antenna power constrained beamforming and a modified Lloyd codebook generation algorithm was derived with the capacity loss as the performance metric. Other papers that consider the design of per-antenna power constrained transmission schemes include [32], [33], [38], [39] and [88].

There is an inherent similarity between classical source coding and channel quantization, since the channel instantiation can be thought of as a random source that needs to be compressed. This similarity is exploited to study the performance of quantized EGT systems when the feedback channel is noisy, with the loss in SNR relative to perfect feedback as performance measure. The extension to quantized feedback based beamforming systems is non-trivial because of two main reasons: (1) the source (channel instantiation) and the quantized vector (phase angles of the beamforming vector) lie in different manifolds and (2) the source has several extra random parameters (the real-valued antenna gains) that do not need to be quantized, but that can be used as side information at the encoder.

The second application of the theory developed in this chapter is in the

area of quantizing the Linear Predictive Coding (LPC) filter coefficients for speech compression. Here, the with the Log Spectral Distortion (LSD) as performance metric, the sensitivity of speech compression to SEC is analyzed. Monte Carlo methods have to be employed to evaluate some of the integrals, as a closed-form expression is not available. For verifying the results, a structured quantizer has to be employed in practice, as the number of code points is immense (about 50 bits must be used for high quality compression). This makes it impractical to implement unconstrained, full-search VQ, hence structured quantizers that offer low, rate independent complexity, such as that proposed in [89] have to be employed. It is shown that the theory can correctly predict the maximum error rate that can be tolerated, given an acceptable level (1dB^2) of distortion.

This chapter is organized as follows. In Section 5.2, the system model is set down and the noisy channel model is described. In Section 5.3, the main result that will be used to analyze the performance of the systems of interest is reiterated. The high rate result is applied to quantized EGT systems in Section 5.4. In Section 5.5, the performance of wideband speech spectrum compression with the log spectral distortion measure is derived. Simulation results to verify the accuracy of the analysis are presented in Section 5.6, and some concluding remarks are offered in Section 5.7.

5.2 Source and Channel Model

Let $\mathbf{x} \in \mathcal{D}_{\mathbf{x}} \subset \mathbb{R}^n$ be a random source with a continuous pdf $f_{\mathbf{x}}(\mathbf{x})$, where $\mathcal{D}_{\mathbf{x}}$ is the domain of $\mathbf{x}$. The vector quantization encoder is described by N partition regions $\bar{\mathcal{R}}_i, 1 \leq i \leq N$ that tile $\mathcal{D}_{\mathbf{x}}$. Associated with each partition region $\bar{\mathcal{R}}_i$ is a code-vector $\hat{\mathbf{x}}_i$. In the case of centroid quantizers considered here, $\hat{\mathbf{x}}_i$ is the centroid of the random vector $\mathbf{x}$ conditioned on $\mathbf{x} \in \bar{\mathcal{R}}_i$ under the distortion measure of interest. Now, whenever $\mathbf{x} \in \bar{\mathcal{R}}_i$, the quantizer outputs index i, which is mapped to a point in some constellation and sent over a noisy channel. At the receiver,

the index i is received as a possibly different index j with probability $P_{j|i}$, upon which it outputs $\hat{\mathbf{x}}_j$ as its estimate of $\mathbf{x}$. Finally, let $d(\mathbf{x}, \hat{\mathbf{x}})$ represent the distortion incurred in representing a source instantiation $\mathbf{x}$ by $\hat{\mathbf{x}}$. The distortion function is assumed to be non-negative, twice continuously differentiable and bounded, and with $d(\mathbf{x}, \mathbf{x}) = 0$ regardless of $\mathbf{x}$.

5.2.1 Discrete Symmetric Channels

As in the previous chapter, the noisy channel is modeled as a Discrete Symmetric Channel (DSC) with random index assignment, which can be represented by the equivalent Symmetric Error Channel (SEC)

$$P_{j|i} = \begin{cases} \epsilon(N), & j \neq i \\ 1 - (N-1)\epsilon(N), & j = i \end{cases}, \tag{5.1}$$

where $\epsilon(N)$ is the index error probability with N points in the codebook. Detailed explanation and justification of this channel model has been provided in the previous chapter.

Note that set-up assumes that as N increases, more (or less) energy is used to transmit the index in order to maintain the probability of correct reception $P_{i|i} = 1 - (N-1)\epsilon(N)$. For example, one simple model is obtained by assuming that as N is increased, the per-index transmit power is increased to maintain a constant probability of correct index reception, that is, $\epsilon(N) = \rho/(N-1)$. In this case, $P_{i|i} = 1 - \rho$ is independent of N. Another example is when the index is mapped to a $L \triangleq \log_2(N)$ bit symbol, and each bit is transmitted over a binary symmetric channel with cross-over probability q. In this case, with random index assignment, the probability of correct reception $P_{i|i} = (1-q)^L$, and thus $\epsilon(N) = \left(1 - (1-q)^L\right) / (N-1)$.

5.3 High-Rate Performance of Vector Quantization

In this section, the high-rate performance of vector quantization for the case of discrete symmetric channels with random index assignment is stated, and is applied to the two specific cases of interest in the following two sections.

The expected distortion performance is obtained by taking a triple expectation of the distortion over the n-dimensional source distribution, the channel transition probabilities $P_{j|i}$ and the index assignments $\pi(\cdot)$, as follows

$$E_d = \frac{1}{N!} \sum_{\pi} \sum_{i=1}^{N} \int_{\mathbf{x} \in \bar{\mathcal{R}}_i} \sum_{j=1}^{N} P_{\pi(j)|\pi(i)} d(\mathbf{x}, \hat{\mathbf{x}}_j) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}. \tag{5.2}$$

It has been shown in the previous chapter that the expected distortion is given by

$$E_d = \int_{\mathbf{x}} E_{d,\mathbf{x}} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \tag{5.3}$$

where, $E_{d,\mathbf{x}}$, the expected distortion conditioned on the source instantiation $\mathbf{x}$,

$$\begin{aligned} E_{d,\mathbf{x}} &\approx N\epsilon(N) \int_{\mathbf{y}} d(\mathbf{x}, \mathbf{y}) \lambda(\mathbf{y}) d\mathbf{y} \\ &+ \frac{N^{\frac{-2}{n}} \kappa_n^{\frac{-2}{n}}}{2(n+2)} N\epsilon(N) \lambda^{\frac{-2}{n}}(\mathbf{x}) \left| D(\mathbf{x}, \mathbf{x}) \right|^{\frac{1}{n}} \operatorname{tr} \left(D^{-1}(\mathbf{x}, \mathbf{x}) \left[\int_{\mathbf{y}} D(\mathbf{x}, \mathbf{y}) \lambda(\mathbf{y}) d\mathbf{y} \right] \right) \\ &+ \frac{n N^{\frac{-2}{n}} \kappa_n^{\frac{-2}{n}}}{2(n+2)} \left(1 - N\epsilon(N) \right) \lambda^{\frac{-2}{n}}(\mathbf{x}) \left| D(\mathbf{x}, \mathbf{x}) \right|^{\frac{1}{n}}. \end{aligned} \tag{5.4}$$

In the above equation, κ_n is the volume of an n-dimensional unit sphere, $D(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is an n by n dimensional matrix with j, k-th element defined by

$$D_{k,j}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \left. \frac{\partial^2 d(\mathbf{x}, \hat{\mathbf{y}})}{\partial x_j \partial x_k} \right|_{\mathbf{x} = \hat{\mathbf{x}}}, \tag{5.5}$$

and $\lambda(\mathbf{x})$ is the so-called *fractional point density*, and is defined as follows. The *specific point density* [76] is given by

$$\lambda_N(\mathbf{x}) \triangleq \frac{1}{N V(\bar{\mathcal{S}}_i)}, \text{if } \mathbf{x} \in \bar{\mathcal{S}}_i, \text{ for } i = 1, 2, \ldots, N. \tag{5.6}$$

Then, when N is very large, $\lambda_N(\mathbf{x})$ approximates a continuous nonnegative density function $\lambda(\mathbf{x})$ having a unit integral [77]. Note that the distortion expression (5.4)

can be rewritten as

$$\begin{aligned} E_{d,\mathbf{x}} &\approx N\epsilon(N)\left\{\int_{\mathbf{y}} d(\mathbf{x},\mathbf{y})\lambda(\mathbf{y})d\mathbf{y}\right. \\ &+ \frac{N^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}}{2(n+2)}\lambda^{\frac{-2}{n}}(\mathbf{x})\left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n}} \\ &\quad \left.\cdot\mathrm{tr}\left(D^{-1}(\mathbf{x},\mathbf{x})\left[\int_{\mathbf{y}}\left(D(\mathbf{x},\mathbf{y})-D(\mathbf{x},\mathbf{x})\right)\lambda(\mathbf{y})d\mathbf{y}\right]\right)\right\} \\ &+ \frac{nN^{\frac{-2}{n}}\kappa_n^{\frac{-2}{n}}}{2(n+2)}\lambda^{\frac{-2}{n}}(\mathbf{x})\left|D(\mathbf{x},\mathbf{x})\right|^{\frac{1}{n}}, \end{aligned} \tag{5.7}$$

where the last term is now the asymptotic distortion in the absence of channel errors (i.e., when $\epsilon(N) = 0$). The expected distortion is thus the sum of three terms. The first term represents the distortion incurred as a result of channel errors, the second term is an interdependence term, and the last term is the source quantization-induced distortion.

5.4 Multiple Antenna Systems with Finite-Rate Feedback

5.4.1 System Model

In this section, a multiple input, single output (MISO) system with t antennas at the transmitter is considered, as represented by Fig. 5.1. The multiple antenna flat-fading channel is modelled by the channel vector $\mathbf{h} \in \mathbb{C}^t$. Then, the channel input $\mathbf{x} \in \mathbb{C}^t$ and the channel output $y \in \mathbb{C}$ have the relationship

$$y = \mathbf{h}^H\mathbf{x} + \eta, \tag{5.8}$$

where $\eta \in \mathbb{C}$ is the zero mean, unit variance complex Gaussian noise at the receiver. The CSI $\mathbf{h}$ is assumed to be known perfectly at the receiver, and partially at the transmitter through a limited-rate noisy feedback channel. Note that under the block fading assumption, the time index is unimportant to the derivations, hence, the dependence on time is not explicitly shown. Now, the transmitted vector $\mathbf{x}$ is obtained by multiplying the data symbol $s \in \mathbb{C}$ by a beamforming vector $\mathbf{w}$ to get

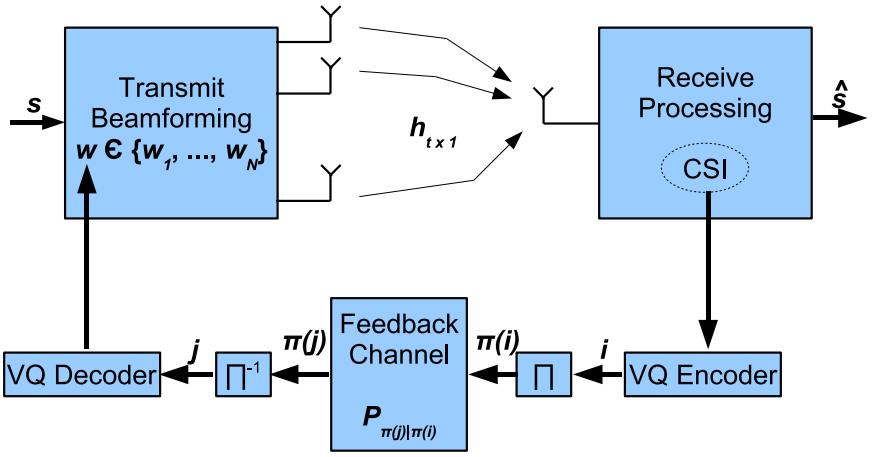


Figure 5.1: Schematic representation of a MISO system with beamforming at the transmitter.

$\mathbf{x} = \mathbf{w}s$. The power in the data symbol is denoted $P_s \triangleq \mathbf{E}\left\{|s|^2\right\}$. A quantized beamforming-vector codebook $\mathcal{C} \triangleq \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\}$ is known to both the receiver and the transmitter, where $N = 2^B$. Based on the knowledge of $\mathbf{h}$, the receiver selects the best beamforming vector $\mathbf{w}_i \in \mathcal{C}$ and sends the corresponding index i to the transmitter through the noisy feedback channel. The transition probability of receiving index j given that the transmitted index is i, is given by (5.1). When the transmitter receives index j, it employs beamforming vector $\mathbf{w}_j$ for transmission.

The transmitter is assumed to employ EGT beamforming [8]. In general, an EGT beamforming vector has the form $\mathbf{w} = [1, \exp j\theta_2, \exp j\theta_3, \ldots, \exp j\theta_t]^T$. Thus, the EGT vector contains $n = t - 1$ real-valued phase parameters that need to be made available to the transmitter to enable EGT.

5.4.2 Distortion Measure

It is well known that the received SNR with perfect feedback (i.e., $\mathbf{h}$ known at the transmitter) and optimum EGT is given by $P_s\|\mathbf{h}\|_1^2$ where P_s is the per-antenna power constraint. Also, the received SNR when the beamforming vector $\hat{\mathbf{w}}$ is employed at the transmitter is given by $P_s\left|\mathbf{h}^H\mathbf{w}\right|^2$. Thus, the loss in

received SNR relative to the received SNR with perfect feedback, which will be the distortion function of interest, is given by

$$d(\mathbf{h}, \hat{\mathbf{w}}) = \left(1 - \frac{\left|\mathbf{h}^H \hat{\mathbf{w}}\right|^2}{\|\mathbf{h}\|_1^2}\right), \tag{5.9}$$

which can be simplified using the above notation as

$$d(\mathbf{h}, \hat{\mathbf{w}}) = d_1(\mathbf{x}, \hat{\mathbf{x}}; \mathbf{r}) = \left(1 - \frac{\left|\left[1 \exp\left(j\left(\hat{\mathbf{x}} - \mathbf{x}\right)\right)\right] \mathbf{r}\right|^2}{s_{\mathbf{r}}^2}\right), \tag{5.10}$$

since the distortion function is independent of θ_1, the phase of h_1. It can be shown that the above expression closely approximates the ergodic capacity loss incurred by using $\hat{\mathbf{w}}$ as the beamforming vector at the transmitter instead of $\mathbf{w}_o$, when the transmit power P_s and the number of quantization levels N are large [88].

5.4.3 High-Rate Performance Analysis

In this subsection, the result in Sec. 5.3 is used to derive analytical expressions for the performance of QEGT in the case of an *i.i.d.* Rayleigh flat-fading channel with unit-variance complex Gaussian entries, when the quantizer output (index) is sent to the transmitter over a DSC with random index assignment. Specifically, the relative loss in SNR (or capacity loss under the high-SNR assumption) [88] incurred due to the quantization of the EGT beamforming vector by a finite number of bits is analytically characterized. Let $r_i \triangleq |h_i|$ and $x_i \triangleq \angle h_i - \angle h_1$, for $1 \leq i \leq t$. Also, let $\theta_1 \triangleq \angle h_1$, $\mathbf{x} \triangleq [x_2, \ldots, x_t]$, $\mathbf{r} \triangleq [r_1, \ldots r_t]^T$, and $s_{\mathbf{r}} \triangleq \sum_{i=1}^{t} r_i = \|\mathbf{h}\|_1$. Thus, $\mathbf{h}$ can be rewritten as

$$\mathbf{h} = \exp(j\theta_1)\mathrm{diag}([1 \exp(j\mathbf{x})])\mathbf{r}. \tag{5.11}$$

Note that $x_1 = 0$, and as observed earlier, only the $t-1$ phase angle differences given by the entries of $\mathbf{x}$ need to be quantized. Therefore, the codebook can be equivalently descibed by $N = 2^B$ vectors $\{\hat{\mathbf{x}}_1, \ldots \hat{\mathbf{x}}_N\}$, with $\hat{\mathbf{x}}_i \in \mathbb{R}^{t-1}$.

When the channel $\mathbf{h}$ is i.i.d. Rayleigh distributed, the gain vector $\mathbf{r}$ contains i.i.d. standard Chi-distributed entries, and the phase vector $\mathbf{x}$ contains i.i.d.

entries that are uniformly distributed on $[-\pi, \pi)$, and the gain vector and the phase vector are statistically independent. Then, it can be shown that the conventionally optimized point density (i.e., optimum when there are no feedback channel errors) is uniform, i.e., $\lambda(\mathbf{x}) = 1/(2\pi)^{t-1}$. In this subsection, the relative SNR loss performance of multiple antenna systems with EGT due the noisy feedback channel is derived by evaluating each of the three terms in (5.4). For convenience, define the notation $E_d \approx E_d^{(1)} + E_{d,1}^{(3)} + E_{d,2}^{(3)}$, where the three terms are

$$E_d^{(1)} = N\epsilon(N) \int_{\mathbf{r}} \int_{\mathbf{x}} \int_{\mathbf{y}} d_1(\mathbf{x}, \mathbf{y}; \mathbf{r}) f_{\mathbf{x}}(\mathbf{x}) \lambda(\mathbf{y}) f_{\mathbf{r}}(\mathbf{r}) d\mathbf{y} d\mathbf{x} d\mathbf{r}. \tag{5.12}$$

$$\begin{aligned} E_{d,1}^{(3)} = {} & \frac{N^{\frac{-2}{t-1}} \kappa_{t-1}^{\frac{-2}{t-1}}}{2(t+1)} N\epsilon(N) \int_{\mathbf{r}} \int_{\mathbf{x}} \lambda^{\frac{-2}{t-1}}(\mathbf{x}) \left| D_1(\mathbf{x}, \mathbf{x}; \mathbf{r}) \right|^{\frac{1}{t-1}} \\ & \operatorname{tr} \left(D_1^{-1}(\mathbf{x}, \mathbf{x}; \mathbf{r}) \left[\int_{\mathbf{y}} D_1(\mathbf{x}, \mathbf{y}; \mathbf{r}) \lambda(\mathbf{y}) d\mathbf{y} \right] \right) f_{\mathbf{x}}(\mathbf{x}) f_{\mathbf{r}}(\mathbf{r}) \, d\mathbf{x} d\mathbf{r} \end{aligned} \tag{5.13}$$

$$\begin{aligned} E_{d,2}^{(3)} = {} & \frac{(t-1) N^{\frac{-2}{t-1}} \kappa_{t-1}^{\frac{-2}{t-1}}}{2(t+1)} \left(1 - N\epsilon(N)\right) \\ & \cdot \int_{\mathbf{r}} \int_{\mathbf{x}} \lambda^{\frac{-2}{t-1}}(\mathbf{x}) \left| D_1(\mathbf{x}, \mathbf{x}; \mathbf{r}) \right|^{\frac{1}{t-1}} f_{\mathbf{x}}(\mathbf{x}) f_{\mathbf{r}}(\mathbf{r}) \, d\mathbf{x} d\mathbf{r}. \end{aligned} \tag{5.14}$$

Since the phase angles are i.i.d. and uniformly distributed on $[-\pi, \pi)$, hence, $f_{\mathbf{x}}(\mathbf{x}) = 1/(2\pi)^{(t-1)}$ for $x_i \in [-\pi, \pi)$. Also, with the uniform point generating density, $\lambda(\mathbf{x}) = 1/(2\pi)^{(t-1)}$. Thus,

$$E_d^{(1)} = \frac{N\epsilon(N)}{(2\pi)^{2(t-1)}} \int_{\mathbf{r}} \int_{x_i, y_i} \left(1 - \frac{\left| [1 \exp\left(j \left(\mathbf{x} - \mathbf{y}\right)\right)] \mathbf{r} \right|^2}{s_{\mathbf{r}}^2} \right) f_{\mathbf{r}}(\mathbf{r}) d\mathbf{y} d\mathbf{x} d\mathbf{r}, \tag{5.15}$$

where the inner integral is over the range $x_i, y_i \in [-\pi, \pi)$. After simplification, the above expression reduces to

$$E_d^{(1)} = N\epsilon(N) \left(1 - L_t\right), \tag{5.16}$$

where the constant L_t is defined as

$$L_t \triangleq \int_{\mathbf{r}} \frac{\sum_{i=1}^{t} r_i^2}{\left(\sum_{i=1}^{t} r_i\right)^2} f_{\mathbf{r}}(\mathbf{r}) \, d\mathbf{r}. \tag{5.17}$$

The integral in the above expression can be easily evaluated by numerical integration for different values of t, since $\mathbf{r}$ is a t-dimensional i.i.d. standard Chi-distributed random vector, and is listed in Table 5.1 for several values of t.

Next, to simplify the $E_{d,1}^{(3)}$ term, given by (5.13), the $(t-1)\times(t-1)$ Hessian matrix $D_1(\mathbf{x},\mathbf{y};\mathbf{r})$ needs to be evaluated. Define $z(\mathbf{v}) \triangleq [\underline{\mathbf{0}}\ \mathrm{diag}(\mathbf{v})]$ where $\mathbf{v} \in \mathbb{C}^{t-1}$ and $\underline{\mathbf{0}}$ is a $(t-1)\times 1$ vector of zeros, and $u(\mathbf{x};\mathbf{r}) \triangleq [1\ \exp(j\mathbf{x})]\mathbf{r} \in \mathbb{C}^1$ and finally $\mathbf{r}_s = [r_2,\ldots,r_t]^T$, then, through straightforward albeit tedious differentiation it can be shown that

$$\begin{aligned} D_1(\mathbf{x},\mathbf{y};\mathbf{r}) &= \frac{z(\exp(j(\mathbf{x}-\mathbf{y})))z\left(u\left(\mathbf{x}-\mathbf{y};\mathbf{r}\right)\mathbf{r}_s\right)^T}{s_{\mathbf{r}}^2} \\ &- \frac{z(\exp(j(\mathbf{x}-\mathbf{y})))\mathbf{r}\mathbf{r}^T z(\exp(j(\mathbf{y}-\mathbf{x})))^T}{s_{\mathbf{r}}^2} \\ &+ \frac{z(\exp(j(\mathbf{y}-\mathbf{x})))z\left(u\left(\mathbf{y}-\mathbf{x};\mathbf{r}\right)\mathbf{r}_s\right)^T}{s_{\mathbf{r}}^2} \\ &- \frac{z(\exp(j(\mathbf{y}-\mathbf{x})))\mathbf{r}\mathbf{r}^T z(\exp(j(\mathbf{x}-\mathbf{y})))^T}{s_{\mathbf{r}}^2} \end{aligned}$$

From the above expression, it is can be shown that

$$\int_{\mathbf{y}} D_1(\mathbf{x},\mathbf{y};\mathbf{r})\lambda(\mathbf{y})d\mathbf{y} = \underline{\mathbf{0}} \tag{5.18}$$

where $\underline{\mathbf{0}}$ here is a $(t-1)\times(t-1)$ matrix of zeros. Thus, the $E_{d,1}^{(3)}$ term given by (5.13) is equal to zero. It remains to find the $E_{d,2}^{(3)}$ term given by (5.14). Since $D_1(\mathbf{x},\mathbf{y};\mathbf{r})$ depends only on $\mathbf{x}-\mathbf{y}$, $D_1(\mathbf{x},\mathbf{x};\mathbf{r}) = D_1(\underline{\mathbf{0}},\underline{\mathbf{0}},\mathbf{r}) \triangleq D_2(\mathbf{r})$, where

$$D_2(\mathbf{r}) = 2\left(\frac{s_{\mathbf{r}}\mathrm{diag}(\mathbf{r}_s) - \mathbf{r}_s\mathbf{r}_s^T}{s_{\mathbf{r}}^2}\right), \tag{5.19}$$

where $\mathbf{r}_s \triangleq [r_2,\ldots,r_t]^T$, as before. From this, it can be shown that

$$|D_2(\mathbf{r})|^{\frac{1}{t-1}} = 2\left(\prod_{i=1}^{t}\frac{r_i}{s_{\mathbf{r}}}\right)^{\frac{1}{t-1}} \tag{5.20}$$

Substituting in (5.14),

$$\begin{aligned} E_{d,2}^{(3)} &= \frac{(t-1)N^{\frac{-2}{t-1}}\kappa_{t-1}^{\frac{-2}{t-1}}}{2(t+1)}\left(1-N\epsilon(N)\right)\left[\int_{\mathbf{r}}|D_2(\mathbf{r})|^{\frac{1}{t-1}}\,d\mathbf{r}\right]\int_{\mathbf{x}}\lambda^{\frac{-2}{t-1}}(\mathbf{x})f_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x} \\ &= \frac{(2\pi)^2(t-1)N^{\frac{-2}{t-1}}\kappa_{t-1}^{\frac{-2}{t-1}}M_t}{(t+1)}\left(1-N\epsilon(N)\right), \end{aligned} \tag{5.21}$$

where M_t is defined as

$$M_t \triangleq \int_{\mathbf{r}} \left(\prod_{i=1}^{t} \frac{r_i}{s_{\mathbf{r}}} \right)^{\frac{1}{t-1}} f_{\mathbf{r}}(\mathbf{r})\, d\mathbf{r} \tag{5.22}$$

and can be evaluated using numerical integration without difficulty, and is tabulated in Table 5.1 for several values of t. In summary, the relative loss in SNR with of quantized EGT with feedback over an SEC is given by

$$E_d \approx N\epsilon(N)\left(1 - L_t\right) + N^{\frac{-2}{t-1}}\left(1 - N\epsilon(N)\right)\left(\frac{(2\pi)^2 (t-1)\kappa_{t-1}^{\frac{-2}{t-1}} M_t}{(t+1)}\right), \tag{5.23}$$

where the constants L_t and M_t are determined by numerical integration,and are listed in Table 5.1 for a few values of t.

When $\epsilon(N) = 0$, i.e., for a noiseless feedback channel, the above equation yields the distortion due to the source quantization only, as follows

$$E_d \approx N^{\frac{-2}{t-1}}\left(\frac{(2\pi)^2 (t-1)\kappa_{t-1}^{\frac{-2}{t-1}} M_t}{(t+1)}\right). \tag{5.24}$$

This is an alternative expression for the high-rate result compared to that in Chapter 3, where the high-rate distortion is derived using an entirely different approximation for the quantization cell (as an ellipsoidal cap in the $\mathbf{w}$-space), and is reproduced here for convenience

$$E_{d,\text{old}} \approx N^{\frac{-2}{t-1}}\left(\frac{2(t-1)}{t+1}\right). \tag{5.25}$$

Although the expressions in (5.24) and (5.25) look different, they yield similar values. If the ratio of E_d to $E_{d,\text{old}}$ is defined as $U_t \triangleq 2\pi^2 \kappa_{t-1}^{\frac{-2}{t-1}} M_t$, clearly, if $U_t \approx 1$, the two expressions above would yield approximately the same value. Table 5.1 lists the value of U_t for a few values of t, and it is seen that the error is less than 2% for $t \leq 32$, although the two expressions were obtained using completely different approximations.

Notice from (5.23) that the overall distortion is not exactly described by the sum of the distortions incurred due to the channel errors (inter-codepoint

distortion) and that incurred due to quantization errors (intra-codepoint distortion), due to the $\epsilon(N)$ factor in the second term. This is unlike the case when mean-squared error is used as performance metric [81].

Table 5.1: Values of L_t, M_t and U_t for different values of t. L_t and M_t are coefficients that determine the high-rate performance of the VQ, and $U_t \approx 1$ shows that the expression in this chapter is in agreement with the one in Chapter 3.

t	2	3	4	5	6	8	16	32
L_t	0.5708	0.3964	0.3028	0.2448	0.2055	0.1553	0.0787	0.0396
M_t	0.2148	0.1638	0.1336	0.1134	0.0986	0.0787	0.0439	0.0236
U_t	1.0599	1.0290	1.0155	1.0072	1.0021	0.9961	0.9857	0.9817

5.4.4 Asymptotic Behavior

From the previous subsection, as N increases, the overall distortion (5.23) is dominated by the behavior of $N\epsilon(N)$ relative to $N^{\frac{-2}{t-1}}$. That is,

1. If $N\epsilon(N) = o\left(N^{\frac{-2}{t-1}}\right)$, the error is dominated by the second term in (5.23), i.e., channel errors play an insignificant role in the asymptotic distortion, and the high-rate distortion is given by (5.24).

2. If $N^{\frac{-2}{t-1}} = o\left(N\epsilon(N)\right)$, the error is dominated by the channel errors (i.e., the first term in (5.23)).

   For example, if a constant probability of index reception is maintained as N increases, $P_{i|i} = 1 - \rho$, then $\epsilon(N) = \rho/(N-1)$. In this case, it is interesting to note that for large N the error is approximately given by $E_d \approx \rho\,(1 - L_t)$, and it can be seen from (5.15) by substituting for any continuous $\lambda(\mathbf{x})$ instead of the uniform point density, that the point density does not affect the asymptotic performance, as long as it is continuous.

   As another example, when the feedback channel is a BSC, as seen earlier, $a(N) = \left(1 - (1-q)^B\right)/(N-1)$ after averaging the performance over

all possible index assignments. It can be readily verified that as N gets large, $N\epsilon(N)$ is $O(1)$. Thus, the distortion is asymptotically given by $E_d \approx (1 - L_t)$. This distortion is in fact exactly that obtained if the transmitter were to ignore the feedback information and employ a *single, fixed* beamforming vector. Indeed, the distortion with fixed beamforming is given by

$$\begin{aligned} E_d^{\text{fix}} &= \mathbf{E}\left\{1 - \frac{\left|\underline{1}^T\mathbf{h}\right|^2}{\|\mathbf{h}\|_1^2}\right\}, \\ &= 1 - \frac{1}{(2\pi)^{(t-1)}}\int_{\mathbf{r}}\int_{\mathbf{x}}\left(\frac{\left|\underline{1}^T\text{diag}\left([1\,\exp\left(j\mathbf{x}\right)]\right)\mathbf{r}\right|^2}{s_{\mathbf{r}}^2}\right) f_{\mathbf{r}}(\mathbf{r})d\mathbf{x}d\mathbf{r}, \end{aligned}$$

where without loss of generality, the fixed beamforming vector has been chosen to be all-ones vector $\underline{1}$ (due to the left rotational invariance property of the density of $\mathbf{h}$, the choice of the fixed beamforming vector does not matter). The expression for $\mathbf{h}$ in (5.11) was used to obtain the last expression above. Integrating over $\mathbf{x}$ and simplifying,

$$\begin{aligned} E_d^{\text{fix}} &= 1 - \int_{\mathbf{r}}\left(\frac{\sum_{i=1}^t r_i^2}{s_{\mathbf{r}}^2}\right) f_{\mathbf{r}}(\mathbf{r})d\mathbf{r}, \\ &= 1 - L_t, \end{aligned} \tag{5.26}$$

where L_t is defined as in (5.17). Thus, the performance of quantized EGT systems when the index is sent over a BSC with cross-over probability $q > 0$ and random index assignment asymptotically approaches that achieved by using a fixed beamforming vector at the transmitter, i.e., feedback is useless when N is large, if no channel coding is employed. The reason for this behavior is because as N increases, the probability of successful index reception, $(1-q)^B$, becomes small. Then, it is no longer optimum to use a uniformly distributed point density; i.e., it is more efficient to trade off quantization error for a better channel coding gain, i.e., the uniform (or continuous) point density is inefficient for large N when the feedback is sent over a BSC. This is in agreement with a related result derived in [83] in the source coding context, where the authors show that for mean squared

error distortion, with a BSC and random index assignment, the distortion asymptotically approaches the source variance.

3. Finally, if $N\epsilon(N) = \Theta\left(N^{\frac{-2}{t-1}}\right)$, then both terms decrease at the same rate $N^{\frac{-2}{t-1}}$ as N increases.

Interestingly, the above observations imply that asymptotically, the overall distortion can in fact be tightly bounded by the sum of the distortion due to channel errors and the distortion due to quantization errors, as follows

$$\begin{aligned} E_d &\approx N\epsilon(N)\left(1-L_t\right) + N^{\frac{-2}{t-1}}\left(1 - N\epsilon(N)\right)H_t \\ &\lessapprox N\epsilon(N)\left(1-L_t\right) + N^{\frac{-2}{t-1}}H_t, \end{aligned} \tag{5.27}$$

where $H_t \triangleq \left(\frac{(2\pi)^2(t-1)\kappa_{t-1}^{\frac{-2}{t-1}}M_t}{(t+1)}\right)$. The upper bound is tight because, as N gets large, $N^{\frac{-2}{t-1}}N\epsilon(N)$ will always be dominated by either $N\epsilon(N)$ or $N^{\frac{-2}{t-1}}$. For the case of the BSC, the upper bound can be used to determine the value of N that minimizes the overall distortion. Indeed, when q is small,

$$N\epsilon(N) = \frac{N\left(1-(1-q)^B\right)}{N-1} \approx qB, \tag{5.28}$$

and thus, $E_d \approx q\log_2(N)\left(1-L_t\right) + N^{\frac{-2}{t-1}}H_t$. Differentiating with respect to N and setting equal to zero, it can be shown that

$$N_{\text{opt}} = \left(\frac{2H_t\log 2}{q(t-1)(1-L_t)}\right)^{\frac{t-1}{2}}, \tag{5.29}$$

and the optimum number of bits per dimension is $B_{\text{opt}} = \log_2(N_{\text{opt}})/(t-1)$, as there are $t-1$ free dimensions per vector. Table 5.2 shows the value of B_{opt} for different values of t and q. As the bit error probability and the number of transmit antennas get large, channel quantization and feedback result in no performance improvement due to the channel errors, hence employing a fixed beamforming vector irrespective of the channel state achieves lowest distortion. Finally, it is also interesting to note that in the case where the probability of an *index error* is

Table 5.2: The optimum number of bits per dimension to minimize the overall distortion, for a BSC with different values of the cross-over probability q.

B_{opt}	t = 2	3	4	5	6	8	16	32	48	64
$q = 10^{-5}$	9	8	8	8	8	8	7	7	6	6
10^{-4}	7	7	7	6	6	6	5	5	5	4
10^{-3}	6	5	5	5	5	4	4	3	3	3
10^{-2}	4	3	3	3	3	3	2	2	1	1
10^{-1}	2	2	2	1	1	1	1	0	0	0

kept fixed as N increases, i.e., $N\epsilon(N) \approx \rho$ independent of N, the overall distortion is approximately given by

$$E_d \approx \rho\,(1 - L_t) + N^{\frac{-2}{t-1}}\,(1 - \rho)\,H_t, \tag{5.30}$$

from which it is clear that the asymptotic distortion decreases monotonically as N increases, and approaches $\rho(1 - L_t)$.

5.5 Wideband Speech Spectrum Compression

In this section, the high-rate vector quantization results from Sec. 5.3 are applied to the quantization the LPC parameters of speech systems for noisy SEC. For LPC quantization, the LSD is often cited [90] as a measure that correlates well with speech quality. Hence, in this section, the distortion is measured as the LSD, and the sensitivity matrices with respect to the LSD measure are computed. Here, a very brief overview of the LPC quantization is provided, and interested readers are referred to [91] for complete motivations and theoretical details. Let the set of filter taps corresponding to a v-th order LPC filter be denoted by the vector $\mathbf{x} = [x_1, x_2, \ldots, x_v]^T$. Note that the zeroth tap, normally constrained to be equal to unity, is not included in $\mathbf{x}$. The filter taps are obtained in a straightforward manner from the autocorrelation of the speech samples; the details are omitted

here. Then, the LSD in dB^2 incurred by quantizing the vector $\mathbf{x}$ to $\hat{\mathbf{x}}$ is given by

$$L(\mathbf{x}, \hat{\mathbf{x}}) \triangleq \frac{\beta}{2\pi} \int_{-\pi}^{\pi} \left[\log \left(\frac{\mathbf{x}^T B(\omega) \mathbf{x}}{\hat{\mathbf{x}}^T B(\omega) \hat{\mathbf{x}}} \right) \right]^2 d\omega, \tag{5.31}$$

where $\beta = (10/\log(10))^2$ and the $v \times v$ matrix $B(\omega)$ has $\cos(\omega(i-j))$ as its (i,j)-th element.

5.5.1 Sensitivity Matrices for LPC Coefficients

Let $X(z)$ denote the z-transform of the discrete-time filter with tap coefficients given by $\mathbf{x}$, and let $h(n)$ denote the impulse response of the discrete-time filter $1/X(z)$. In [78], it is shown that

$$D_{k,l}(\hat{\mathbf{x}}, \hat{\mathbf{x}}) \triangleq \left. \frac{\partial^2 L(\mathbf{x}, \hat{\mathbf{x}})}{\partial x_k \partial x_l} \right|_{\mathbf{x}=\hat{\mathbf{x}}} = 4\beta R_X(k-l), \tag{5.32}$$

where $R_X(k)$ is the autocorrelation function of the impulse response $h(n)$, i.e.,

$$R_X(k) = \sum_{n=0}^{\infty} h(n)h(n+k). \tag{5.33}$$

However, in order to apply the results from the previous section, in addition to $D(\mathbf{x}, \mathbf{x})$, the cross-sensitivity matrix $D(\mathbf{x}, \mathbf{y})$ with k, l-th element defined by

$$D_{k,l}(\mathbf{x}, \mathbf{y}) = \left. \frac{\partial^2 L(\hat{\mathbf{x}}, \mathbf{y})}{\partial \hat{x}_k \partial \hat{x}_l} \right|_{\hat{\mathbf{x}}=\mathbf{x}} \tag{5.34}$$

needs to be computed. In the Appendix 5.8.1, it is shown that,

$$D(\mathbf{x}, \mathbf{y}) = D(\mathbf{x}, \mathbf{x}) + \frac{\beta}{2\pi} \int_{-\pi}^{\pi} \frac{B(\omega)}{\mathbf{x}^T B(\omega) \mathbf{x}} \log \left(\frac{\mathbf{x}^T B(\omega) \mathbf{x}}{\mathbf{y}^T B(\omega) \mathbf{y}} \right) d\omega \tag{5.35}$$

Substituting this into (5.3) and (5.7), the distortion with channel errors is obtained. However, since the source density is unknown, the expectations in (5.3) and (5.7) must be evaluated via a Monte-Carlo method, which will be described in greater detail later. In Sec. 5.6, simulation results show that although the above derived theoretical expressions can be utilized to correctly estimate the amount of channel error that can be tolerated, given an upper-limit on the allowable distortion (about 1dB^2 is considered transparent quality for speech).

5.6 Simulation Results

5.6.1 Equal Gain Transmission

First, the problem of quantized equal gain beamforming described in Sec. 5.4 is considered, and it is shown that the theoretical curves agree well with the simulation based ones. The $1 \times t$ MISO channel is assumed to be i.i.d. Rayleigh flat fading, and 10,000 random instantiations were used with the Lloyd algorithm proposed in Chapter 3 to generate the beamforming codebook.

First, consider the noiseless case, i.e., $\epsilon(N) = 0$. Fig.5.2 shows the loss in SNR relative to perfect feedback versus the number of feedback bits B. The simulation results agree well with the theoretical expression of (5.23). Also, note that about 2 bits per dimension is a good rule-of-thumb to ensure that the high rate approximations become accurate. Next, consider the feedback channel with a fixed probability of success, i.e., $a(N) = \rho/(N-1)$, where $0 \leq \rho \leq (N-1)/N$ is a parameter. In this case, from (5.23), the relative loss in SNR is given by

$$E_d \approx \frac{N\rho}{N-1}\left(1 - L_t\right) + N^{\frac{-2}{t-1}}\left(1 - \frac{N\rho}{N-1}\right)\left(\frac{(2\pi)^2\,(t-1)\kappa_{t-1}^{\frac{-2}{t-1}} M_t}{(t+1)}\right). \qquad (5.36)$$

Fig.5.3 plots the relative loss in SNR versus the parameter ρ with the number of quantization levels fixed at $N = 16$ (i.e., $B = 4$ bits). Notice that as ρ gets close to 1, for large N the distortion approaches $E_d \approx \rho(1 - L_t)$, i.e., it is linear in ρ. Also, for a given N, as ρ gets smaller, the performance improves initially, but after a point, it is determined by the $E_{d,2}^{(3)}$ term (i.e., due to the source distortion). In Fig. 5.4, the relative loss in SNR versus the number of feedback bits B is plotted, with ρ fixed at $10^{-1.5} \approx 0.0316$. The theoretical curves are generated using the expression in (5.36). Again, we see that the theoretical curves agree well with the experimental curves as B increases. Also, for a fixed ρ, the performance is eventually dominated by the $E_d^{(1)}$ term of (5.16) as B gets large.

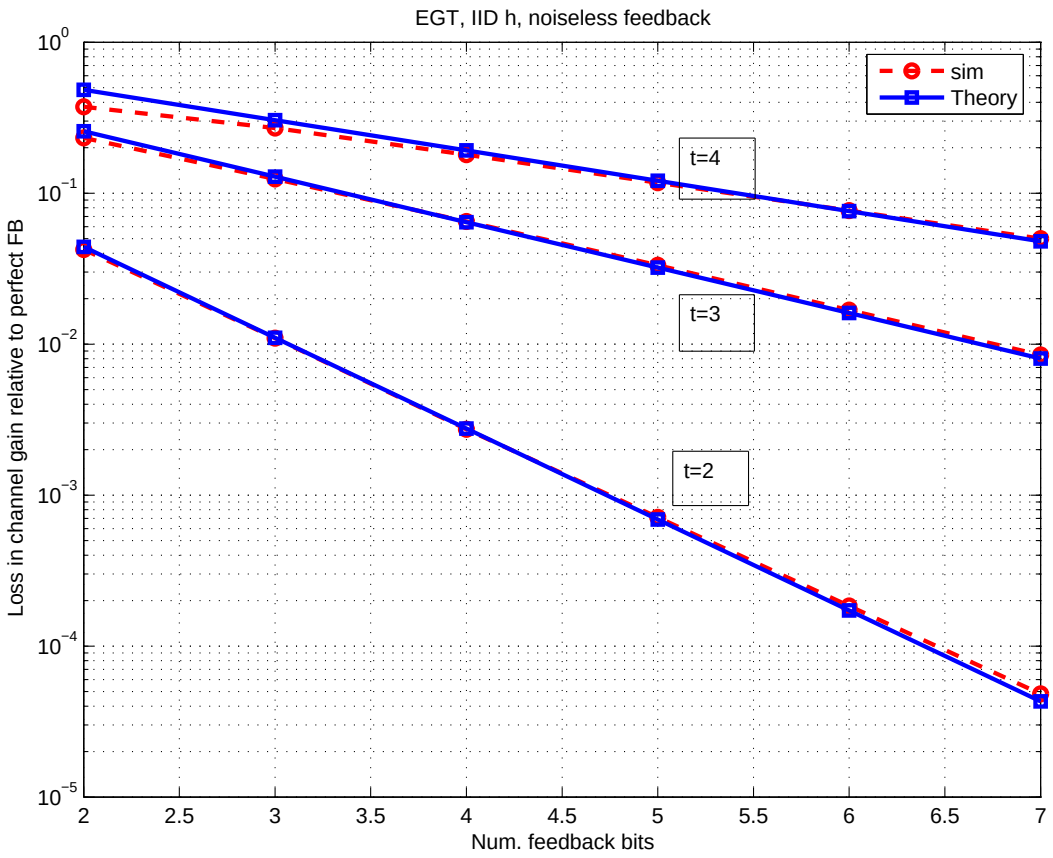


Figure 5.2: Loss in gain relative to perfect feedback, with a noiseless feedback channel, versus the number of feedback bits B.

5.6.2 Wideband Speech Compression

Next, an experiment is performed on wideband speech spectrum coding, under the LSD measure. The goal in this context is to achieve an average distortion of 1dB^2. This experiment is designed to determine at what error rates this goal is feasible. A database of 16-dimensional wideband speech spectrum vectors is gathered, and their sensitivities are evaluated using the method described in Sec. 5.5. For sources with such a large dimension, the codebook sizes become very large (around 50 bits). This rules out the use of full-search vector quantizers, and structured systems must instead be used to reduce the complexity. To this end, the Gaussian Mixture Model (GMM) based VQ described in [89] is employed. This

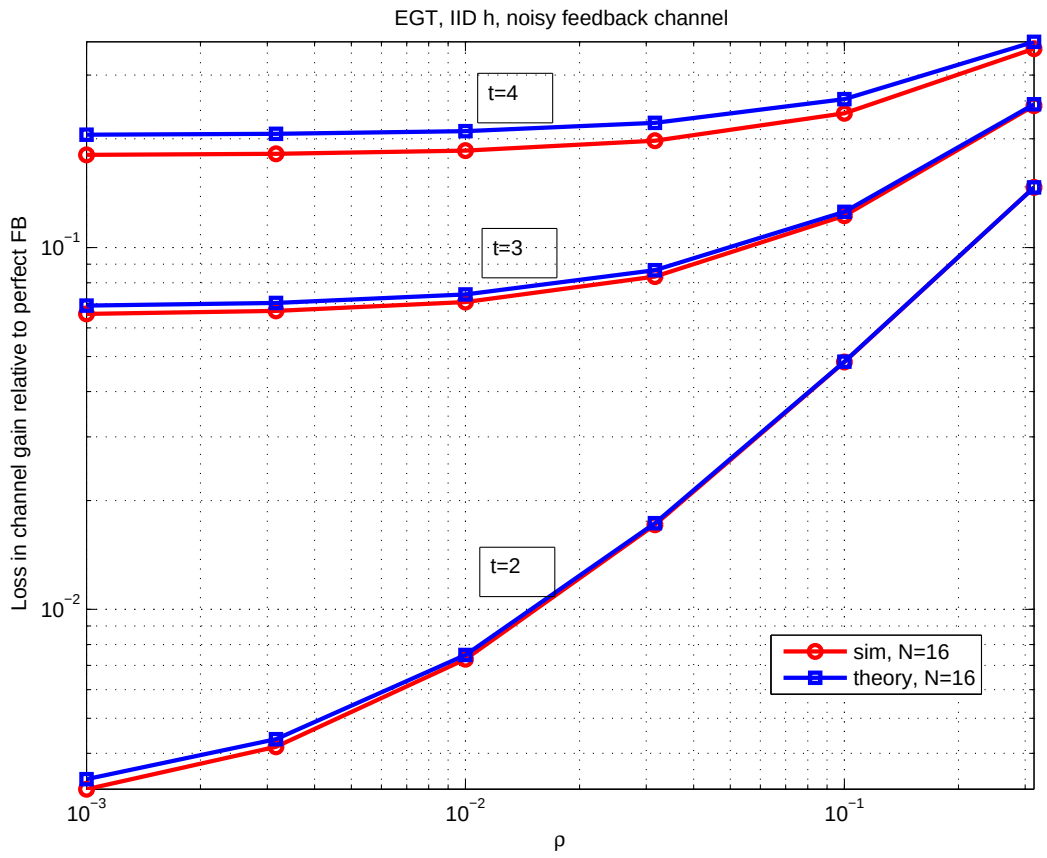


Figure 5.3: Loss in gain relative to perfect feedback versus ρ, where ρ is the parameter that determines the transition probability of the SEC of (5.1) when $\epsilon(N) = \rho/(N-1)$. Here, the number of quantization levels N is kept fixed at 16.

system is able to operate with a low, rate-independent complexity, although it is suboptimal in the sense that its cells are not ellipsoidal; and as a result there is a small gap between the theoretical and experimental distortion curves.

The point density of the quantizer is itself a GMM, with parameters specified through the training process. Thus, the integral in (5.7) over $\mathbf{y}$ can be approximated by averages over data $\mathbf{y}$ drawn randomly according to the point density. A database of 65000 source vectors $\mathbf{x}_i$ is employed, along with a database of 65000 "error" vectors $\mathbf{y}_i$, drawn according to $\lambda(\mathbf{y})$ for the Monte-Carlo estimation. From figure 5.7, it can be seen that the theory is good at predicting the true high-

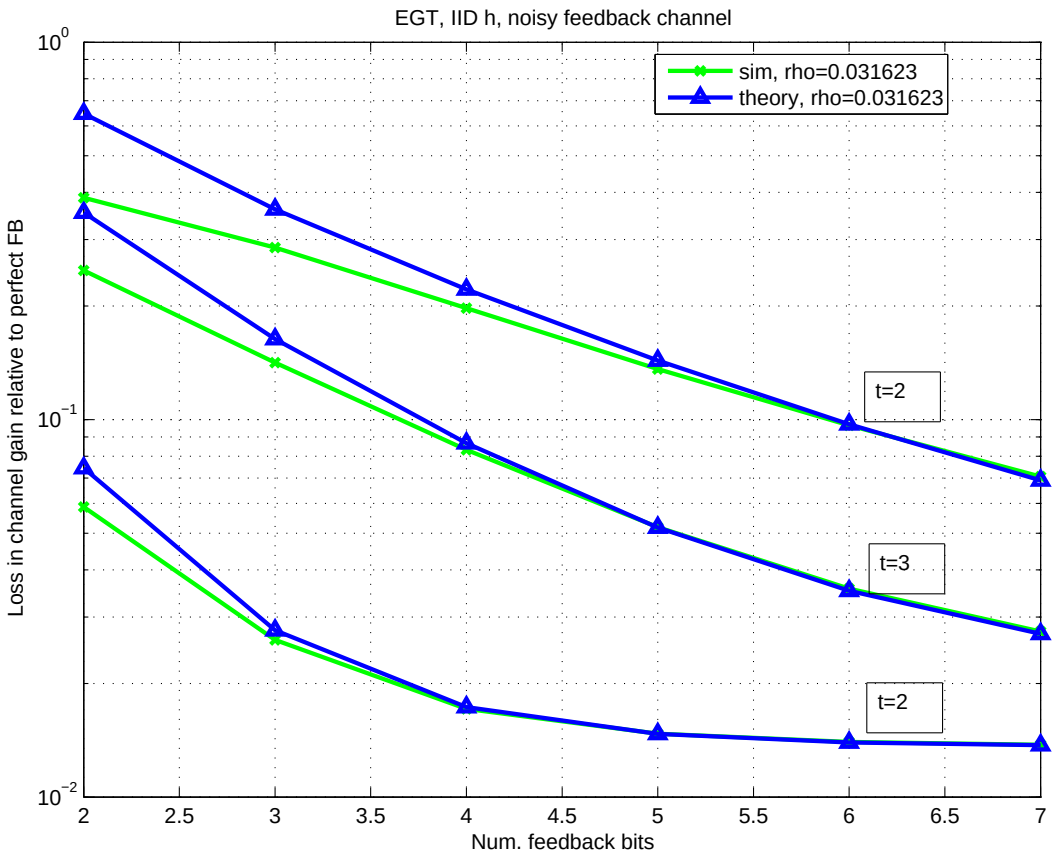


Figure 5.4: Loss in gain relative to perfect feedback versus the number of feedback bits B, where ρ is the parameter that determines the transition probability of the SEC of (5.1) when $\epsilon(N) = \rho/(N-1)$. Here, ρ is kept fixed at $10^{-1.5}$.

rate distortion. Observe that for high values of the error probability, it is impossible to perform high-quality quantization of this source, with the error levelling off at around 3dB^2. For moderate error probabilities, 1dB^2 LSD can be achieved, although a few extra bits will be required to compensate for channel errors. At low values of the error probability, there is no penalty, as the channel effects do not become significant until well beyond the desired 1dB^2 operating point. Thus, a channel error probability of at most 0.001 is judged to be permissible for the wideband speech spectrum quantization problem.

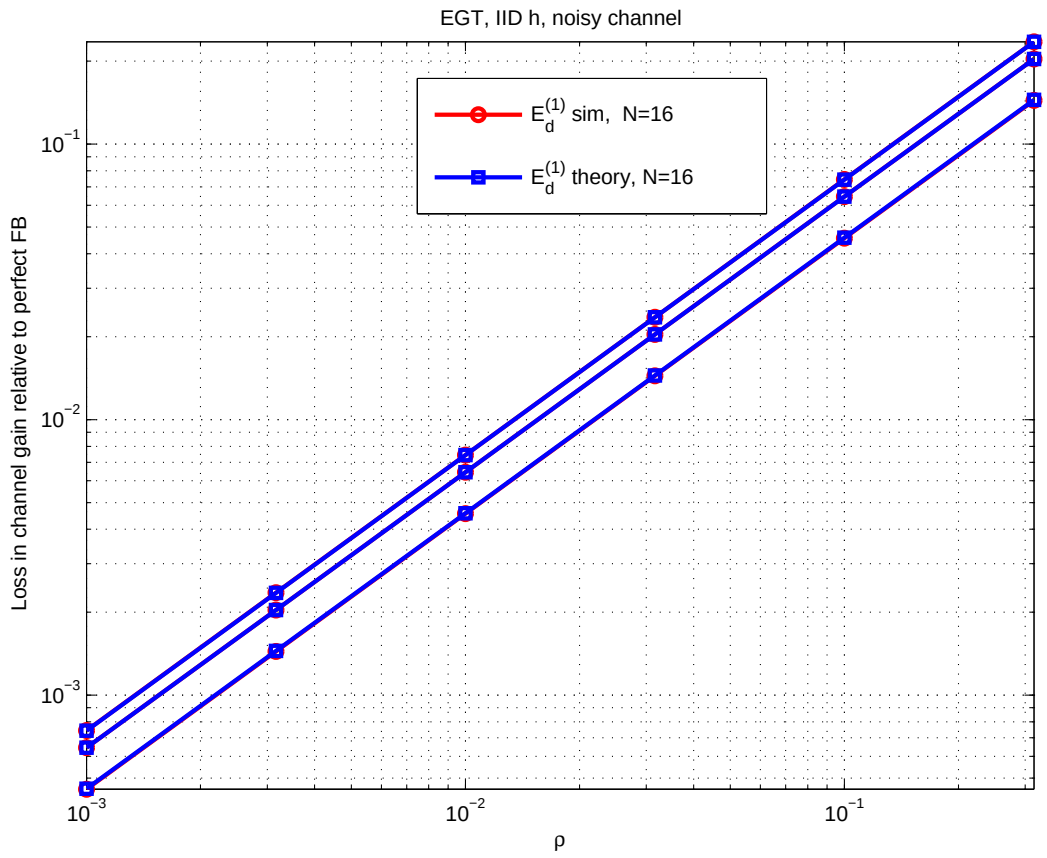


Figure 5.5: The term $E_d^{(1)}$ versus ρ, where ρ is the parameter that determines the transition probability of the SEC of (5.1) when $\epsilon(N) = \rho/(N-1)$. Here, the number of quantization levels N is kept fixed at 16.

5.7 Conclusions

In this chapter, the source quantization problem where the quantized index is sent over a noisy channel before being reproduced at the receiver was considered. For the case of the simplex error channel, a theoretical framework for the asymptotic performance analysis with channel errors and arbitrary distortion functions was presented, and applied to the problem of quantizing the phase angle information in equal gain beamforming. Theoretical expressions were derived for the asymptotic loss in SNR performance in the presence of channel errors. The

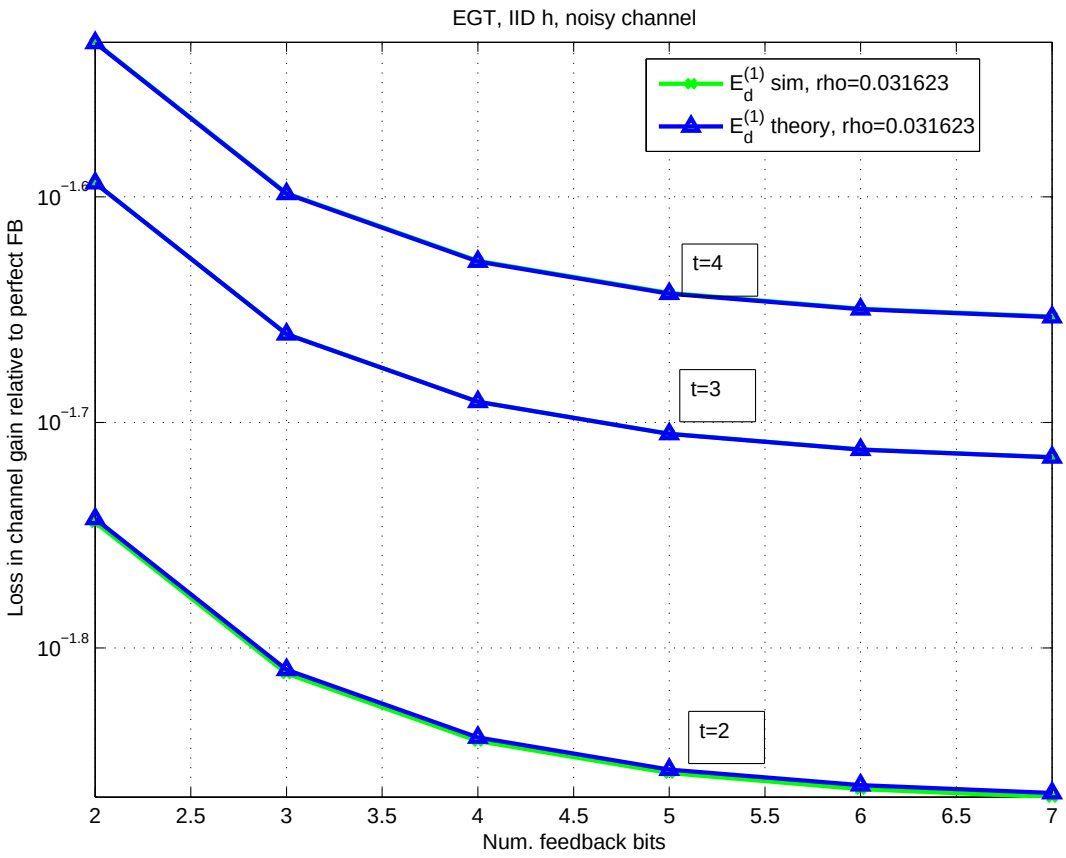


Figure 5.6: The term $E_d^{(1)}$ versus the number of feedback bits B, where ρ is the parameter that determines the transition probability of the SEC of (5.1) when $\epsilon(N) = \rho/(N-1)$. Here, ρ is kept fixed at $10^{-1.5}$.

framework was also applied to the problem of wideband speech spectrum quantization under the LSD measure, and seen to accurately characterize the effects of the source, the quantizer and the channel. The accuracy of the theoretical results were further illustrated through Monte-Carlo simulation.

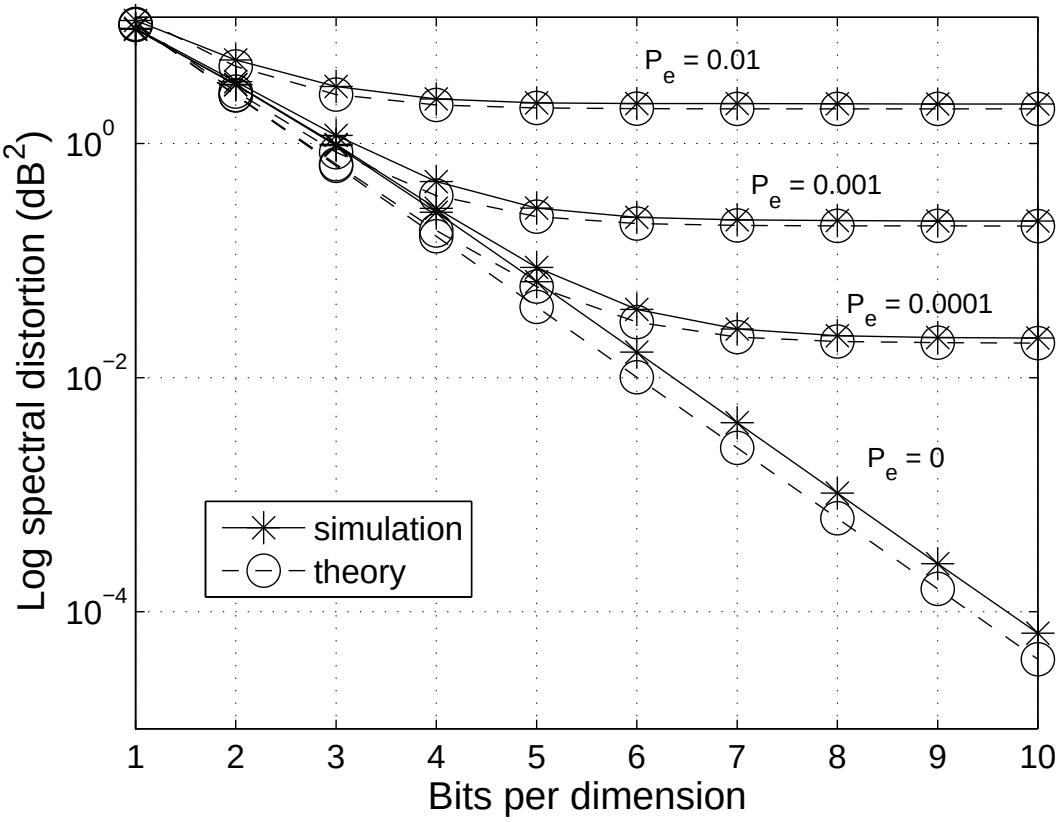


Figure 5.7: Log Spectral Distortion on Wideband Speech LSF vectors versus B. Both predicted and actual distortions are shown for several values of P_e, the total probability of an index error.

5.8 Appendix

5.8.1 Derivation of the LPC Sensitivity Matrices

In this section, it is shown that, when the distortion is given by

$$L(\mathbf{x},\mathbf{y}) \triangleq \frac{\beta}{2\pi}\int_{-\pi}^{\pi}\left[\log\left(\frac{\mathbf{x}^T B(\omega)\mathbf{x}}{\mathbf{y}^T B(\omega)\mathbf{y}}\right)\right]^2 d\omega, \tag{5.37}$$

the sensitivity matrix $D(\mathbf{x},\mathbf{y})$ defined in 5.5 is given by

$$D(\mathbf{x},\mathbf{y}) = D(\mathbf{x},\mathbf{x}) + \frac{\beta}{2\pi}\int_{-\pi}^{\pi}\frac{B(\omega)}{\mathbf{x}^T B(\omega)\mathbf{x}}\log\left(\frac{\mathbf{x}^T B(\omega)\mathbf{x}}{\mathbf{y}^T B(\omega)\mathbf{y}}\right) d\omega. \tag{5.38}$$

The proof is straightforward from first principles. Define $B_{\mathbf{y}}(\omega)$ as

$$B_{\mathbf{y}}(\omega) \triangleq \frac{B(\omega)}{\mathbf{y}^T B(\omega)\mathbf{y}}. \tag{5.39}$$

Then,

$$L(\mathbf{x},\mathbf{y}) = \frac{\beta}{2\pi}\int_{-\pi}^{\pi}\left[\log\left(\mathbf{x}^T B_{\mathbf{y}}(\omega)\mathbf{x}\right)\right]^2 d\omega. \tag{5.40}$$

To find $D(\mathbf{x},\mathbf{y}) \triangleq \partial^2 L(\hat{\mathbf{x}},\mathbf{y})/\partial\hat{\mathbf{x}}^2|_{\hat{\mathbf{x}}=\mathbf{x}}$, consider the expansion of $L(\hat{\mathbf{x}},\mathbf{y})$ around $\mathbf{x}$, where $\mathbf{x}$ is a vector that is "close to" $\hat{\mathbf{x}}$, and let $\mathbf{e} \triangleq \hat{\mathbf{x}} - \mathbf{x}$. Only the integrand of the above expression is written here for simplicity.

$$\begin{aligned}\log\left(\hat{\mathbf{x}}^T B_{\mathbf{y}}(\omega)\hat{\mathbf{x}}\right) &= \log\Big((\mathbf{x}+\mathbf{e})^T B_{\mathbf{y}}(\omega)(\mathbf{x}+\mathbf{e})\Big)\\ &= \log\left(\mathbf{x}^T B_{\mathbf{y}}(\omega)\mathbf{x}\right) + \log\left(1 + \frac{2\mathbf{e}^T B_{\mathbf{y}}(\omega)\mathbf{x}}{\mathbf{x}^T B_{\mathbf{y}}(\omega)\mathbf{x}} + \frac{\mathbf{e}^T B_{\mathbf{y}}(\omega)\mathbf{e}}{\mathbf{x}^T B_{\mathbf{y}}(\omega)\mathbf{x}}\right)\end{aligned}$$

Note that $B_{\mathbf{y}}(\omega)/(\mathbf{x}^T B_{\mathbf{y}}(\omega)\mathbf{x}) = B_{\mathbf{x}}(\omega)$, and considering a second-order approximation to $\log\left(\hat{\mathbf{x}}^T B_{\mathbf{y}}(\omega)\hat{\mathbf{x}}\right)$,

$$\log\left(\hat{\mathbf{x}}^T B_{\mathbf{y}}(\omega)\hat{\mathbf{x}}\right) = \log\left(\mathbf{x}^T B_{\mathbf{y}}(\omega)\mathbf{x}\right) + 2\mathbf{e}^T B_{\mathbf{x}}(\omega)\mathbf{x} + \mathbf{e}^T B_{\mathbf{x}}(\omega)\mathbf{e}. \tag{5.41}$$

Squaring and isolating just the second order terms to obtain the Hessian matrix,

$$\begin{aligned}\left[\log\left(\hat{\mathbf{x}}^T B_{\mathbf{y}}(\omega)\hat{\mathbf{x}}\right)\right]^2 &\approx \text{Constant and First order terms}\\ &+ 4\mathbf{e}^T B_{\mathbf{x}}(\omega)\mathbf{x}\mathbf{x}^T B_{\mathbf{x}}(\omega)\mathbf{e} + 2\log\left(\mathbf{x}^T B_{\mathbf{y}}(\omega)\mathbf{x}\right)\mathbf{e}^T B_{\mathbf{x}}(\omega)\mathbf{e}\end{aligned}$$

Thus, we have the sensitivity matrix

$$D(\mathbf{x},\mathbf{y}) = \frac{\beta}{2\pi}\int_{-\pi}^{\pi}\left(4B_{\mathbf{x}}(\omega)\mathbf{x}\mathbf{x}^T B_{\mathbf{x}}(\omega) + 2\log\left(\mathbf{x}^T B_{\mathbf{y}}(\omega)\mathbf{x}\right)B_{\mathbf{x}}(\omega)\right)d\omega \tag{5.42}$$

Clearly, when $\mathbf{y} = \mathbf{x}$, we get $D(\mathbf{x},\mathbf{x}) = 4B_{\mathbf{x}}(\omega)\mathbf{x}\mathbf{x}^T B_{\mathbf{x}}(\omega)$, therefore,

$$\begin{aligned}D(\mathbf{x},\mathbf{y}) &= D(\mathbf{x},\mathbf{x}) + 2\frac{\beta}{2\pi}\int_{-\pi}^{\pi} B_{\mathbf{x}}(\omega)\log\left(\mathbf{x}^T B_{\mathbf{y}}(\omega)\mathbf{x}\right)d\omega\\ &= D(\mathbf{x},\mathbf{x}) + 2\frac{\beta}{2\pi}\int_{-\pi}^{\pi}\frac{B(\omega)}{\mathbf{x}^T B(\omega)\mathbf{x}}\log\left(\frac{\mathbf{x}^T B(\omega)\mathbf{x}}{\mathbf{y}^T B(\omega)\mathbf{y}}\right)d\omega,\end{aligned} \tag{5.43}$$

which completes the proof.

Acknowledgement

This chapter, in part, is a reprint of the material which has appeared as C. R. Murthy and B. D. Rao, "Effect of feedback errors on quantized equal gain transmission", *Int. Conf. on Communications (ICC)*, Istanbul, Turkey, Jun. 2006.

6 Conclusions

In this thesis, several aspects of feedback based communication with multiple antennas were considered, primarily in the areas of channel estimation and quantization, and the main contributions are summarized below.

6.1 Contributions of this Thesis

As pointed out earlier, channel estimation is doubly important in feedback-based communication because inaccurate CSI affects not only the receiver performance, but also results in sub-optimal transmission. In this context, MIMO flat-fading channel estimation when the transmitter employs Maximum Ratio Transmission (MRT) was studied. Two competing schemes for estimating the transmit and receive beamforming vectors of the channel matrix were analyzed: a training based conventional least squares estimation (CLSE) scheme and a closed-form semi-blind (CFSB) scheme that employs training followed by information-bearing spectrally white data symbols. Employing matrix perturbation theory, expressions for the mean squared error (MSE) in the beamforming vector, the average received SNR and the symbol error rate (SER) performance of both the semi-blind and the conventional schemes were derived. A weighted linear combiner of the CFSB and CLSE estimates for additional improvement in performance was also proposed.

Another important issue in beamforming-based communication with multiple antennas is the quantization and feedback of CSI. Hence, this dissertation also considered the design and analysis of quantizers for Equal Gain Transmission

(EGT) systems with finite rate feedback-based communication in flat-fading MISO systems. EGT is a beamforming technique that maximizes the MISO channel capacity when there is an equal power-per-antenna constraint at the transmitter, and requires the feedback of $t-1$ phase angles, when there are t antennas at the transmitter. Two popular approaches for quantizing the phase angles were contrasted: vector quantization (VQ) and scalar quantization (SQ). On the VQ side, using the capacity loss with respect to EGT with perfect channel information at transmitter as performance metric, a criterion for designing the beamforming codebook for quantized EGT (Q-EGT) was developed. An iterative algorithm based on the well-known generalized Lloyd algorithm was proposed for computing the beamforming vector codebook. On the analytical side, closed-form expressions were derived for the performance of quantized feedback in terms of capacity loss and outage probability in the case of i.i.d. Rayleigh flat-fading channels.

The next issue addressed in this thesis dealt with the effect of having a noisy feedback channel. Errors in the feedback channel can adversely affect the performance of a quantized-feedback based transmission scheme, because the beamforming vector employed by the transmitter could be very different from the intended beamforming vector due to index errors. With the view to understand the effect of errors on channel quantization, the more general problem of characterizing the high-rate performance of source coding for noisy discrete symmetric channels with random index assignment was considered. Theoretical expressions for the performance of source coding were derived for a large class of distortion measures. It was shown that when the point density is continuous, the high-rate distortion can be approximately expressed as the sum of the source quantization distortion and the channel-error induced distortion, result known previously only for the case of the mean-squared error distortion. Optimization of the point density was also considered. For general distortion functions, assuming that the point density is continuous, expressions were derived for the point density that minimizes the expected distortion. For the mean squared error distortion, an upper bound on the

asymptotic (i.e., high-rate) distortion was derived by assuming a certain structure on the codebook. This structure enabled the extension of the analysis to source coders with *singular* point densities. It was shown that, for channels with small errors, the point density that minimizes the upper bound is continuous, while as the error rate increases, the point density becomes singular, and the extent of the singularity was analytically characterized.

In the final chapter of this thesis, new results on the performance of the high-rate vector quantization of random sources when the quantized index is transmitted over a noisy channel were derived for two specific applications. The first was the quantization of the channel state information for multiple-input, single output systems with beamforming at the transmitter. Here, it was assumed that there exists a per-antenna power constraint at the transmitter, hence, the Equal Gain Transmission (EGT) beamforming vector is quantized and sent from the receiver to the transmitter over a noisy discrete symmetric channel with random index assignment. The loss in received SNR was analytically characterized, and it was shown that at high rates, the overall distortion can be expressed as the sum of the quantization-induced distortion and the channel error-induced distortion. The optimum density of codepoints (also known as the point density) that minimizes the overall distortion subject to a boundedness constraint was shown to be the uniform density. Also, it was found that the asymptotic performance depends on the behavior of the noisy feedback channel as the number of codepoints gets large. The binary symmetric channel with random index assignment was a special case of the analysis, and it was shown that the asymptotic distortion as the number of quantized bits gets large approaches the distortion with random beamforming. The second application was in the wideband speech compression problem, i.e., that of quantizing the linear predictive coding parameters in speech coding systems with the log spectral distortion as performance metric. It was shown that the theory is able to correctly predict the channel error rate that is permissible for operation at a particular distortion level.

6.2 Future Work

Several interesting and important problems remain, some of which are listed below.

- If the channel changes is time-varying, the performance of feedback based communication can be very sensitive to feedback delays. The exact characterization of the effect of the delay in the feedback channel is therefore an interesting issue to be addressed.

- In this dissertation, the channel was assumed to be flat-fading, i.e., it is applicable for narrowband communication. Extension of the channel quantization results to wideband communication scenarios such as OFDM is yet another open topic of research.

- The results derived here assumed point-to-point communication, hence, it would be interesting to consider multi-user, broadcast or multiple-access communication with feedback, and consider the problem of CSI quantization for these scenarios.

- In the channel estimation area, it is interesting to consider the effect of time variation, and analytically characterize the efficacy of using semi-blind estimation schemes as a function of the doppler or mobile speed.

- The design of optimum training sequences with feedback-based communication (since part of the channel is known at the transmitter and receiver, the training sequence needs to help estimate only the unknown part) is another interesting problem.

Solving all of the problems listed above should result in several years of fruitful research.

Bibliography

[1] J. H. Winters, "On the capacity of radio communication systems with diversity in a rayleigh fading environment," *IEEE Jounal on Select. Areas Commn.*, vol. 5, pp. 871 – 878, June 1987.

[2] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, pp. 311–335, Mar. 1998.

[3] I. E. Telatar, "Capacity of multi-antenna Gaussian channels." *AT&T Bell Labs Tech. Memo.*, 1995.

[4] 3GPP TS25.308, "UTRA high speed downlink packet access (HSDPA); overall description; stage 2," *version 5.5.0*.

[5] IEEE P802.11, "TGn sync proposal technical specification," *doc.: IEEE 802.11-04/0889r4*.

[6] IEEE P802.16, "IEEE standard for local and metropolitan area networks part 16: Air interface for fixed broadband wireless access systems," *IEEE Std 802.16-2004*.

[7] T. K. Y. Lo, "Maximum ratio transmission," *IEEE Trans. Commun.*, vol. 47, pp. 1458–1461, Oct. 1999.

[8] D. J. Love and R. W. Heath, Jr., "Equal gain transmission in multiple-input multiple-output wireless systems," *IEEE Trans. Commn.*, vol. 51, pp. 1102–1110, July 2003.

[9] D. Pal, "Fractionally spaced semi-blind equalization of wireless channels," *The Twenty-Sixth Asilomar Conference*, vol. 2, pp. 642–645, 1992.

[10] D. Pal, "Fractionally spaced equalization of multipath channels: a semi-blind approach," *1993 International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. 9–12, 1993.

[11] E. de Carvalho and D. T. M. Slock, "Cramer-Rao bounds for semi-blind and training sequence based channel estimation," *First IEEE Workshop on SPAWC*, pp. 129–32, 1997.

[12] E. de Carvalho and D. T. M. Slock, "Asymptotic performance of ML methods for semi-blind channel estimation," *Thirty-First Asilomar Conference*, vol. 2, pp. 1624–8, 1998.

[13] S. A. Jafar, S. Vishwanath, and A. Goldsmith, "Channel capacity and beamforming for multiple transmit and receive antennas with covariance feedback," in *Proc. IEEE Int. Conf. on Commun. 2001*, pp. 2266–2270, June 2001.

[14] E. Visotsky and U. Madhow, "Space-time transmit precoding with imperfect feedback," *IEEE Trans. Info. Theory*, vol. 47, pp. 2632–2639, Sept. 2001.

[15] E. G. Larsson, G. Ganesan, P. Stoica, and W.-H. Wong, "On the performance of orthogonal space-time block coding with quantized feedback," *IEEE Commun. Letters*, vol. 6, pp. 487–489, Nov. 2002.

[16] S. Zhou and G. B. Giannakis, "Adaptive modulation for multiantenna transmissions with channel mean feedback," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 1626–1636, Sept. 2004.

[17] A. L. Moustakas and S. H. Simon, "Optimizing multiantenna systems with partial channel knowledge," in *Proc. Seventh Int. Symposium on Sig. Proc. and Its Applications*, vol. 1, (Hong Kong), pp. 217–220, July 2003.

[18] M. Skoglund and G. Jongren, "On the capacity of a multiple-antenna communication link with channel side information," *IEEE J. on Select. Areas in Commun.*, vol. 21, pp. 395–405, Apr. 2003.

[19] E. A. Jorsweick and H. Boche, "Optimal transmission strategies and impact of correlation in multiantenna systems with different types of channel state information," *IEEE Trans. on Sig. Proc.*, vol. 52, pp. 3440–3453, Dec. 2004.

[20] A. Narula, M. J. Lopez, M. D. Trott, and G. W. Wornell, "Efficient use of side information in multiple-antenna data transmission over fading channels," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1423 – 1436, Oct. 1998.

[21] S. Bhashyam, A. Sabharwal, and B. Aazhang, "Feedback gain in multiple antenna systems," *IEEE Trans. on Commun.*, vol. 50, pp. 785–798, May 2002.

[22] R. S. Blum, "MIMO with limited feedback of channel state information," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, vol. 4, pp. 89–92, Apr. 2003.

[23] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Trans. Info. Theory*, vol. 49, pp. 2562–2579, Oct. 2003.

[24] D. J. Love, R. W. Heath, Jr., and T. Strohmer, "Quantized maximal ratio transmission for multiple-input multiple-output wireless systems," in *Proc. Asilomar Conf. 2002*, (Pacific Grove, CA), Nov. 2002.

[25] D. J. Love, R. W. Heath, Jr., and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. Info. Theory*, vol. 49, pp. 2735–2747, Oct. 2003.

[26] W. Santipach and M. L. Honig, "Asymptotic performance of MIMO wireless channels with limited feedback," in *Proc. MILCOM*, (Boston, MA), pp. 141–146, Oct. 2003.

[27] J. C. Roh and B. D. Rao, "Performance analysis of multiple antenna systems with VQ-based feedback," in *Proc. Asilomar Conf. 2004*, (Pacific Grove, CA), Nov. 2004.

[28] K. N. Lau, Y. Liu, and T. A. Chen, "On the design of MIMO block-fading channels with feedback-link capacity constraint," *IEEE Trans. Commn.*, vol. 52, pp. 62–70, Jan. 2004.

[29] J. Choi and R. W. Heath, Jr., "Interpolation based transmit beamforming for MIMO-OFDM with limited feedback," in *Proc. IEEE ICC*, vol. 1, pp. 249 – 253, June 2004.

[30] P. Xia, S. Zhou, and G. B. Giannakis, "Multiantenna adaptive modulation with beamforming based on bandwidth- constrained feedback," *IEEE Trans. on Commn.*, vol. 53, pp. 526–536, Mar. 2005.

[31] P. Xia and G. B. Giannakis, "Design and analysis of transmit-beamforming based on limited-rate feedback," in *Proc. IEEE Vechicular Technol. Conference*, (Los Angeles, CA), Sept. 2004.

[32] G. Jongren and M. Skoglund, "Quantized feedback information in orthogonal space-time block coding," *IEEE Trans. Info. Theory*, vol. 50, pp. 2473–2486, Oct. 2004.

[33] D. J. Love and R. W. Heath, Jr., "Limited feedback unitary precoding for orthogonal space-time block codes," *IEEE Trans. on Sig. Proc.*, vol. 53, pp. 64–73, Jan. 2005.

[34] D. J. Love, R. W. Heath, Jr., W. Santipach, and M. L. Honig, "What is the value of limited feedback for mimo channels?," *IEEE Comm. Mag.*, vol. 42, pp. 54–59, Oct. 2004.

[35] J. C. Roh and B. D. Rao, "Channel feedback quantization methods for MISO and MIMO systems," in *IEEE International symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) 2004*, (Barcelona, Spain), Sept. 2004.

[36] P. Xia and G. B. Giannakis, "Design and analysis of transmit-beamforming based on limited-rate feedback," *IEEE Trans. on Sig. Proc.*, 2006 (to appear).

[37] R. W. Heath, Jr. and A. J. Paulraj, "A simple scheme for transmit diversity using partial channel feedback," in *Proc. Asilomar Conf. 1998*, vol. 2, (Pacific Grove, CA), Nov. 1998.

[38] B. M. Hochwald, T. Marzetta, T. Richardson, W. Sweldens, and R. Urbanke, "Systematic design of unitary space-time constellations," *IEEE Trans. Info. Theory*, vol. 46, pp. 1962 – 1973, Sept. 2000.

[39] P. Xia, S. Zhou, and G. B. Giannakis, "Achieving the welch bound with difference sets," *IEEE Trans. Info. Theory*, vol. 51, pp. 1900–1907, May 2005.

[40] C. R. Murthy and B. D. Rao, "On antenna selection with maximum ratio transmission," in *Conf. Record of the Thirty-Seventh Asilomar Conf. on Signals, Systems and Computers*, vol. 1, pp. 228 – 232, Nov. 2003.

[41] A. Medles, D. T. M. Slock, and E. de Carvalho, "Linear prediction based semi-blind estimation of MIMO FIR channels," *Third IEEE SPAWC, Taiwan*, pp. 58–61, Mar. 2001.

[42] A. K. Jagannatham and B. D. Rao, "A semi-blind technique for MIMO channel matrix estimation," in *Proc. SPAWC*, (Rome, Italy), pp. 304–308, June 2003.

[43] A. K. Jagannatham and B. D. Rao, "Whitening rotation based semi-blind MIMO channel estimation," *IEEE Trans. on Sig. Proc.*, Submitted.

[44] T. Marzetta, "BLAST training: Estimating channel characteristics for high-capacity space-time wireless," *Proc. 37th Annual Allerton Conference on Communications, Control, and Computing, Monticello, IL*, pp. 22–24, Sept. 1999.

[45] S. M. Kay, *Fundamentals of Statistical Signal Processing,Vol I: Estimation Theory*. Prentice Hall PTR, first ed., 1993.

[46] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, ch. 11. John Wiley & Sons, 1971.

[47] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*. Walton St., Oxford: Oxford University Press, first ed., 1965.

[48] M. Kaveh and A. J. Barabell, "The statistical performance of the MUSIC and the minimum-norm algorithms in resolving plane waves in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 2, pp. 331–341, 1986.

[49] J. G. Proakis, *Digital Communications*. NewYork,NY-10020: McGraw-Hill Higher Education, international ed., 2001.

[50] A. K. Jagannatham and B. D. Rao, "Complex constrained CRB and its application to semi-blind MIMO and OFDM channel estimation," in *Proc. of IEEE SAM Workshop*, (Sitges, Barcelona), 2004.

[51] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1451 – 1458, Oct. 1998.

[52] D. Gu and C. Leung, "Performance analysis of transmit diversity scheme with imperfect channel estimation," *IEE Electronics Letters*, vol. 39, pp. 402–403, Feb. 2003.

[53] T. Baykas and A. Yongacoglu, "Robustness of transmit diversity schemes with multiple receive antennas at imperfect channel state information," in *IEEE CCECE 2003*, vol. 1, pp. 191–194, may 2003.

[54] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic, 1992.

[55] J. C. Roh, *Multiple-Antenna Communication with Finite Rate Feedback*. PhD thesis, Univ. of California, San Diego, 2005.

[56] M. K. Simon and M. Alouini, *Digital Communication over Fading Channels*. Wiley-IEEE Press, 2nd ed., 2004.

[57] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*. Academic Press, Inc, 5th ed., 1994.

[58] M. Nakagami, "The m-distribution - A general formula of intensity distribution of rapid fading," *Statistical Methods in Radio Wave Propagation*, 1960.

[59] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. Journal*, vol. 27, pp. 446 – 472, July 1948.

[60] S. Na and D. Neuhoff, "Bennett's integral for vector quantizers," *IEEE Trans. Info. Theory*, vol. 41, pp. 886 – 900, July 1995.

[61] D. Astély, "On antenna arrays in mobile communication systems: Fast fading and GSM base station receiver algorithms," Master's thesis, Royal Inst. Technology, Sweden, March 1996.

[62] S. A. Jafar and S. Srinivasa, "Capacity of the isotropic fading vector channel with quantized channel direction feedback," in *Conf. Record of the 38th Asilomar Conf. on Signals, Systems and Computers*, (Pacific Grove, CA, USA), pp. 1178–1182, Nov. 2004.

[63] S. Srinivasa and S. A. Jafar, "Vector channel capacity with quantized feedback," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, (Seoul, S. Korea), May 2005.

[64] D. G. Luenberger, *Linear and Nonlinear Programming*. Addison-Wesley Publishing Company, 2nd ed., 1989.

[65] N. Farvardin and V. Vaishampayan, "Optimal quantizer design for noisy channels: An approach to combined source-channel coding," *IEEE Trans. Info. Theory*, vol. 33, pp. 155 – 159, Nov. 1987.

[66] N. Farvardin, "A study of vector quantization for noisy channels," *IEEE Trans. Info. Theory*, vol. 36, pp. 799 – 809, July 1990.

[67] A. Gersho and K. A. Zeger, "Vector quantizer design for memoryless noisy channels," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, (Philadelphia, PA, USA), pp. 1593–1597, June 1998.

[68] K. Zeger and V. Manzella, "Asymptotic bounds on optimal noisy channel quantization via random coding," *IEEE Trans. Info. Theory*, vol. 40, pp. 1926 – 1938, Nov. 1994.

[69] J. R. B. DeMarca and N. S. Jayant, "An algorithm for assigning binary indices to the codevectors of multidimensional quantizers," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, pp. 1128–1132, June 1987.

[70] K. Zeger and A. Gersho, "Pseudo-Gray coding," *IEEE Trans. Comm.*, vol. 38, pp. 2147 – 2156, Dec. 1990.

[71] D. J. Goodman and T. J. Moulsley, "Using simulated annealing to design transmission codes for analogue sources," *Electronics Letters*, vol. 24, pp. 617 – 618, May 1988.

[72] P. Knagenhjelm and E. Agrell, "The Hadamard transform - a tool for index assignment," *IEEE Trans. Info. Theory*, vol. 42, pp. 1139 – 1151, July 1996.

[73] S. W. McLaughlin and D. L. Neuhoff, "Asymptotic quantization for noisy channels," in *Proc. IEEE Interational Symposium on Information Theory*, pp. 442–442, Jan. 1993.

[74] G. Ben-David and D. Malah, "Bounds on the performance of vector-quantizers under channel errors," *IEEE Trans. Info. Theory*, vol. 51, pp. 2227 – 2235, June 2005.

[75] B. Hochwald and K. Zeger, "Tradeoff between source and channel coding," *IEEE Trans. Info. Theory*, vol. 43, pp. 1412 – 1424, Sept. 1997.

[76] A. Gersho, "Asymptotically optimum block quantization," *IEEE Trans. Info. Theory*, vol. IT-25, pp. 373–380, July 1979.

[77] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Info. Theory*, vol. IT-28, pp. 129–137, Mar. 1982.

[78] W. R. Gardner and B. D. Rao, "Theoretical analysis of high-rate vector quantization of LPC parameters," *IEEE Trans. on Speech and Audio Proc.*, vol. 3, pp. 367–381, Sept. 1995.

[79] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Willy, 1991.

[80] P. M. Pardalos, F. Rendl, and H.Wolkowicz, "The quadratic assignment problem: A survey and recent development," *DIMACS series in Discrete Mathematics and Theoretical Computer Science*, vol. 16, pp. 1–42, 1994.

[81] R. Totty and G. Clark, "Reconstruction error in waveform transmission," *IEEE Trans. Info. Theory*, vol. 13, pp. 336 – 338, Apr. 1967.

[82] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, 1st ed., 1969.

[83] A. Méhes and K. Zeger, "Randomly chosen index assignments are asymptotically bad for uniform sources," *IEEE Trans. Info. Theory*, vol. 45, pp. 788 – 794, Mar. 1999.

[84] H. Scheffé, "A useful convergence theorem for probability distributions," *Annals of Mathematical Statistics*, vol. 18, pp. 434–458, 1947.

[85] J. Li, N. Chaddha, and R. M. Gray, "Asymptotic performance of vector quantizers with a perceptual distortion measure," *IEEE Trans. Info. Theory*, vol. 45, pp. 1082–1091, May 1999.

[86] J. H. Conway and N. J. A. Sloane, "Voronoi regions of lattices, second moments of polytopes, and quantization," *IEEE Trans. Info. Theory*, pp. 211 – 226, Mar. 1982.

[87] C. R. Murthy, J. Zheng, and B. D. Rao, "Multiple antenna systems with finite rate feedback," in *Proc. MILCOM*, (Atlantic City, NJ), Oct. 2005.

[88] C. R. Murthy and B. D. Rao, "A vector quantization based approach for equal gain transmission," in *Proc. Globecom*, (St. Louis, MO), Nov. 2005.

[89] A. D. Subramaniam and B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, pp. 130 – 142, Mar. 2003.

[90] F. K. Soong and B.-H. Juang, "Optimal quantization of LSP parameters," in *Proc. Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, pp. 394–397, 1988.

[91] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.