THE ROLES OF MAJORIZATION AND GENERALIZED TRIANGULAR DECOMPOSITION IN COMMUNICATION AND SIGNAL PROCESSING

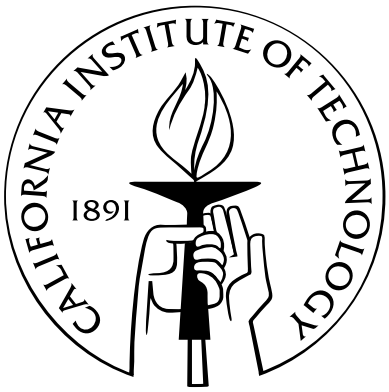
Thesis by

Ching-Chih Weng

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2011

(Defended June 2nd, 2011)

Acknowledgments

I still remember the day I first met my advisor, Professor P. P. Vaidyanathan, during an admission interview back in February of 2006. I was nervous, but hopeful at the same time, for the potential to study under such an esteemed scholar. When he told me that he would gladly take me under his wing, I was ecstatic—my parents thought I won a secret lottery of some sort because I was overfilled with joy. It is therefore now, five years later, that I express my most sincere gratitude to Prof. Vaidyanathan. He is a true gentleman, whose considerate guidance and careful nurturing led me to complete one of the most important milestones in my life. Without his advice and inspiration, my academic career would not have been the same. In every respect, he is the perfect teacher and a role model and I know all of his students will continue to learn from as we journey through our lives.

I would also like to thank members of my defense and candidacy examining committee: Professor Yaser Abu-Mostafa, Professor Babak Hassibi, Dr. Andre Tkacenko, and Dr. Kevin Quirk. Their knowledge and expertise have been instrumental to my study at Caltech. I studied information theory from Yaser, and stochastic signal processing from Babak; I learned communication theory with Kevin, and Andre's excellent papers on filter bank theory built a solid basis for my own academic research.

In regard to providing me with the financial resources to pursue this degree, I would like to thank the Office of Naval Research (ONR) and Taiwan's TMS scholarship from the National Science Council. Because of their generous support, I was able to join Caltech's excellent academic environment. This is a unique place on Earth because of all the researchers and scholars that have contributed their knowledge to better human lives, and continue to do so with uncompromising dedication. It has been an honor to be a part of their extraordinary community.

Speaking of scholars, my personal appreciations go to my current and former labmates, Professor Byung-Jun Yoon, Professor Borching Su, Dr. Chun-Yang Chen, Piya Pal, and Chih-Hao Liu. I

was very fortunate to work with these smart people, and enjoyed all the stimulating discussions and joyful moments that we shared. In particular, Borching and Chun-Yang are like two big brothers to me. Coming from a similar cultural background, their encouragement and friendship are treasures that I will continue to cherish.

My special thanks go to Christina Lin. She has brought me infinite happiness since our eyes met the very first time. I am so grateful to have her in my life. Also, Christina's parents, Alex and Irene Lin, have been more than supportive, treating me like their own son.

Being raised in a traditional Taiwanese family, I sometimes find it difficult to express love and affection verbally to my dear parents, Chang-Yi Weng and Li-Chuan Huang. However, it is only with their unconditional love that I grew to become the man that I am today. They have worked hard to provide me with my education and all the resources that I may have taken for granted. A simple "thanks" is not enough to show them my deepest appreciation, but I hope to make them proud. In addition, I would like to thank my brother, Pei-Chao, and sister, I-Han, for their support and for taking care of my parents when I am abroad. I am also grateful to my grandfather Shu-Gen Weng, and grandmother Tsui-Hsiu Weng Lu. I know that grandma has always watched me with a loving eye from above.

Lastly, I would like to thank all those who have helped me, knowingly or unknowingly, in this universe. There might be forces out there beyond my comprehension, but I am truly grateful for all the encounters and opportunities. Life may be short by the universe standard, but I will continue to take this positive energy and embark on new endeavors. Thank you!

Abstract

Signal processing is an art that deals with the representation, transformation, and manipulation of the signals and the information they contain based on their specific features. The field of signal processing has always benefited from the interaction between theory, applications, and technologies for implementing the systems. The development of signal processing theory, in particular, relies heavily on mathematical tools including analysis, probability theory, matrix theory, and many others.

Recently, the theory of majorization, which is an extremely useful tool for deriving inequalities, was introduced to the signal processing society in the context of MIMO communication system design. This also led many researchers to develop a fundamental matrix decomposition called generalized triangular decomposition (GTD), which was general enough to include many existing matrix orthogonal decompositions as special cases.

The main contribution of this thesis is toward the use of majorization and GTD to the theory and many applications of signal processing. In particular, the focus is on developing new signal processing methods based on these mathematical tools for digital communication, data compression, and filter bank design. We revisit some classical problems and show that the theories of majorization and GTD provide a general framework for solving these problems. For many important new problems not solved earlier, they also provide elegant solutions.

The first part of the thesis focuses on transceiver design for multiple-input multiple-output (MIMO) communications. The first problem considered is the joint optimization of transceivers with linear precoders, decision feedback equalizers (DFEs), and bit allocation schemes for frequency flat MIMO channels. We show that the generalized triangular decomposition offers an optimal family of solutions to this problem. This general framework incorporates many existing designs, such as the optimal linear transceiver, the ZF-VBLAST system, and the geometric mean decomposition (GMD) transceiver, as special cases. It also predicts many novel optimal solutions that

have not been observed before. We also discuss the use of each of these theoretical solutions under practical considerations. In addition to total power constraints, we also consider the transceiver optimization under individual power constraints and other linear constraints on the transmitting covariance matrix, which includes a more realistic individual power constraint on each antenna. We show the use of semidefinite programming (SDP), and the theory of majorization again provides a general framework for optimizing the linear transceivers as well as the DFE transceivers. The transceiver design for frequency selective MIMO channels is then considered. Block diagonal GMD (BD-GMD), which is a special instance of GTD with block diagonal structure in one of the semi-unitary matrices, is used to design transceivers that have many desirable properties in both performance and computation.

The second part of the thesis focuses on signal processing algorithms for data compressions and filter bank designs. We revisit the classical transform coding problem (for optimizing the theoretical coding gain in the high bit rate regime) from the view point of GTD and majorization theory. A general family of optimal transform coders is introduced based on GTD. This family includes the Karhunen-Loéve transform (KLT), and the prediction-based lower triangular transform (PLT) as special cases. The coding gain of the entire family, with optimal bit allocation, is maximized and equal to those of the KLT and the PLT. Other special cases of the GTD-TC are the GMD (geometric mean decomposition) and the BID (bidiagonal transform). The GMD in particular has the property that the optimum bit allocation is a uniform allocation. We also propose using dither quantization in the GMD transform coder. Under the uniform bit loading scheme, it is shown that the proposed dithered GMD transform coders perform significantly better than the original GMD coder in the low rate regime.

Another important signal processing problem, namely the filter bank optimization based on the knowledge of input signal statistics, is then considered. GTD and the theory of majorization are again used to give a new look to this classical problem. We propose GTD filter banks as subband coders for optimizing the theoretical coding gain. The orthonormal GTD filter bank and the biorthogonal GTD filter bank are discussed in detail. We show that in both cases there are two fundamental properties in the optimal solutions, namely, *total decorrelation* and *spectrum equalization*. The optimal solutions can be obtained by performing the frequency dependent GTD on the Cholesky factor of the input power spectrum density matrices. We also show that in both theory and numerical simulations, the optimal GTD subband coders have superior performance than op-

timal traditional subband coders. In addition, the uniform bit loading scheme can be used in the optimal biorthogonal GTD coders with no loss of optimality. This solves the granularity problem in the conventional optimum bit loading formula. The use of the GTD filter banks in frequency selective MIMO communication systems is also discussed. Finally, the connection between the GTD filter bank and the traditional filter bank is clearly indicated.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Signal processing is an art that deals with the representation, transformation, and manipulation of signals and the information they contain based on their specific features. It is a technology that spans an immense set of disciplines. The field of signal processing has always benefited from the interaction between theory, applications, and technologies for implementing the systems. The development of signal processing theory, in particular, relies heavily on mathematical tools including analysis, probability theory, matrix theory, and many others. The theory of majorization, which is an extremely useful tool for deriving inequalities, was recently introduced to signal processing society. This also led researchers to develop a fundamental matrix decomposition called generalized triangular decomposition (GTD), which was shown to be general enough to include many existing orthogonal matrix decompositions as special cases.

The present thesis is a contribution towards the use of these newly developed mathematical tools in several important signal processing problems. In particular, we focus on the signal processing for communications and filter bank designs. With these powerful mathematical tools at hand, we revisit some classical problems and show that the theories of majorization and GTD provide a general framework for solving these problems. For some important new problems, these new tools also provide elegant solutions.

In this introductory chapter, we review the history and development of the problems of focus in this thesis, namely, the MIMO transceiver optimization and signal-adapted filter bank optimization. Every attempt is to make the present text as self-contained as possible, and the introduction is meant to serve this purpose. Due to the large volume of the literature, the summary here is only directly related to the current thesis and is by no means a complete treatment of all past work. The readers interested in more comprehensive treatments are referred to [65, 75, 107], and especially

Chapter 9 of [111].

1.1 MIMO Transceiver Optimization

1.1.1 MIMO Channel Models

The first part of this thesis focuses on the communication system design, in particular, transceiver design for multiple-input multiple-output (MIMO) channels. We consider modeling the communication channels as linear time invariant (LTI) systems with additive noise. The focus is on MIMO channel models because they represent a unified way to model a wide variety of many different types of communication scenarios. Also, the matrix-vector notation can be conveniently used to handled MIMO channel models. This fact is crucial for applying elegant matrix theory results to many communication system design problems.

The models we considered are the MIMO frequency flat channel model, and the MIMO frequency selective channel model, which are shown in Fig. 1.1(a) and Fig. 1.1(b), respectively. Here $\mathbf{x}(n)$ is the $N_T \times 1$ transmitted signal, $\mathbf{y}(n)$ is the $N_R \times 1$ received signal, and $\mathbf{e}(n)$ is the $N_R \times 1$ additive noise introduced by the channel. The $N_R \times N_T$ channel matrix is modeled as $\mathbf{H}(e^{j\omega})$ for the frequency selective case. When the channel is memoryless, the channel is modeled as the constant matrix $\mathbf{H}$. In these linear models, the received signal is the channel noise plus the transmitted signal after being linearly distorted by the channel. For frequency flat channels, the input/output relation of the system is

$$\mathbf{y}(n) = \mathbf{H}\mathbf{x}(n) + \mathbf{e}(n), \forall \text{ integer } n.$$

For frequency selective channels, the received signal $\mathbf{y}(n)$ is the noise $\mathbf{e}(n)$ plus the output of the filter $\mathbf{H}(e^{j\omega})$ in response to $\mathbf{x}(n)$. The MIMO channel models are general enough to model many different communication scenarios, as we will see from several examples in the following.

1. *Channels of wireless multi-antenna systems:* The recent growth of wireless communication systems has drawn much attention to the problem of increasing system capacity. Since the radio spectrum available for wireless service is limited, spectral efficiency is therefore of primary concern to develop modern systems with bandwidth and power constraints. The multi-antenna system is a way to generate spatial diversity and greatly increase capacity in the

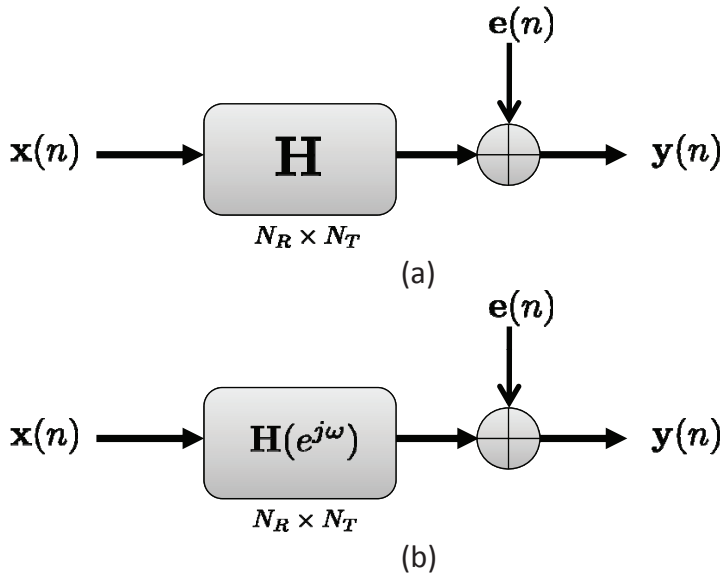


Figure 1.1: (a) Frequency flat MIMO channel model. (b) Frequency selective MIMO channel model.

wireless systems. If spatial diversity is simultaneously exploited at both the transmitter and the receiver, it is natural to use a MIMO representation. In this setting, N_R is the number of receiving antennas, and N_T is the number of transmitting antennas.

2. *Blocked scalar channels with finite memory:* The scalar channels with memory can be converted to MIMO channels without memory in a number of ways. Two of the most common techniques for this are the *zero-padding* and *cyclic-prefix* precoding techniques [89]. The zero-padding precoding produces the effective frequency flat channel matrix $\mathbf{H}$ having Toeplitz structure, and the cyclic-prefix precoding produces circulant matrix $\mathbf{H}$. The elements of $\mathbf{H}$ in each case can be obtained from the time domain samples of the channel coefficients. The cyclic-prefix precoding is in particular of great interests since it leads to OFDM and DMT systems, which have great performance advantages in wireless communication systems and digital subscriber line (DSL) systems.

In addition to the two examples mentioned above, there are many other common communication scenarios that can be modeled appropriately as MIMO channel models. These include multi-carrier systems on frequency selective channels, systems exploiting polarization diversity, and code division multiple access (CDMA) channels. Note that the structures of $\mathbf{H}$ and $\mathbf{H}(e^{j\omega})$ depend completely on the specific application at hand. In this thesis, we consider the generic transceiver opti-

mization problem with any given $\mathbf{H}$ or $\mathbf{H}(e^{j\omega})$

1.1.2 Transceiver Optimization and History

Transceiver optimization has had a long history since the 1950s. Because of the technological breakthrough of DSL, MIMO, and wireless communications, research in this area has become very intense since the 1990s. However, much of the recent work has its roots in the mathematical methods and signal models in earlier papers. This is the reason why we shall review the history of transceiver optimization in somewhat detailed fashion. The readers interested in more comprehensive treatments are referred to Chapter 9 in [111].

Fig. 1.2 shows the transceiver models we consider in this thesis. The channel $\mathbf{H}(e^{j\omega})$ (frequency selective in general) is assumed to be with dimension $N_R \times N_T$, and $\mathbf{e}(n)$ is the additive channel noise. Here $\mathbf{s}(n)$ is the $M \times 1$ symbols, $\mathbf{x}(n)$ is the transmitted signal, $\mathbf{y}(n)$ is the received signal, and $\hat{\mathbf{s}}(n)$ is the input to the scalar decision devices. Fig. 1.2(a) shows the linear transceiver with precoder $\mathbf{F}(e^{j\omega})$, and equalizer $\mathbf{G}(e^{j\omega})$. Fig. 1.2(b) shows the DFE transceiver with precoder $\mathbf{F}(e^{j\omega})$, feedforward filter $\mathbf{G}(e^{j\omega})$, and feedback filter $\mathbf{B}(e^{j\omega})$. Note that the successive decision feedback and decoding is performed at the receiver. The matrix $\mathbf{B}(e^{j\omega})$ is therefore restricted to be strictly upper triangular for casual implementation. In both systems, the transmitted signal $\mathbf{x}(n)$ is the symbol $\mathbf{s}(n)$ after linearly precoded by $\mathbf{F}(e^{j\omega})$, and the received signal $\mathbf{y}(n)$ is formed by channel noise $\mathbf{e}(n)$ plus the transmitted signal distorted by the channel $\mathbf{H}(e^{j\omega})$. Also note that the model in Fig. 1.2 is general enough to model the frequency flat case. If the channel is frequency flat, where the channel is a constant matrix $\mathbf{H}$, then all the other matrices in Fig. 1.2 are also constant matrices.

The transceiver optimization problem is to optimize $\{\mathbf{F}(e^{j\omega}), \mathbf{G}(e^{j\omega})\}$ for the linear transceiver, or $\{\mathbf{F}(e^{j\omega}), \mathbf{G}(e^{j\omega}), \mathbf{B}(e^{j\omega})\}$ for the DFE transceiver, subject to appropriate constraints under some assumptions of the channel state information available, such that some measure of performance is optimized. This simple model leads to a multitude of interesting optimization problems depending upon the applications. For example, one may wish to minimize the bit error rate under a total average power constraint or individual antenna power constraints. One may also consider the quality of service problem to minimize the transmitted power under some bit rate and bit error rate constraints for subchannels.

Before the MIMO model was introduced to the communication society, researchers had been attempting to optimize the single-input single-output (SISO) transceivers under different assump-

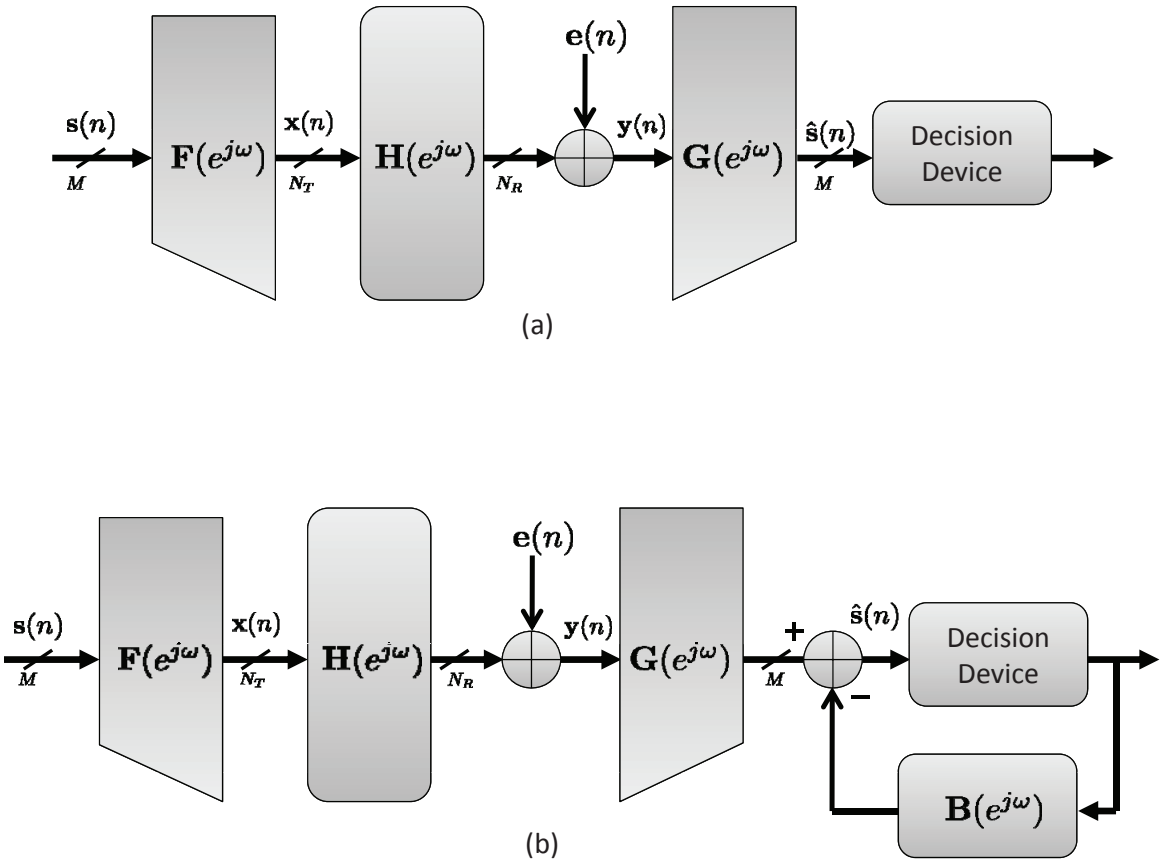


Figure 1.2: (a) The general form of linear transceivers with channel $\mathbf{H}(e^{j\omega})$, precoder $\mathbf{F}(e^{j\omega})$, and equalizer $\mathbf{G}(e^{j\omega})$. (b) The general form of DFE transceivers with channel $\mathbf{H}(e^{j\omega})$, precoder $\mathbf{F}(e^{j\omega})$, feedforward filter $\mathbf{G}(e^{j\omega})$, and feedback filter $\mathbf{B}(e^{j\omega})$. Note that the successive decision feedback and decoding is performed at the receiver. The matrix $\mathbf{B}(e^{j\omega})$ is restricted to be strictly upper triangular for casual implementation.

tions on the channel state information (CSI). CSI at the receiver (CSIR) is traditionally obtained via the transmission of pilot symbols that allows the estimation of the channel. CSI at the transmitter (CSIT), on the other hand, cannot be directly obtained as such. One scheme to obtain CSIT is to send the transmitter the quantized version of channel coefficients once the receiver estimates the channel. Another popular scheme is the so-called *limited feedback* technique, in which the receiver feeds back the index of the precoder in the predetermined codebook to inform the transmitter which precoding scheme to use. Although the CSIT may not be perfect in practice, it is still crucial to discuss theoretically how to jointly optimize the transmitter and receiver assuming perfect CSIT and CSIR are both at hand. This serves as a performance upper bound and gives insight to practical designs. At the same time, it is also important to consider the more robust designs, i.e., the situation where perfect CSIT and/or CSIR is not available.

The history of linear transceiver optimization (also see Chapter 9 of [111] for more detailed review) can be dated back to the paper by Costas [16] in 1952, in which for the identity continuous channel the author addressed the problem of optimizing *prefilter* and *postfilter* to minimize the mean square error of the reconstructed signal. It was shown that the optimal scheme is of the form that adopts some Wiener-type receiver and some power loading on the transmitted signal across frequency. Optimization of the transceivers for discrete symbol streams under channel distortions was then considered by Berger and Tufs [6]. The Wiener-type receiver was again shown to be optimal, and the transmission scheme is similar to the fashion of Costas but with some modifications. In 1971, Chan and Donaldson [11] extended the optimization to cover sampling and quantization as in digital communication systems. The design of filters for the case of bandlimited channels was later addressed by Chevillat and Ungerboeck [13].

Another line of research is on replacing the linear receiver by the decision feedback equalizer (DFE) to combat the inter-symbol interference (ISI). In general, DFEs are superior to linear equalizers—slight for good channels, moderate for channels with severe attenuation distortion, and substantial for channels with spectral nulls. Price [84] assumed a zero-forcing condition and found the joint optimum DFE receiver and linear transmitter. The joint MMSE transmitter and receiver were later obtained by Salz [87]. Several other authors of important papers on DFE transceiver optimizations include Falconer and Foschini [22], Messerschmitt [66], and Witsenhausen [118].

Optimization of transceivers for MIMO channels was considered as early as the 1970s [45]. The milestone paper by Salz in 1985 [88] considered the problem of optimizing continuous-time

square *MIMO filters* for the case of transmitting a discrete time sequence through a continuous time channel under the average power constraint. Salz showed that the optimal equalizer was identified to be a Wiener-type filter and the transmitter was obtained from the Karush-Kuhn-Tucker (KKT) condition for constrained optimization. By using a theorem based on the concept of Schur-convexity provided by Witsenhausen, Salz showed that the optimal solution can be obtained by diagonalizing the channel using singular value decomposition (SVD), and deriving the optimal filters for the diagonal channel. This diagonalization was later to be observed ubiquitously in various MIMO transceiver optimization problems. In 1988 Malvar and Staelin [63] addressed the transceiver optimization for rectangular channel and transceiver matrices. Instead of total power constraints, individual antenna power constraints are also considered in [45] and [63]. Besides linear transceivers, the MIMO DFE transceiver optimization was addressed by Yang and Roy in [138], in which the authors proposed the optimal system that minimizes the MSE simultaneously minimizes the geometric MSE.

In 2003 Palomar *et al.* [73] officially introduced the theory of majorization and Schur-convexity to the linear transceiver optimization, and thus unified many existing works in the literature. In this milestone paper, the authors proposed a formulation that covers a wide range of objective functions and constraints for MIMO transceiver optimization problems. The solutions can be divided into different classes depending upon whether the objective function is Schur-convex or Schur-concave. If the objective function is Schur-concave in the mean square errors of the subchannels, the diagonalizing structure is always optimal. If the objective is Schur-convex, the optimal solution diagonalizes the channel after a very specific rotation of the transmitted symbols. Many of the existing works were shown to be special cases within this framework. Since this paper, the theory of majorization and convex optimization became more heavily used in this field.

In 2005, Jiang *et al.* [36] considered the DFE transceiver optimization and proposed the geometric mean decomposition (GMD) to decompose the channel matrix as parallel subchannels with equal gains. It was shown that the GMD system is optimal within the class of zero-forcing DFE linear precoded transceivers. Jiang *et al.* further considered the MMSE version and proposed uniform channel decomposition (UCD) design [37]. The UCD system is particularly attractive since it could provide substantial gain compared to the GMD system when there is channel null. At the same time, it maximizes the diversity as well as the mutual information between the input and output of the channel. Independent results on the DFE transceiver optimization at the same time were also

published by Zhang *et al.* [142], and Xu *et al.* [137]. A unified framework for DFE transceivers under total average power constraints, which can be seen as a parallel counterpart of Palomar et al.'s 2003 linear transceiver paper, was also reported in [41], and independently in [90]. Based on the concept of multiplicative majorization, the framework covers a wide range of objective functions and the solutions are divided into different classes depending upon whether the objective function is Schur-convex and Schur-concave in the logarithms of the MSEs in the subchannels. The UCD scheme was shown to be optimal for the class of Schur-convex functions. For the class of Schur-concave functions, the optimal solution becomes the degenerate linear transceiver.

It is sometimes desirable to optimize the transceiver system with specific quality of service (QoS) for each of the subbands, which could be assigned to different users. Pandharipande and Dasgupta [18] extended the work of [50] and used majorization theory to established some results for digital multitone (DMT) systems. In 2004, Palomar *et al.* extended their 2003 milestone paper and considered the QoS problem [77] for linear transceivers. Jiang et al. on the other hand considered the QoS problem for DFE transceivers and proposed the so-called *tunable channel decomposition* (TCD) [39]. This later led them to the discovery of generalized triangular decomposition (GTD) [38].

Most of the results described above assume the channel, precoder, and equalizer to be constant matrices. For frequency selective channels, some of the early papers considered frequency dependent precoder and equalizers. The solution ended up being ideal unrealizable filters [88, 138]. These papers still have great value since they provide a theoretical upper bound and give insight to design practical systems. More recently some researchers considered the transceivers with finite memory and showed how to design these filters for practical applications. For example, Mertins [67] constrained the precoder to be a FIR matrix but the receiver can in general to be IIR. The work of Phoong *et al.* [81] restrained both the precoder and equalizer to be FIR filters. The paper by Vijaya Krishna and Hari [113] considered the optimization of minimum redundancy precoders, which was originally proposed by Lin and Phoong [80].

As mentioned previously, most of the results described above assume perfect channel state information at both transmitters and receivers (perfect CSIT and CSIR). More recently the robust designs, which focus on assuming perfect CSIR but imperfect CSIT, also attracted attention. Limited feedback is the most popular way to obtain approximate CSIT. An excellent review on this topic up to 2008 can be found in [55]. In particular, in the series papers by Love and Heath [56, 57], *Grassmannian codebook* was found to optimize some performance upper bounds of linear

transceivers for many different communication scenarios. This framework was later extended to the DFE transceivers in [91], and Grassmannian codebook was again found to be useful.

Such is the very brief history of transceiver optimization. Researchers across many decades have devoted themselves to this field. As mentioned above, the 2003 paper by Palomar *et al.* officially brought the theory of majorization and Schur-convexity into this field and changed the way people view these problems. By considering the DFE transceiver counter part, Jiang *et al.* discovered a novel matrix decomposition—generalized triangular decomposition (GTD) that was shown to be useful in the MIMO transceiver QoS problem. This thesis continues this line, and further shows that the theory of majorization and GTD is useful not only in the scenarios described in these earlier papers, but also in broader signal processing applications including other important transceiver optimization problems, transform coding problems, and filter bank optimization.

1.2 Transform Coder and Signal Adapted Filter Bank Optimization

A filter bank (FB) is used to decompose a signal into several bands, which are then processed independently and combined. Processing resources can be allocated according to specific features in each of the subbands. The optimization of filter banks based on knowledge of input statistics has been of great interest in signal processing for a long time. The transform coder optimization, before the time of the filter bank, was first considered by Huang and Schultheiss [33] in the 1960s. Since then, there have been many advances in the theory of filter banks, wavelets, and their ubiquitous signal processing applications including data compressions, signal denoising, and digital communications.

Fig. 1.3 shows the standard M-channel filter banks which can be found in many signal processing books, e.g., [107]. The subband processors P_i can represent many kinds of linear or nonlinear operations, such as a hard threshold device, a linear multiplier, and a quantizer. This structure is said to be a *uniform filter bank* because all the decimators are identical. The uniform filter bank is the focus of this thesis. Using the polyphase notation introduced in [107], we can redraw the uniform filter bank structure in the form of Fig. 1.4. The system is said to be *biorthogonal* if the filters are

such that the polyphase matrices $\mathbf{R}(e^{j\omega})$ and $\mathbf{E}(e^{j\omega})$ satisfy

$$\mathbf{R}(e^{j\omega})\mathbf{E}(e^{j\omega}) = \mathbf{I}$$

for all ω. This is also called the perfect reconstruction (PR) property. The reason is that in the absence of any subband processing, the PR property implies $\hat{x}(n) = x(n)$ for all n. For the special case where the polyphase matrix $\mathbf{E}(e^{j\omega})$ is paraunitary (i.e., unitary for all ω) and $\mathbf{R}(e^{j\omega}) = \mathbf{E}^{\dagger}(e^{j\omega})$, the filter bank is called an *orthonormal filter bank*. In this case, the set of M filters $\{H_k(e^{j\omega})\}$ is said to be orthonormal, and the set of synthesis filters can be shown to be $F_k(e^{j\omega}) = H_k^*(e^{j\omega})$.

A filter bank whose filters depend on knowledge of the input statistics is called a signal-adapted filter bank. The subband coding problem, which is an instance of the signal-adapted filter bank design problems, is to quantize the subband signals rather than to quantize the original signal directly, based on the knowledge of input statistics. While other performance measures in the rate-distortion sense are possible, one measure that receives much attention is the *theoretical coding gain*. The coding gain maximization problem is equivalent to the minimization of the average mean-square error of the reconstructed quantized signal by designing the filter bank and the bit allocation scheme.

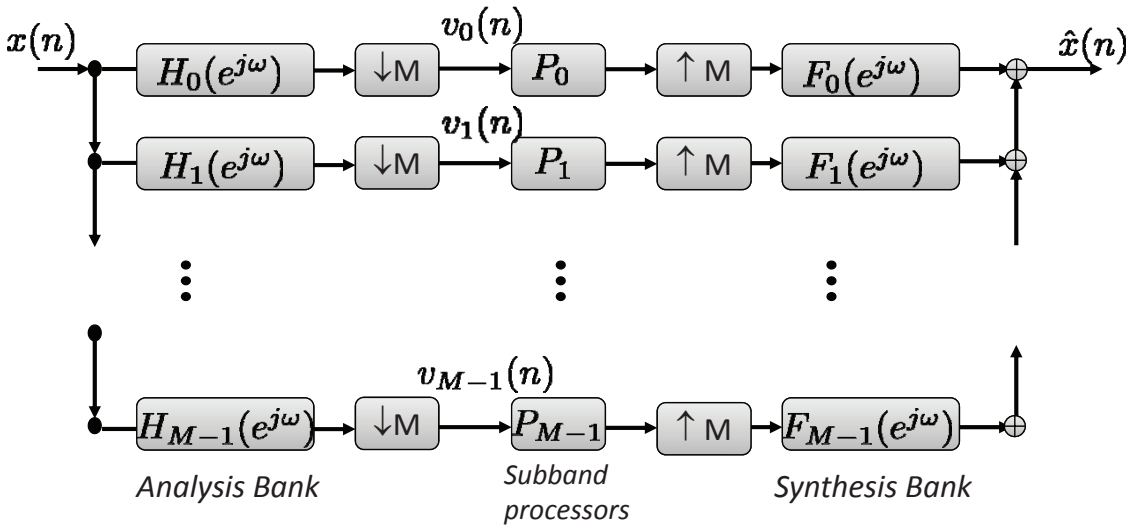


Figure 1.3: The M-channel maximally decimated filter bank with uniform decimation ratio M.

In the literature, the usual assumption is that the input signal $x(n)$ is a cyclo wide sense stationary process with period M (abbreviated CWSS(M)). It should be noted that $x(n)$ is CWSS(M)

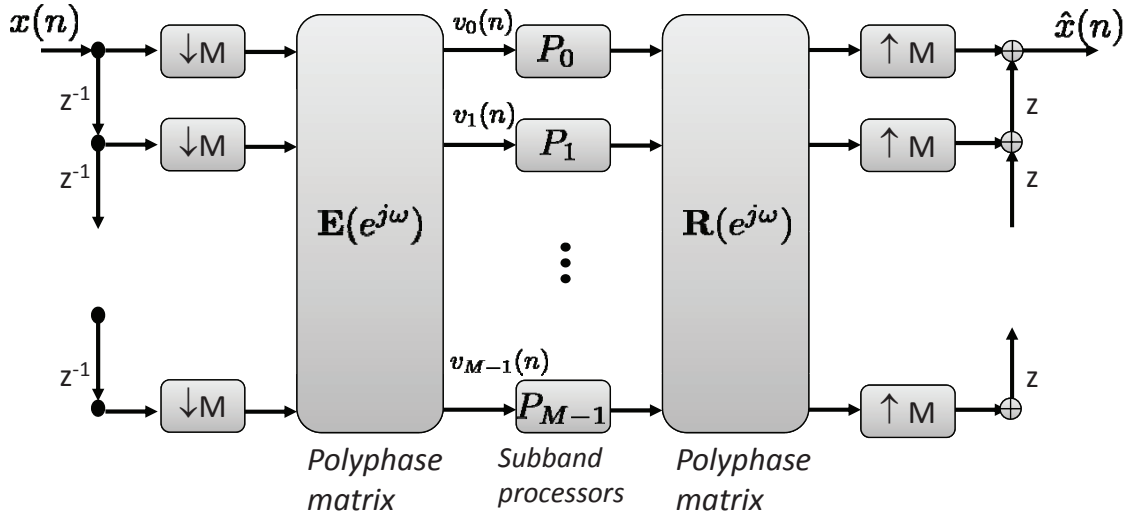


Figure 1.4: The polyphase representation of the M-channel maximally decimated filter bank.

if and only if the M-fold blocked version $\mathbf{x}(n)$ is wide sense stationary (WSS), meaning that the mean and autocorrelation of $\mathbf{x}(n)$ do not depend on n. In this thesis, we will adopt this assumption and also assume that $x(n)$ and hence $\mathbf{x}(n)$ are zero mean. In addition, we will assume that we only have knowledge of the second order statistics of $\mathbf{x}(n)$, namely the autocorrelation function $\mathbf{R_{xx}}(k) = E[\mathbf{x}(n)\mathbf{x}^\dagger(n-k)]$. The power spectral density (psd) matrix $\mathbf{S_{xx}}(e^{j\omega})$, which is simply the Fourier transform of $\mathbf{R_{xx}}(k)$, will often appear in the discussion.

We now give a brief overview of the past work in this field. For the special case where the matrices $\mathbf{E}(e^{j\omega})$ and $\mathbf{R}(e^{j\omega})$ are constant, the system in Fig. 1.4 is said to be a *transform coder*. In 1963, Huang and Schulthess [33] formulated the transform coding problem, and showed that the Karhunen-Loéve transform (KLT) with some bit loading formula maximizes the coding gain. The KLT essentially performs eigenvalue decomposition (EVD) of the input covariance matrix, and produces uncorrelated subband signals. In 2000, Phoong and Lin [79] revisited this problem and proposed the prediction-based lower triangular transform (PLT) that gives the same maximized coding gain. The PLT coder performs Cholesky decomposition of the input covariance matrix. Using linear prediction, the PLT coder produces uncorrelated subband signals and only performs the quantization on the *innovations* after subtracting linearly predicted signals. For the unconstrained filter order case, the research started in the late 1980s. Theoretical results on the optimization of a two channel orthonormal filter bank were developed by Unser in [103]. A closely related idea called

the *principal component filter banks* (PCFB) was introduced and developed in [100]. Theoretical results for the optimality of an orthonormal filter bank with unconstrained filter order was developed by Vaidyanathan in [106]. It was shown that there are two necessary and sufficient conditions for the optimal orthonormal subband coder, namely, *total decorrelation* and *spectrum majorization*. For the case of biorthogonal filter banks, it was conjectured by Vaidyanathan and Kirac that the optimal structure is the cascade of the optimal orthonormal filter bank, and a set of half-whitening filters applied to the signal in each individual subband [109]. This conjecture was later proven to be true by Moulin *et al.* in [70]. The authors in [70] also showed the two fundamental properties, total decorrelation and spectrum majorization, are also two necessary conditions for optimality of biorthogonal filter banks. The same group of researchers also extended their work and derived the optimal subband coders when there is no perfect reconstruction constraint [68].

These theoretical results provide a nice performance bound on the filter banks with unconstrained filter order. However, finite length filter implementation methods are needed for practice. The finite impulse response (FIR) solutions to the orthonormal and biorthogonal filter banks are also discussed extensively in the literature [44, 58, 69, 70, 97, 98, 102].

As mentioned earlier, principal component filter bank (PCFB) is a closely related concept. PCFB was shown to be simultaneously optimal for a variety of objective functions within the class of orthonormal filter banks. By definition, a PCFB for an input psd $\mathbf{S_{xx}}(e^{j\omega})$ and for a class C of filter banks, if it exists, is one whose subband variance vector

$$\boldsymbol{\sigma} = [\sigma_{v_0}^2 \ \sigma_{v_1}^2 \ \cdots \ \sigma_{v_{M-1}}^2]^T$$

additively majorizes any other subband variance vector arising from any other filter bank in C. In addition to being optimal for coding gain and mean-squared error in the presence of quantization noise, the PCFB has also been shown to be optimal for any concave function of $\boldsymbol{\sigma}$ [3]. This fact was later again explained by Jahromi *et al.* in the language of majorization and Schur-convexity [34]. Although PCFB has optimal characteristics, they are only known to exist for some special cases of filter banks. It is known that for $M = 2$, PCFB always exists for any class of orthonormal filter banks. For general M, however, the existence of PCFB is no longer guaranteed. If C is the class of all transform coders, the PCFB is the KLT coder. If C is the class of filters with unconstrained order, the PCFB is the optimal orthonormal subband coder that performs the frequency dependent KLT for $x(n)$. Although as such these filters are in general unrealizable, they serve to compute the

upper bound on the performance we can expect from paraunitary filter banks.

Such is the very brief review of the filter bank optimization problem. The research in this field, since as early as Huang and Schultheiss' transform coding paper, has been continued for almost five decades. In this thesis, we introduce the generalized triangular decomposition (GTD), which was originally developed in the MIMO communication society, to this field. We will show that the concept of GTD and the theory of multiplicative majorization give this classical problem a completely new look. Many novel coder structures are proposed, and several theoretical results are established. The connection of the GTD filter bank to the PCFB will also be indicated.

1.3 Outline and Scope of the Thesis

There are two major signal processing problems considered in detail in this thesis. The first problem of focus is transceiver designs for MIMO communications (Chapter 3, 4). We consider many different MIMO communication scenarios, including the optimization of DFE transceivers when bit allocation is allowed, QoS problem for DFE transceivers, transceiver design under individual antenna power constraints, and also the transceiver design for frequency selective channels. Based on the concept of majorization and the use of generalized triangular decomposition, many novel designs are proposed, and performance analyses are provided. The second problem of focus is the data compression and signal-adapted filter bank optimization (Chapter 5, 6). We first revisit the classical transform coding problem. Then, based on GTD, we propose and optimize a new subband coding structure, which can be shown to have superior performance than the existing ones. Many theoretical results as well as practical concerns will be presented. The connection between the current thesis and existing work in literature will also be clearly indicated. In the following we briefly discuss the scope of each chapter.

1.3.1 Review of Majorization, Matrix Theory, and Generalized Triangular Decomposition (Chapter 2)

In Chapter 2 we review the concepts of majorization, matrix theory, and generalized triangular decomposition, on which many results of this thesis are based. We start by introducing the intuition and mathematical formulation of additive majorization. The closely related concept of Schur-convexity and Schur-concavity is then introduced. Besides additive majorization, multiplicative

majorization is introduced as well.

Then, we review the connections between majorization and the matrix theory. Additive majorization is closely related to the properties of diagonal elements and eigenvalues of Hermitian matrices. On the other hand, multiplicative majorization is the complete characterization of the relation between singular values and eigenvalues of complex-valued square matrices. We then introduce the generalized triangular decomposition (GTD) developed by Jiang *et al.*. This decomposition is very fundamental and incorporates many existing well-known matrix orthogonal decompositions as special cases. Several examples of the GTD will be discussed in this chapter as well. Finally, we will review some mathematical properties of the block diagonal geometric mean decomposition (BD-GMD).

1.3.2 Transceiver Designs for MIMO Frequency Flat Channels (Chapter 3)

In Chapter 3, we consider several transceiver design problems for frequency flat MIMO channels using GTD and majorization theory. Mainly two scenarios are considered in detail.

The first part of this chapter considers the joint optimization of MIMO transceivers with linear precoders, decision feedback equalizers (DFEs), and bit allocation schemes. It is shown that the generalized triangular decomposition (GTD) offers an optimal family of solutions. The optimal linear transceiver (which has a linear equalizer rather than a DFE) with optimal bit allocation, as well as the DFE transceiver using the geometric mean decomposition (GMD), are members of this family. The QR-based system used in the VBLAST system is yet another member of the optimal family and is particularly well suited when limited feedback is allowed from receiver to transmitter. Thus, GTD provides a general theoretical framework for this optimization problem, and gives insight on the practical designs.

While most of the literature of transceiver designs focus on the total power constraints for wireless communications, the second part of this chapter considers the joint transceiver optimization problem for MIMO channels under more realistic individual antenna power constraints. The linear transceiver as well as the transceiver with linear precoding and MMSE-DFE are considered. For both types of transceivers, we show that the optimization problems for a wide range of objective functions can be formulated as semi-definite programming (SDP) optimization problems. Based on the result of majorization, after solving the SDPs, specific rotation is then performed to obtain the optimal solution. For both types of transceivers, the framework developed here is general enough

to also incorporate any finite number of linear constraints to the covariance matrix of the input.

1.3.3 Transceiver Designs for MIMO Frequency Selective Channels (Chapter 4)

While the previous chapter focuses on frequency flat MIMO channels, this chapter is devoted to transceiver designs for MIMO frequency selective channels. We consider using the block-diagonal GMD (BD-GMD), a novel matrix decomposition that was originally proposed by Lin *et al.* in 2008 [47] for MIMO broadcast channel, to design the DFE transceivers. Two new BD-GMD transceivers are proposed: the ZF-BD-GMD system, where the receiver is a zero-forcing DFE (ZF-DFE), and the MMSE-BD-GMD system, where the receiver is a minimum-mean-square-error DFE (MMSE-DFE). We show that the BD-GMD systems have many optimal properties and at the same time computationally efficient. These make the proposed BD-GMD favorable designs for frequency selective MIMO channels.

1.3.4 The Role of GTD in Transform Coding (Chapter 5)

In this chapter we revisit the classical transform coding problem from the view point of GTD and majorization theory. In the first part of this chapter, a general family of optimal transform coders (TC) is introduced based on GTD. This family includes the Karhunen-Loéve transform (KLT), and the generalized version of the prediction-based lower triangular transform (PLT) introduced by Phoong and Lin in 2000 [79], as special cases. The coding gain of the entire family, with optimal bit allocation, is equal to those of the KLT and the PLT. Other special cases of the GTD-TC are the GMD (geometric mean decomposition) and the BID (bidiagonal transform). The GMD in particular has the property that the optimum bit allocation is a uniform allocation; this is because all its transform domain coefficients have the same variance, implying thereby that the dynamic ranges of the coefficients to be quantized are identical.

The theoretical results established in the first part of this chapter are based on the high bit rate assumption. However, the performance of the GMD transform coder is degraded significantly in the low rate case. In the second part of this chapter, we introduce dither quantization to tackle this problem. The precoders and predictors in the GMD transform coders are redesigned accordingly. Two modified transform coders are proposed: the GMD subtractive dithered (GMD-SD) transform coder where the decoder has access to the dither information, and the GMD non-subtractive

dithered (GMD-NSD) transform coder where the decoder has no knowledge about the dither. Under the uniform bit loading scheme, it is shown that the proposed dithered GMD transform coders perform significantly better than the original GMD coder in the low rate regime.

1.3.5 The Role of GTD in Filter Bank Optimization (Chapter 6)

In this chapter we consider the filter bank optimization based on the knowledge of input signal statistics. GTD and the theory of majorization is again used to give a new look to this classical problem. We propose the GTD filter bank as a subband coder for optimizing the theoretical coding gain. The focus is on perfect reconstruction orthonormal GTD filter banks and biorthognal GTD filter banks. In both cases, we show that there are two fundamental properties in the optimal solutions, namely, *total decorrelation* and *spectrum equalization*. The optimal solutions can be obtained by performing the frequency dependent GTD on the Cholesky factor of the input power spectrum density matrices. We also show that in both theory and numerical simulations, the optimal GTD subband coders have superior performance to optimal traditional subband coders. In addition, the uniform bit loading scheme, with no loss of optimality, can be used in the optimal biorthogonal GTD coders. This solves the granularity problem in the conventional optimum bit loading formula.

The proposed GTD filter banks can also be used in MIMO communication systems. We consider the transceiver with linear precoding and zero-forcing decision feedback equalization for MMO frequency selective channels. The quality of service (QoS) problem of minimizing the transmitted power subject to the bit error rate and total bit rate constraints is considered. Optimal systems with orthonormal precoder and unconstrained precoder are both derived and shown to be related to the frequency dependent GTD of the channel frequency response.

1.4 Notations

The notations used throughout this thesis are defined as follows. Boldfaced lower case letters represent column vectors. Boldfaced upper case letters and calligraphic upper case letters are reserved for matrices. Superscripts *, and ^{T}, as in a^*, $\mathbf{A}^T$ denote the conjugate and the transpose, respectively. Superscripts $\mathbf{A}^\dagger$ and $\mathbf{A}^H$ both denote transpose-conjugate operations. $\mathbf{A}^\#$ represents the pseudo-inverse of $\mathbf{A}$. $[\mathbf{v}]_i$ denotes the ith element of vector $\mathbf{v}$, $[\mathbf{A}]_i$ denotes the ith row of matrix $\mathbf{A}$, and $[\mathbf{A}]_{ij}$ denotes the entry at the ith row and the jth column of matrix $\mathbf{A}$. For vector $\mathbf{x}$,

the notation diag($\mathbf{x}$) denotes the diagonal matrix with diagonal terms equal to the elements in $\mathbf{x}$. In figures, "$\uparrow N$" and "$\downarrow N$" denote the signal upsampler and downsampler, respectively [107]. For any $\mathbf{x} \in \mathbf{R}^n$, $x_{[1]} \geq x_{[2]} \geq \cdots \geq x_{[n]}$ denote the elements of $\mathbf{x}$ in descending order, and $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ denote the elements of $\mathbf{x}$ in ascending order. For two real vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x} \succ_+ \mathbf{y}$ or $\mathbf{y} \prec_+ \mathbf{x}$ denotes that $\mathbf{x}$ additively majorizes $\mathbf{y}$ [65]. For two complex vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x} \succ_\times \mathbf{y}$ or $\mathbf{y} \prec_\times \mathbf{x}$ denotes that $\mathbf{x}$ multiplicatively majorizes $\mathbf{y}$ [75].

Chapter 2

Review of Majorization, Matrix Theory, and Generalized Triangular Decomposition

In this chapter we will give a brief overview of the necessary mathematical preliminaries, on which many of the results in this thesis are based. We will first introduce majorization theory, Schur-convexity, and the relation to matrix theory. Then, we will review the generalized triangular decomposition (GTD). A special case of GTD, namely the block-diagonal geometric mean decomposition (BD-GMD), will also be introduced.

2.1 Review of Majorization and Schur Convexity

The idea of majorization and Schur-convexity is very fundamental to many problems in linear algebra and optimization. One of the earliest references on this topic is the book by Hardy, Littlewood, and Pólya [29]. More recent references include Marshall and Olkin [65]. The relation between majorization theory and matrix theory is discussed extensively in [32]. Many of the results reviewed in this chapter are used in the thesis. The readers interested in more comprehensive treatments are also referred to [111].

2.1.1 Additive Majorization and Schur Convexity

The idea and motivation of introducing majorization might be best explained by the following sentence in Marshall and Olkin [65]: *There is a certain intuitive appeal to the vague notion that the components of a vector x are "less spread out" or "more nearly equal" than are the components of a vector y.*

The notion of majorization makes this precise.

Definition 2.1: (*Additive Majorization*.) For any $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$, the vector $\mathbf{x}$ is said to additively majorize $\mathbf{y}$ (or $\mathbf{y}$ is additively majorized by $\mathbf{x}$), and is denoted as $\mathbf{x} \succ_+ \mathbf{y}$ if and only if[1]

$$\sum_{i=1}^{k} \mathbf{x}_{[i]} \geq \sum_{i=1}^{k} \mathbf{y}_{[i]}, \ \forall k = 1, 2, \cdots, n-1$$

and

$$\sum_{i=1}^{n} \mathbf{x}_{[i]} = \sum_{i=1}^{n} \mathbf{y}_{[i]}$$

□

From the above definition it can be seen that the notion of majorization is invariant to any permutation of the elements in the vector, i.e.,

$$\mathbf{x} \succ_+ \mathbf{y} \text{ if and only if } \mathbf{\Pi}_1 \mathbf{x} \succ_+ \mathbf{\Pi}_2 \mathbf{y}$$

for any permutation matrix $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$. The ordering "$\succ_+$" defined on $\mathbf{R}^n$ is a *preordering* but not *partial ordering* (see p.13 of [65]). However it is also *partial proper ordering* if it is regarded as an ordering of sets of numbers rather than as an ordering of vectors. It it important to remember that two sequences may not have any majorization relationship.

One simple observation is that the vector of the arithmetic mean of the elements is always additively majorized, which is shown in the following example.

Example 2.1: Let $\mathbf{x} \in \mathbf{R}^n$, and $\overline{\mathbf{x}}$ denote the constant vector with the value equal to the arithmetic mean of the elements in $\mathbf{x}$, i.e., $\frac{1}{n}\sum_{i=1}^{n} x_i$. Then, $\mathbf{x} \succ_+ \overline{\mathbf{x}}$. ♢

The notion of majorization is closely related to Schur-convexity. Specifically, Schur-convexity characterizes the differentiable functions that preserve the ordering $\succ_+$ on $\mathbf{R}^n$ (see p.53 of [65]).

Definition 2.2: (*Schur-Convexity*.) A real-valued function ϕ defined on a set $A \subseteq \mathbf{R}^n$ is said to be Schur-convex on A if

$$\mathbf{x} \prec_+ \mathbf{y} \text{ on } A \ \Rightarrow \phi(\mathbf{x}) \leq \phi(\mathbf{y}).$$

[1]The notation $x_{[i]}$ is defined in Sec. 1.4.

In addition, ϕ is said to be Schur-concave if and only if $-\phi$ is Schur-convex. □

Note that the sets of Schur-concave and Schur-convex functions do not form a partition of the set of all functions. This fact will be illustrated later in Fig. 2.1.

There are many beautiful examples of Schur-convex/concave functions that arise in optimization problems in signal processing and communications. Some of the examples follow from the theorems presented in this section. The first theorem shows the relation between convex functions and Schur-convex functions [29, 65].

Theorem 2.1.1 *The inequality*

$$\sum_{i=1}^{n} g(x_i) \geq \sum_{i=1}^{n} g(y_i)$$

holds for all continuous convex functions $g : \mathbf{R} \to \mathbf{R}$ *if and only if* $\mathbf{x} \succ_+ \mathbf{y}$. *Therefore, the function* $f(\mathbf{x}) = \sum_{i=1}^{n} g(x_i)$ *is a Schur-convex function in* $\mathbf{x}$. ♢

Example 2.2: (*Log-Product*.) Let $g(x) = \log(x)$ for any positive x and thus g is convex. The function

$$f(\mathbf{x}) = \sum_{i=1}^{N} g(x) = \sum_{i=1}^{N} \log(x) = \log\left(\prod_{i=1}^{N} x_i\right)$$

is Schur-convex. ♢

Example 2.3: (*The average probability of error of MIMO communication systems*.) The average symbol error probability of a scalar Gaussian communication channel is given by (see p.383 - p.385 of [111])

$$P_e(y) = cQ(A/\sqrt{y}),$$

where c and A are constants depending on the constellations (the size of PAM or QAM signaling) used and the signal power, y is the error variance, and $Q(\cdot)$ is the Q-function. It was shown that $Q(A/\sqrt{y})$ is convex in y for $y < A^2/3$. Therefore, the average symbol error probability of a M-channel MIMO communication system is given by

$$P_e(\mathbf{y}) = \frac{c}{M} \sum_{k=1}^{M} Q(\frac{A}{\sqrt{y_k}}).$$

Using Thm. 2.1.1, it can be shown that $P_e(\mathbf{y})$ is Schur-convex if $y_k < A^2/3$ for all k. $\diamondsuit$

There are a number of simple but useful facts relating to compositions that involve Schur-convex and Schur-concave functions (p.61 of [65]). If $g(\mathbf{x})$ is Schur-convex then $f(\mathbf{x}) = h(g(\mathbf{x}))$ is Schur-convex, as long as $h(\cdot)$ is a non-decreasing real function of its argument. More relations are shown as follows:

$$g(\mathbf{x}) \text{ is Schur-convex and } h(y) \text{ is non-decreasing} \Rightarrow f(\mathbf{x}) \text{ is Schur-convex;}$$

$$g(\mathbf{x}) \text{ is Schur-convex and } h(y) \text{ is non-increasing} \Rightarrow f(\mathbf{x}) \text{ is Schur-concave;}$$

$$g(\mathbf{x}) \text{ is Schur-concave and } h(y) \text{ is non-decreasing} \Rightarrow f(\mathbf{x}) \text{ is Schur-concave;}$$

$$g(\mathbf{x}) \text{ is Schur-concave and } h(y) \text{ is non-increasing} \Rightarrow f(\mathbf{x}) \text{ is Schur-convex.}$$

2.1.2 Multiplicative Majorization

The notion parallel to additive majorization is *multiplicative majorization*, which finds application in various signal processing problems [133, 28, 38].

Definition 2.3: : (*Multiplicative Majorization* [65, 38].) For any $\mathbf{x}, \mathbf{y} \in \mathbf{R}_+^n$, the vector $\mathbf{x}$ is said to multiplicatively majorize $\mathbf{y}$ (denoted as $\mathbf{x} \succ_\times \mathbf{y}$) if and only if

$$\prod_{i=1}^{k} \mathbf{x}_{[i]} \geq \prod_{i=1}^{k} \mathbf{y}_{[i]}, \ \forall k = 1, 2, \cdots, n-1$$

and

$$\prod_{i=1}^{n} \mathbf{x}_{[i]} = \prod_{i=1}^{n} \mathbf{y}_{[i]}.$$

$\square$

From these two definitions, we observe that if all elements of $\mathbf{x}$ and $\mathbf{y}$ are positive, then

$$\mathbf{x} \succ_\times \mathbf{y} \Longleftrightarrow \ln(\mathbf{x}) \succ_+ \ln(\mathbf{y})$$

In the following we give few examples of multiplicative majorization.

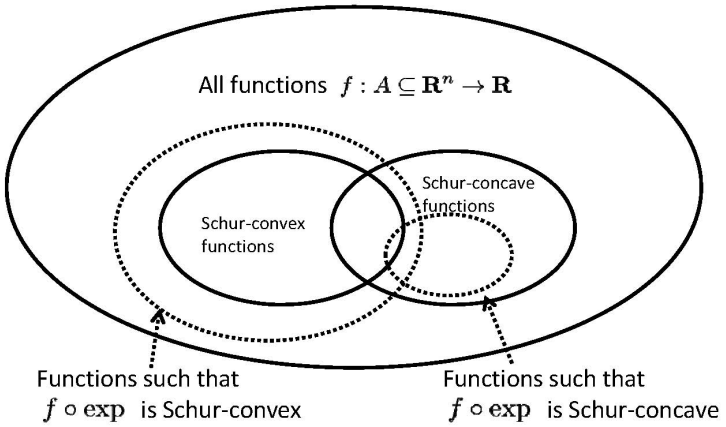


Figure 2.1: The illustration of the relations between sets of functions [41].

Example 2.3: Let $\mathbf{x} \in \mathbf{R}_+^n$ and $\hat{\mathbf{x}}$ denotes the constant vector with the value equals to the geometric mean of the elements in $\mathbf{x}$, i.e. $\sqrt[n]{\prod_{i=1}^n x_i}$. Then

$$\mathbf{x} \succ_\times \hat{\mathbf{x}}.$$

$\diamondsuit$

A type of function closely related to the notion of multiplicative majorization is the composition of Schur-convex functions and the exponential function. The composition $f \circ \exp : \mathbf{R}^N \to \mathbf{R}$ is defined as

$$f \circ \exp(\mathbf{x}) \equiv f(e^{x_1}, e^{x_2}, \cdots, e^{x_N}).$$

The following theorem is a direct consequence of the composition rule of Schur-convex functions with increasing convex functions.

Theorem 2.1.2 *(a) The composite function $f \circ exp$ is Schur-convex if function $f : \mathbf{R} \subseteq \mathbf{R}^n$ is Schur-convex.*

(b) If the composite function $f \circ exp$ is Schur-concave, then $f : \mathbf{R} \subseteq \mathbf{R}^n$ is Schur-concave. $\diamondsuit$

Theorem 2.1.2 is very useful in optimizing the MIMO transceiver with the linear precoder and decision feedback equalizer [41]. Fig. 2.1, which was first presented in [41], illustrates the relation between functions that map arguments from $\mathbf{R}^n$ to $\mathbf{R}$.

2.2 Relation to Matrix Theory

There are several important facts connecting matrix theory and the notion of majorization. These beautiful results in the matrix singular values and eigenvalues of a square matrix, or eigenvalues and diagonal terms of a Hermitian matrix, serve as the foundation of the results established in this thesis.

2.2.1 Hermitian Matrices

The first result we present here was proved by Schur in 1923. It relates the diagonal elements of a Hermitian matrix to its eigenvalues [65].

Theorem 2.2.1 *(Diagonal elements and eigenvalues of a Hermitian matrix [31, 65].) Let* $\mathbf{R}$ *be an* $n \times n$ *Hermitian matrix with diagonal elements denoted by the vector* $\mathbf{d}$ *and eigenvalues denoted by* $\boldsymbol{\lambda}$*. Then*

$$\boldsymbol{\lambda} \succ_+ \mathbf{d}. \tag{2.1}$$

That is, for a Hermitian matrix, the vector of eigenvalues majorizes the vector of diagonal elements. $\diamondsuit$

For Theorem 2.2.1, the converse is also true. Therefore, the notion of additive majorization is the strongest relation we can have between the diagonal elements and the eigenvalues of Hermitian matrices.

Theorem 2.2.2 *(Existence of a particular Hermitian matrix, see Thm 4.3.32 in [31].) Let* $\boldsymbol{\lambda}, \mathbf{d} \in \mathbf{R}^n$*, and satisfy (2.1), then there exists a Hermitian matrix* $\mathbf{M}$ *such that* $\mathbf{d}$ *is the vector of diagonal elements of* $\mathbf{M}$*, and* $\boldsymbol{\lambda}$ *is the vector of eigenvalues of* $\mathbf{M}$*.*

The above two theorems are very important in the optimization of transceivers for MIMO channels. One nice application is Witsenhausen's observation used in [88]. The other use of these two theorems is in the unified framework of linear transceiver optimization provided by Palomar *et. al* in [73]. The following example is very crucial in establishing the results of [73].

Example 2.4: Let $\mathbf{d}, \boldsymbol{\lambda} \in \mathbf{R}^n$ denote the vector of diagonal elements and the vector of eigenvalues of a Hermitian matrix $\mathbf{M}$, respectively. Let $\bar{\mathbf{d}} = \bar{d} \times [1, 1, \cdots, 1]^T$ denote the vector with all elements equal to the arithmetic mean of elements in $\mathbf{d}$. Thus, $\boldsymbol{\lambda} \succ_+ \mathbf{d} \succ_+ \bar{\mathbf{d}}$, and by Theorem 2.2.2, there exists a Hermitian matrix $\mathbf{M}'$ such that the vector of the eigenvalues is $\boldsymbol{\lambda}$ and the vector of diagonal

elements is $\bar{\mathbf{d}}$. Here is a way to obtain such matrix. Suppose the eigenvalue decomposition of $\mathbf{M}$ is $\mathbf{M} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^\dagger$. Let $\mathbf{Q}$ be any unitary matrix with identical magnitudes for all its elements, i.e., $|[\mathbf{Q}]_{ij}| = 1/\sqrt{n}$. Examples of such matrices are the normalized DFT matrix and the normalized Hadamard matrix for certain values of n [49, 73]. Let $\mathbf{M}' = \mathbf{Q}^\dagger\mathbf{T}^\dagger\mathbf{\Lambda}\mathbf{T}\mathbf{Q}$, then for any diagonal of $\mathbf{M}'$,

$$[\mathbf{M}']_{kk} = \sum_{i=1}^{n}[\mathbf{Q}^\dagger]_{ki}\lambda_i[\mathbf{Q}^\dagger]_{ik} = \frac{1}{n}\sum_{i=1}^{n}\lambda_i = \bar{d}.$$

Therefore, $\mathbf{M}'$ is an example of a Hermitian matrix with diagonal elements $\bar{\mathbf{d}}$ and eigenvalues $\boldsymbol{\lambda}$. $\diamondsuit$

2.2.2 Complex-Valued Square Matrices

The fundamental results on the relation between eigenvalues and singular values of a square complex-valued matrix was first established by Weyl in 1949 [134]. The findings can be presented as follows:

Theorem 2.2.3 *(Eigenvalues and singular values [134, 32, 140].) Let $\mathbf{M} \in \mathbf{C}^{n\times n}$, and let λ_i and σ_i denote the eigenvalues and singular values of $\mathbf{M}$, respectively. Then*

$$[\ \sigma_1^2, \cdots, \sigma_n^2\ ]^T \succ_\times [\ |\lambda_1|^2, \cdots, |\lambda_n|^2\ ]^T. \tag{2.2}$$

That is, multiplicative majorization relationship exists in eigenvalues and singular values of a complex valued matrix. $\diamondsuit$

It is surprising that the converse is also true. This important result was established by Horn in 1954 [30].

Theorem 2.2.4 *(The converse of Theorem 2.2.3, see [30] or Thm 3.6.6 in [32].) Let $\boldsymbol{\sigma} \in \mathbf{R}_+^n$, and $\boldsymbol{\lambda} \in \mathbf{C}^n$. Suppose $\boldsymbol{\sigma}$ and $\boldsymbol{\lambda}$ satisfy (2.2), then there exists a square matrix $\mathbf{M} \in \mathbf{C}^{n\times n}$ such that $\boldsymbol{\sigma}$ is the set of singular values of $\mathbf{M}$, and $\boldsymbol{\lambda}$ is the set of eigenvalues of $\mathbf{M}$.* $\diamondsuit$

These two theorems establish the complete characterization of the relation between eigenvalues and singular values of a square matrix. The algorithms to obtain such matrices were given in [30, 32].

2.3 Generalized Triangular Decomposition

Generalized triangular decomposition (GTD) was developed by Jiang *et. al.* in 2007 [38]. It utilizes the multiplicative majorization relation of eigenvalues and singular values in a square matrix, and also generalizes this relation to rectangular matrices. To be more specific, the theorem statement of GTD is as follows.

Theorem 2.3.1 *(Generalized triangular decomposition.) Let* $\mathbf{H} \in \mathcal{C}^{m\times n}$ *be a rank-*K *matrix with singular values* $\sigma_{h,1}, \sigma_{h,2}, \cdots, \sigma_{h,K}$ *in descending order. Let* $\mathbf{r} = [r_1, r_2, \cdots, r_K]$ *be any vector which satisfies*

$$\mathbf{a} \prec_{\times} \mathbf{h}, \tag{2.3}$$

where $\mathbf{a} = [|r_1|, |r_2|, \cdots, |r_K|]$ *and* $\mathbf{h} = [\sigma_{h,1}, \sigma_{h,2}, \cdots, \sigma_{h,K}]$*. Then there exist matrices* $\mathbf{R}$*,* $\mathbf{Q}$*, and* $\mathbf{P}$ *such that* $\mathbf{H}$ *can be decomposed as*

$$\mathbf{H} = \mathbf{Q}\mathbf{R}\mathbf{P}^{\dagger}, \tag{2.4}$$

where $\mathbf{R}$ *is a* $K\times K$ *upper triangular matrix with diagonal terms equal to* r_k*, and* $\mathbf{Q} \in \mathcal{C}^{m\times K}$ *and* $\mathbf{P} \in \mathcal{C}^{n\times K}$ *both have orthonormal columns.* ♢

According to the GTD factorization algorithm described in [38], if $\mathbf{H}$ and $\mathbf{r}$ are real-valued, then the matrices $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{P}$ can be taken to be real-valued.

The GTD can also be viewed as triangularization of the input matrix $\mathbf{H}$ by two semi-unitary matrices ($\mathbf{P}$ and $\mathbf{Q}$) on both sides. There are many standard decompositions that can be regarded as special instances of the GTD. These are listed below. The first five can be found in standard texts [25, 31], while the sixth was proposed by Jiang *et. al.* in the context of MIMO transceiver optimization.

1. The *singular value decomposition* (SVD), $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\dagger}$ where $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values on the diagonal.

2. The *Schur decomposition*, $\mathbf{H} = \mathbf{Q}\mathbf{\Delta}\mathbf{Q}^{\dagger}$ where $\mathbf{\Delta}$ is an upper triangular matrix with eigenvalues of a square matrix $\mathbf{H}$ on the diagonal.

3. The *QR decomposition*, $\mathbf{H} = \mathbf{Q}\mathbf{R}$ where $\mathbf{R}$ is an upper triangular matrix (here $\mathbf{P} = \mathbf{I}$).

4. The *complete orthogonal decomposition*, $\mathbf{H} = \mathbf{Q}_2\mathbf{R}_2\mathbf{Q}_1^\dagger$, where $\mathbf{H}^\dagger = \mathbf{Q}_1\mathbf{R}_1$ is the QR factorization of $\mathbf{H}^\dagger$ and $\mathbf{R}_1^\dagger = \mathbf{Q}_2\mathbf{R}_2$ is the QR factorization of $\mathbf{R}_1^\dagger$.

5. The *bidiagonal decomposition*, $\mathbf{H} = \mathbf{QBP}^\dagger$, where $\mathbf{B}$ is a bidiagonal and upper triangular matrix (page 251 of [25]).

6. The *geometric mean decomposition* (GMD) [36], $\mathbf{H} = \mathbf{QRP}^\dagger$ where $\mathbf{R}$ is an upper triangular matrix with the diagonal elements equal to the geometric means of the positive singular values.

The computation of performing GTD has the same asymptotic complexity as the singular value decomposition [38]. Jiang *et. al.* provided a numerically stable algorithm to compute any GTD with prescribed diagonal elements as long as the multiplicative majorization relation is satisfied. The algorithm starts by performing the singular value decomposition to obtain the diagonal matrix with singular values on the diagonals. Then, successive Givens rotations are performed to produce the middle upper triangular matrix with prescribed diagonal elements. Note that the singular values are invariant under Givens rotations. The details of this algorithm and a computer code implementation can be found in [38].

2.3.1 Block-Diagonal Geometric Mean Decomposition

The block diagonal geometric mean decomposition (BD-GMD) is a special case of the GMD. In BD-GMD, one of the unitary matrices in (2.4) is restricted to be block diagonal. The price paid is that the diagonal elements of the middle triangular matrix can no longer be made equal. Instead, they are block-wise equal. To be more specific, consider the matrix decomposition of the following form. Suppose $\mathbf{H} \in C^{m\times n}$ is with full column rank n and we are seeking the decomposition

$$\mathbf{H}^\dagger = \mathbf{PLQ}^\dagger,$$

where $\mathbf{Q}$ is a $m \times n$ matrix with orthonormal columns, $\mathbf{L}$ is a $n \times n$ lower triangular matrix, and $\mathbf{P}$ is a $n \times n$ block diagonal matrix of the form $\text{diag}(\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_K)$ where each block $\mathbf{P}_i$ is a unitary $n_i \times n_i$ matrix. The task is to find a matrix decomposition such that the diagonal elements of $\mathbf{L}$ are equal in blocks of $n_1, \cdots, n_K$ elements respectively. This decomposition was proposed by Lin *et. al.* in [47] to address the transceiver design problems for MIMO broadcast channels.

The algorithm proposed by [47] is as follows. First, rewrite the decomposition as

$$\begin{bmatrix} \mathbf{H}_1^\dagger \\ \mathbf{H}_r^\dagger \end{bmatrix} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_r \end{bmatrix} \begin{bmatrix} \mathbf{L}_1 & \mathbf{0} \\ \mathbf{A} & \mathbf{L}_r \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1^\dagger \\ \mathbf{Q}_r^\dagger \end{bmatrix} \tag{2.5}$$

where $\mathbf{H}_1^\dagger$ and $\mathbf{Q}_1^\dagger$ are $n_1 \times m$ submatrices, and $\mathbf{L}_1$ and $\mathbf{P}_1$ are $n_1 \times n_1$ square matrices. $\mathbf{H}_r^\dagger$ denotes the remaining lower part of the matrix $\mathbf{H}^\dagger$. Expanding the above equation gives the following two equations

$$\mathbf{H}_1^\dagger = \mathbf{P}_1 \mathbf{L}_1 \mathbf{Q}_1^\dagger \tag{2.6}$$

and

$$\mathbf{H}_r^\dagger = \mathbf{P}_r \mathbf{A} \mathbf{Q}_1^\dagger + \mathbf{P}_r \mathbf{L}_r \mathbf{Q}_r^\dagger. \tag{2.7}$$

From Equation (2.5), it can be seen that by performing the GMD, the diagonal elements of $\mathbf{L}_1$ can be made equal. Since $\mathbf{Q}$ has orthonormal columns, the submatrices $\mathbf{Q}_1$ and $\mathbf{Q}_r$ are orthonormal to each other. Thus, from Equation (2.7), multiplication by the projection matrix $\mathbf{I} - \mathbf{Q}_1 \mathbf{Q}_1^\dagger$ gives

$$\mathbf{H}_r^\dagger (\mathbf{I} - \mathbf{Q}_1 \mathbf{Q}_1^\dagger) = \mathbf{P}_r \mathbf{L}_r \mathbf{Q}_r^\dagger.$$

Here, the right side of the above equation has the same form as in (2.5), so the algorithm proceeds recursively. To solve for $\mathbf{A}$, equation (2.7) is multiplied by $\mathbf{P}_r^\dagger$ and $\mathbf{Q}_1$ on the left and right hand side, respectively, giving

$$\mathbf{A} = \mathbf{P}_r^\dagger \mathbf{H}_r^\dagger \mathbf{Q}_1.$$

This decomposition then has equal diagonal elements in each block of $\mathbf{L}$. After performing the BD-GMD, the matrix $\mathbf{H}^\dagger$ can be decomposed as follows:

$$\mathbf{H}^\dagger = \begin{bmatrix} \mathbf{H}_1^\dagger \\ \mathbf{H}_2^\dagger \\ \vdots \\ \mathbf{H}_K^\dagger \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{P}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{P}_K \end{bmatrix}}_{\mathbf{P}} \underbrace{\begin{bmatrix} \mathbf{L}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \times & \mathbf{L}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \times & \cdots & \times & \mathbf{L}_K \end{bmatrix}}_{\mathbf{L}} \underbrace{\begin{bmatrix} \mathbf{Q}_1^\dagger \\ \mathbf{Q}_2^\dagger \\ \vdots \\ \mathbf{Q}_K^\dagger \end{bmatrix}}_{\mathbf{Q}^\dagger}, \tag{2.8}$$

where $\mathbf{P}$ is unitary and also block diagonal, and each submatrix $\mathbf{L}_i$ has equal diagonal elements.

Chapter 3

Transceiver Designs for MIMO Frequency Flat Channels

In this chapter we consider the optimization of transceivers for frequency flat MIMO channels. The first part of this chapter studies the optimization of MIMO transceivers with linear precoders, decision feedback equalizers, and bit allocation schemes. Considered first is the minimization of the average transmitted power, for a given total bit rate and a specified set of error probabilities for the symbol streams. While this joint optimization has not been addressed in the past, a variety of related transceiver designs have been studied previously. When the transmitter has perfect channel information, four major optimization problems have been considered. First, for fixed precoder and DFE matrices, the optimal bit loading problem has been studied in [9]. Second, for the case where the bit allocation is fixed to be uniform, joint optimization of the precoder and DFE matrices is a well studied problem [36, 37, 41, 90, 137, 142]. Third, for the case of linear transceivers, the joint optimization of the precoder, the (linear) equalizer, and bit allocation has been studied in [48] (under ZF constraint), and in [74] (without ZF constraint). Fourth, if the precoding matrix is restricted to be a diagonal matrix where only power loading applies, the optimization of rate and power allocation for the systems with DFE receiver has been discussed in [14]. If no perfect channel state information is present (only channel statistics known at the transmitter), the optimization of power and rate allocation for the system with DFE receiver was addressed in [83].

As summarized above, bit loading, precoder, and receiver design optimizations have been studied extensively. However, current literature lacks a discussion that reviews bit loading, linear precoder, and DFE *jointly* when perfect channel information is available at both ends of the transceiver. To begin solving this problem, we start with the minimization of transmitted power for a specified set of error probabilities for the symbol streams. We show that the generalized triangular decom-

position (GTD) introduced in [38] offers an optimal solution. The GTD in fact gives rise to a family of solutions, with the bit allocation details changing from solution to solution. We will see in particular that the optimal linear transceiver with optimal bit allocation, which has a *linear* equalizer rather than a DFE, is a member of this family of solutions. This shows formally that, under optimal bit allocation, optimum linear transceivers achieve the same transmitted power as optimum DFE transceivers with bit allocation. These discussions assume that the bit allocation formula is realizable (i.e, the bits are nonnegative integers). The DFE transceiver based on the geometric mean decomposition (GMD) [36] is another member of the above family of optimal solutions, and is such that the optimal bit allocation formula yields *identical* bits for all symbol substreams, when the specified error probabilities are identical for the substreams. DFE with GMD therefore achieves minimum power even without the need for bit allocation. In a way this complements one of the results in [36], namely, when all symbol streams are constrained to have identical bits, the average bit error rate (BER) for fixed power is minimized by the GMD. Other special cases arising from the GTD family of optimal DFE systems include the VBLAST system [135], and a new solution called the bidiagonal (BID) transceiver. Two other optimization problems are then considered: (a) minimization of power for specified set of bit rates and error probabilities (the QoS problem), and (b) maximization of bit rate for a fixed set of error probabilities and power. It is shown in both cases that the GTD yields an optimal family of solutions.

The second part of this chapter considers the joint transceiver optimization problem for frequency flat MIMO channels under other constraints. The linear transceiver as well as the DFE transceiver are considered. Instead of only the total power constraint, in this section we also consider the more realistic per-antenna power constraints on the transmitter [45, 139]. This is because in practice each antenna is limited individually by its equipped power amplifier. In [45], the MMSE problem under individual power constraints is solved sub-optimally using a numerical approach. In [139], the multiuser down-link transceiver design problem is considered. The total power constraint might still be needed since the antennas might rely on a common power supply. Under these constraints, we consider the linear transceiver case and also the simple nonlinear case, i.e., linear precoding with DFE at the equalizer.

The problem of optimizing linear transceivers subject to individual power constraints was addressed in [45, 73, 76], in different contexts, but transceivers with DFE were not considered. In [45], the authors considered the MMSE problem, and solved it sub-optimally using numerical methods.

In [73], the authors considered only Schur-concave objective functions subject to the individual power constraints. However, the problem of optimizing the transceiver for other important objective functions (e.g., Schur-convex functions, including average BER) was not addressed. Direct use of the results of [73] to address this case is nontrivial. In [76] the author considered shaping constraints on the transmitted signal covariance matrix. However, as acknowledged by the author in [76], the paper introduced a stronger artificial constraint that leads to a sub-optimal solution.

Our work is to tackle these unsolved problems in the literature. For the linear transceiver case, we first consider the minimum AM-MSE (Arithmetic Mean of Mean Square Errors) design. We show that it can be reformulated as a semi-definite program (SDP), which can be solved numerically by convex optimization tools. Then, among the family of minimum AM-MSE linear transceivers, we develop a method to find the one that minimizes the average bit error rate as well as many other objective functions. This second step is achieved by appealing to majorization theory. Similarly, for the transceivers with linear precoding and DFE, we first consider the minimum GM-MSE (Geometric Mean of Mean Square Errors) design. We show that it can also be reformulated as an SDP, and solved efficiently. Then, among the family of minimum GM-MSE designs, we develop a method to find the one that minimizes the average bit error rate as well as many other objective functions.

Based on majorization theory [65], we will argue that the minimal average BER transceiver design method developed in this section can also be applied to a wider class of objective functions. Also, we will show that under the framework developed here, any additional linear constraints on the covariance matrix of the transmitted signals can be further added, and the problem is solved both in theory and practice with no difficulty. Examples of such constraints may be spatial power masks. This was formulated but not elaborated in [76]. In addition, it is shown in Section 3.3.5 that our framework includes the case studied in [76].

The content of this chapter is mainly drawn from [123, 128], and portions of it have been presented in [119, 120, 125].

3.1 Outline

This chapter is organized as follows. Sec. 3.2 discusses the joint optimization of linear precoder, DFE receiver, and bit loading for MIMO communication systems. In Sec. 3.3 we consider the linear and DFE transceiver design problems under a finite number of linear constraints on the transmit

covariance matrix, including total power constraint and individual antenna power constraints. Finally, the conclusions are made in Sec. 3.4.

3.2 MIMO Transceivers with Decision Feedback and Bit Loading

In this section, we consider MIMO transceivers with a linear precoder and a decision feedback equalizer (DFE), with bit allocation allowed at the transmitter end. Zero-forcing and QAM signaling are considered throughout, and the perfect channel information is assumed to be known to the transmitter and the receiver. This section is structured as follows. In Sec. 3.2.1, we formulate the power minimization problem. In Sec. 3.2.2, we show that under optimal bit allocation, optimum linear transceivers achieve the same minimum value for transmitted power as optimum DFE transceivers with bit allocation. In Sec. 3.2.3 we present a transceiver structure based on generalized triangular decomposition of the channel matrix, and prove that such a system always achieves the minimum power given in Sec. 3.2.2. We also report several special cases of the optimal solutions developed from GTD. Some of these are known structures (SVD, GMD, and VBLAST or QR-based) and some are new (e.g., the bi-diagonal structure). Two other optimizations are also considered in Sec. 3.2.4: (a) minimization of power for fixed set of bit rates and error probabilities (the QoS problem), and (b) maximization of bit rate for fixed power and error probabilities. It is shown in both cases that the GTD yields optimal solutions. Sec. 3.2.5 and 3.2.6 present the simulation results.

3.2.1 Problem Formulation

The transceiver we consider is shown in Figure 3.11. Here $\mathbf{H}$ is a $J \times N$ memoryless channel matrix, and the additive Gaussian noise $\mathbf{n}$ is assumed to have zero-mean and covariance matrix $\mathrm{E}\{\mathbf{n}\mathbf{n}^\dagger\} = \sigma_n^2\mathbf{I}$. It is assumed that $\mathbf{H}$ is deterministic and known to the transmitter and receiver (perfect CSIT and CSIR). The linear precoder matrix is denoted as $\mathbf{F}$. The vector $\mathbf{s}(n)$ represents the M transmitted symbol streams $s_k(n)$ (with time argument n deleted in all discussions). The received signal is $\mathbf{y} = \mathbf{H}\mathbf{x}+\mathbf{n}$, where $\mathbf{x} = \mathbf{F}\mathbf{s}$. The DFE equalizer consists of the feedforward part $\mathbf{G}$ and feedback part $\mathbf{B}$. Causality of decision feedback is ensured by restricting $\mathbf{B}$ to be strictly upper triangular. With $\tilde{\mathbf{s}}$ denoting the signal vector after the decision device, the input to the decision device has the form $\hat{\mathbf{s}} = \mathbf{G}\mathbf{H}\mathbf{F}\mathbf{s} - \mathbf{B}\mathbf{s} + \mathbf{G}\mathbf{n}$. Under the assumption of correct past decisions, i.e.,

$\tilde{\mathbf{s}} = \mathbf{s}$ (a good assumption in the high SNR regime), this yields

$$\hat{\mathbf{s}} = (\mathbf{GHF} - \mathbf{B})\mathbf{s} + \mathbf{Gn}. \tag{3.1}$$

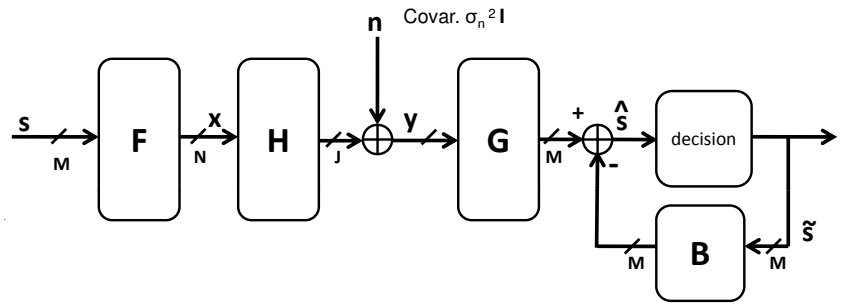


Figure 3.1: The MIMO transceiver with linear precoder and DFE.

Eq. (3.1) shows that the system described above has an effective transfer matrix $\mathbf{GHF} - \mathbf{B}$ from $\mathbf{s}$ to $\hat{\mathbf{s}}$, and an additive noise term $\mathbf{Gn}$. It therefore has the zero-forcing (ZF) property if

$$\mathbf{GHF} - \mathbf{B} = \mathbf{I}. \tag{3.2}$$

Zero-forcing will be assumed throughout the section, so that

$$\hat{\mathbf{s}} = \mathbf{s} + \mathbf{Gn}. \tag{3.3}$$

Without this assumption the problems to be addressed are more difficult, and will be left for the future. Since $\mathbf{GHF} = \mathbf{I} + \mathbf{B}$ is upper triangular with unit diagonal elements, it has rank M. To make the zero-forcing assumption possible, $\mathbf{H}$ is assumed to have rank $K \geq M$.

In the following sections, we will first discuss the problem of minimizing the transmitted power subject to a specified total bit rate and a specified error probability in each substream. Assume the components s_k of $\mathbf{s}$ are zero-mean uncorrelated processes representing independent data streams with power P_k so that the input covariance is

$$\mathbf{\Lambda}_s = \mathrm{E}\{\mathbf{s}\mathbf{s}^\dagger\} = \mathrm{diag}(P_1, P_2, \cdots, P_M). \tag{3.4}$$

We assume the kth data stream is a b_k-bit QAM constellation. From (3.3), since the error vector at the input of the decision device is $\mathbf{e} = \hat{\mathbf{s}} - \mathbf{s} = \mathbf{Gn}$ where $\mathbf{n}$ is zero-mean Gaussian, the error

components e_k are zero-mean Gaussian with variance

$$\sigma_{e_k}^2 = \sigma_n^2[\mathbf{G}\mathbf{G}^\dagger]_{kk}. \tag{3.5}$$

The probability of error for the kth symbol stream is then [85]

$$P_e(k) \approx 4(1-2^{\frac{-b_k}{2}})Q\left(\sqrt{\frac{3P_k}{(2^{b_k}-1)\sigma_{e_k}^2}}\right), \tag{3.6}$$

where $Q(\tau) = \int_\tau^\infty e^{-t^2/2}dt/\sqrt{2\pi}$. Under the high bit rate assumption ($b_k >> 1$) we have $2^{b_k}-1 \approx 2^{b_k}$ and $1-2^{-b_k/2} \approx 1$. By rearranging Eq. (3.6) we then get

$$\frac{P_k}{\sigma_{e_k}^2} \approx \frac{2^{b_k}}{3}\left(Q^{-1}\left(\frac{P_e(k)}{4}\right)\right)^2, \tag{3.7}$$

where $Q^{-1}(\cdot)$ denotes the inverse function of $Q(\cdot)$. This is the average power to noise ratio required for the kth QAM stream to operate at error probability $P_e(k)$ with b_k-bits.

In this section we will regard the error probability $P_e(k)$ as the quality of service (QoS) specification. For the special case of DMT systems one takes all $P_e(k)$ to be equal [48]. The total power transmitted on the channel can be written as

$$P_{trans} = \text{Tr}(\mathbf{F}\mathbf{\Lambda}_s\mathbf{F}^\dagger) = \text{Tr}(\mathbf{F}^\dagger\mathbf{\Lambda}_s\mathbf{F}) = \sum_{k=1}^{M} P_k[\mathbf{F}^\dagger\mathbf{F}]_{kk}.$$

Substituting from (3.7) we can rewrite this as

$$P_{trans} = \sum_{k=1}^{M} d_k 2^{b_k}\sigma_{e_k}^2[\mathbf{F}^\dagger\mathbf{F}]_{kk}, \tag{3.8}$$

where

$$d_k = \frac{1}{3}\left(Q^{-1}\left(\frac{P_e(k)}{4}\right)\right)^2, \tag{3.9}$$

which is determined by the specified probability of error. From (3.8) and (3.5) the transmitted power can then be written as

$$P_{trans} = \sum_{k=1}^{M} c_k 2^{b_k}[\mathbf{F}^\dagger\mathbf{F}]_{kk}[\mathbf{G}\mathbf{G}^\dagger]_{kk}, \tag{3.10}$$

where

$$c_k = \sigma_n^2 d_k = \frac{\sigma_n^2}{3}\left(Q^{-1}\left(\frac{P_e(k)}{4}\right)\right)^2. \tag{3.11}$$

Therefore the problem of minimizing the transmitted power subject to the specified BER and total bit rate constraints, and the zero-forcing constraint can be written as follows:

$$\begin{aligned} \min_{\mathbf{F},\mathbf{G},\mathbf{B},\{b_k\}} \quad & P_{trans} = \sum_{k=1}^{M} c_k 2^{b_k} [\mathbf{F}^\dagger \mathbf{F}]_{kk} [\mathbf{G}\mathbf{G}^\dagger]_{kk} \\ \text{s.t.} \quad & (a) \quad \frac{1}{M}\sum_{k=1}^{M} b_k = b \\ & (b) \quad \mathbf{GHF} - \mathbf{B} = \mathbf{I} \end{aligned} \tag{3.12}$$

Ideally, we should also impose the constraint that b_k be nonnegative integers. But the problem is not analytically tractable in that case. For the high bit rate case (large b) the optimal bit allocation formula derived next yields positive b_k, which can be rounded to integers without severe loss of optimality.

3.2.2 Minimum Power Achieved by DFE Systems

For the minimization of (3.12), we first observe that

$$\begin{aligned} P_{trans} &= \sum_{k=1}^{M} c_k 2^{b_k} [\mathbf{F}^\dagger \mathbf{F}]_{kk} [\mathbf{G}\mathbf{G}^\dagger]_{kk} \\ &\geq M \prod_{k=1}^{M} (c_k 2^{b_k} [\mathbf{F}^\dagger \mathbf{F}]_{kk} [\mathbf{G}\mathbf{G}^\dagger]_{kk})^{\frac{1}{M}} \\ &= c 2^b (\prod_{k=1}^{M} [\mathbf{F}^\dagger \mathbf{F}]_{kk})^{\frac{1}{M}} (\prod_{k=1}^{M} [\mathbf{G}\mathbf{G}^\dagger]_{kk})^{\frac{1}{M}}, \end{aligned}$$

where we have used the AM-GM inequality, and the fact that

$$b = \frac{1}{M}\sum_{k=1}^{M} b_k. \tag{3.13}$$

Here

$$c = M(\prod_{k=1}^{M} c_k)^{\frac{1}{M}}. \tag{3.14}$$

Equality can be achieved in the AM-GM inequality if and only if the terms are identical for all k, that is,

$$c_k 2^{b_k} [\mathbf{F}^\dagger \mathbf{F}]_{kk} [\mathbf{G}\mathbf{G}^\dagger]_{kk} = A$$

for some constant A. Taking logarithms on both sides we get

$$b_k = D - \log_2 c_k - \log_2 [\mathbf{F}^\dagger \mathbf{F}]_{kk} - \log_2 [\mathbf{G}\mathbf{G}^\dagger]_{kk}, \tag{3.15}$$

where D is a constant, chosen such that (3.13) is satisfied. Eq. (3.15) is called the optimum bit loading formula.[1] For any fixed precoder $\mathbf{F}$, receiver $\{\mathbf{G}, \mathbf{B}\}$, and specified probabilities of error $P_e(k)$, the bit allocation that minimizes the transmitted power is given by (3.15). With this b_k, the quantities P_k are computed from setting $\sigma_{e_k}^2$ as in (3.5). With P_k so chosen, the specified probabilities of error are met, and the total power P_{trans} is minimized. This minimized power is

$$P_{trans} = c 2^b (\prod_{k=1}^{M} [\mathbf{F}^\dagger \mathbf{F}]_{kk})^{\frac{1}{M}} (\prod_{k=1}^{M} [\mathbf{G}\mathbf{G}^\dagger]_{kk})^{\frac{1}{M}}), \tag{3.16}$$

and depends only on $\mathbf{F}$ and $\mathbf{G}$, which will be chosen to minimize (3.16) further. First we derive the optimal $\mathbf{G}$:

Lemma 3.2.1 *When the precoder $\mathbf{F}$ and the feedback filter $\mathbf{B}$ are given, the optimal feed-forward filter $\mathbf{G}$ for minimizing the transmitted power in (3.16) subject to the zero forcing condition (3.2) is:*

$$\mathbf{G}_{opt} = (\mathbf{I} + \mathbf{B})(\mathbf{H}\mathbf{F})^\sharp, \tag{3.17}$$

where $(\mathbf{H}\mathbf{F})^\sharp = (\mathbf{F}^\dagger \mathbf{H}^\dagger \mathbf{H}\mathbf{F})^{-1} \mathbf{F}^\dagger \mathbf{H}^\dagger$, which is the minimum-norm pseudo inverse of $(\mathbf{H}\mathbf{F})$. ♢

Proof: See Appendix. □

The zero-forcing constraint yields the form (3.17), which can also be found in other references such as [137], where a different problem is solved (mean square error minimized subject to zero-forcing, without bit allocation). The main point of the lemma is that the pseudo inverse $(\mathbf{H}\mathbf{F})^\sharp$ should be taken to be the *minimum norm* pseudoinverse. This also happens in [137] but the proof

[1] In general (3.15) can yield noninteger or even negative b_k. However in the high-bit-rate case (large b), b_k are large enough to be replaced with integer values without compromising optimality severely. The conclusions derived in the following discussions are valid only under this assumption which has often been made in other papers [50, 48]. Incorporating the positive integer constraint directly into the problem makes it analytically non-tractable.

techniques for the two problems are different.

Substitute for $\mathbf{G}_{opt}$ into (3.16) we get

$$P_{trans} = c2^b(\prod_{k=1}^{M}[\mathbf{F}^\dagger\mathbf{F}]_{kk})^{\frac{1}{M}} \times (\prod_{k=1}^{M}[(\mathbf{I}+\mathbf{B})(\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F})^{-1}(\mathbf{I}+\mathbf{B})^\dagger]_{kk})^{\frac{1}{M}}.$$

Hadamard's inequality for positive definite matrices yields

$$\prod_{k=1}^{M}[\mathbf{F}^\dagger\mathbf{F}]_{kk} \geq \det(\mathbf{F}^\dagger\mathbf{F})$$

and

$$\begin{aligned}&\prod_{k=1}^{M}[(\mathbf{I}+\mathbf{B})(\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F})^{-1}(\mathbf{I}+\mathbf{B})^\dagger]_{kk}\\ &\geq \det((\mathbf{I}+\mathbf{B})(\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F})^{-1}(\mathbf{I}+\mathbf{B})^\dagger)\\ &= \det((\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F})^{-1}),\end{aligned}$$

where we use the fact that $\det(\mathbf{I}+\mathbf{B}) = 1$ since $\mathbf{I}+\mathbf{B}$ is upper triangular with diagonal terms all equal to unity. Substituting the above result into the transmitted power, we have

$$P_{trans} \geq c2^b\left(\frac{\det(\mathbf{F}^\dagger\mathbf{F})}{\det(\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F})}\right)^{\frac{1}{M}}.$$

In the Appendix we prove that

$$P_{trans} \geq P_{min} = c2^b\left(\frac{1}{\prod_{k=1}^{M}\sigma_{h,k}^2}\right)^{\frac{1}{M}}, \tag{3.18}$$

where $\{\sigma_{h,k}\}_{k=1}^{M}$ are the first M dominant channel singular values. Note that P_{min} in (3.18) is exactly equal to the form derived for a linear transceiver with optimal bit loading [48]. This means, the extra freedom provided by the decision feedback receiver structure does not reduce the power needed to achieve the specified bit rate and probability of error. So we have proved:

Theorem 3.2.2 Linear versus DFE transceiver: *Consider the DFE system of Fig. 3.11 and assume the bit rate b and error probabilities $P_e(k)$ are fixed. Then under optimal bit allocation and zero-forcing, the minimum transmitted power obtained by optimizing $\mathbf{F}, \mathbf{G}$, and $\mathbf{B}$ is given by P_{min} defined in Eq. (3.18). This same minimum power can also be achieved by a linear transceiver (a transceiver with $\mathbf{B} = \mathbf{0}$) by*

optimizing $\mathbf{F}, \mathbf{G}$, *and the bit allocation under the zero-forcing constraint.* $\diamondsuit$

Thus, when bit loading is allowed, *DFE with linear precoding has the same performance as linear transceivers!* However, the DFE system with linear precoding actually provides more choices of possible configurations that achieve the P_{min} in (3.18). This interesting observation will be elaborated further in the following subsections.

3.2.3 GTD-Based Transceivers

We now show that the generalized triangular decomposition (GTD) can be used to construct optimal solutions to the problem (3.12). Before diving into any specific realization, we describe in detail the GTD-based method to construct the transceiver matrices $\mathbf{F}, \mathbf{G}, \mathbf{B}$. Here are the steps involved:

1. Given the channel $\mathbf{H}$, we first choose a set of diagonal elements r_k for $\mathbf{R}$ such that (2.3) holds, and express $\mathbf{H}$ in the GTD form (2.4), thereby determining a set of matrices $\mathbf{P}, \mathbf{Q}$ and $\mathbf{R}$.

2. We then show how to choose the precoder $\mathbf{F}$, the receiver matrices $\mathbf{G}$ and $\mathbf{B}$, and the bit allocation such that the transmitted power achieves the minimum value P_{min} given by (3.18).

The first step offers considerable freedom, since any choice for the diagonal elements $r_k = [\mathbf{R}]_{kk}$ is acceptable as long as (2.3) holds. We will choose the K elements r_k as follows: (a) Choose $r_1, r_2, \ldots r_M$ to be any set of positive numbers multiplicatively majorized by the first M dominant singular values $\sigma_{h,1} \geq \ldots \geq \sigma_{h,M}$ of the channel. (b) Choose $r_{M+1}, \ldots r_K$ to be $\sigma_{h,M+1}, \ldots, \sigma_{h,K}$ or any permutation thereof. The choice in (a) implies in particular that

$$\prod_{k=1}^{M} [\mathbf{R}]_{kk}^2 = \prod_{k=1}^{M} \sigma_{h,k}^2. \tag{3.19}$$

With $[\mathbf{R}]_{kk}$ chosen as above, assume the channel has been expressed as in Eq. (2.4). We are now ready for the second step. We begin by choosing $N \times M$ precoder as

$$\mathbf{F} = [\mathbf{P}]_{N \times M}. \tag{3.20}$$

Thus the columns of the precoder are the first M columns of $\mathbf{P}$. We then choose the feedforward matrix as

$$\mathbf{G} = \left(\text{diag}([\mathbf{R}]_{M \times M})\right)^{-1} \mathbf{G}_0, \tag{3.21}$$

where

$$\mathbf{G}_0 = [\mathbf{Q}^\dagger]_{M\times J}. \tag{3.22}$$

Since $\mathbf{P}$ and $\mathbf{Q}$ have orthonormal columns, the columns of $\mathbf{F}$ are orthonormal, and so are the rows of $\mathbf{G}_0$. Finally, the feedback matrix $\mathbf{B}$ is determined by the zero forcing condition $\mathbf{B} = \mathbf{GHF} - \mathbf{I}$. To simplify this, observe first that

$$\begin{aligned}
\mathbf{GHF} &= \left(\text{diag}([\mathbf{R}]_{M\times M})\right)^{-1} \mathbf{G}_0\mathbf{QRP}^\dagger\mathbf{F} \\
&= \left(\text{diag}([\mathbf{R}]_{M\times M})\right)^{-1} \begin{pmatrix} \mathbf{I_M} & \mathbf{0} \end{pmatrix} \mathbf{R} \begin{pmatrix} \mathbf{I_M} \\ \mathbf{0} \end{pmatrix} \\
&= \left(\text{diag}([\mathbf{R}]_{M\times M})\right)^{-1} [\mathbf{R}]_{M\times M}.
\end{aligned}$$

Here we have used the facts that

$$\mathbf{G}_0\mathbf{Q} = (\mathbf{I}_M \;\; \mathbf{0}) \quad \text{and} \quad \mathbf{P}^\dagger\mathbf{F} = \begin{pmatrix} \mathbf{I_M} \\ \mathbf{0} \end{pmatrix}$$

which follow from the choices of (3.20) and (3.22), and the column orthonormality of $\mathbf{P}$ and $\mathbf{Q}$. Thus the expression for the feedback matrix becomes

$$\mathbf{B} = \mathbf{GHF} - \mathbf{I} = \left(\text{diag}([\mathbf{R}]_{M\times M})\right)^{-1} [\mathbf{R}]_{M\times M} - \mathbf{I}. \tag{3.23}$$

This is strictly upper triangular since $\mathbf{R}$ is upper triangular. Fig. 3.2 shows the structure of the GTD transceiver just described.

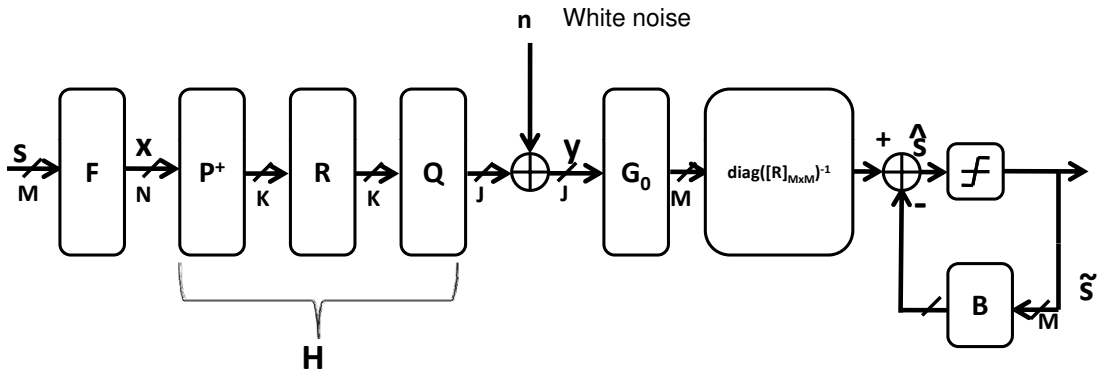


Figure 3.2: The proposed form of optimal solution for the DFE transceiver.

With the above choice of transceiver matrices the error variance (3.5) in the kth substream becomes

$$\sigma_{e_k}^2 = \frac{\sigma_n^2}{[\mathbf{R}]_{kk}^2}. \tag{3.24}$$

Substituting into Eq. (3.8) the transmitted power needed to satisfy the specified QoS and bit rate constraints can be expressed as

$$P_{trans} = \sum_{k=1}^{M} d_k 2^{b_k} [\mathbf{F}^\dagger \mathbf{F}]_{kk} \sigma_{e_k}^2 = \sum_{k=1}^{M} \frac{d_k 2^{b_k}}{[\mathbf{R}]_{kk}^2} \sigma_n^2.$$

Since $\sigma_n^2 d_k = c_k$ (from (3.11)), this simplifies to

$$P_{trans} = \sum_{k=1}^{M} \frac{c_k 2^{b_k}}{[\mathbf{R}]_{kk}^2}. \tag{3.25}$$

We now show that the system in Fig. 3.2 with $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{B}$ chosen as described achieves optimality for problem (3.12), provided the bit allocation is chosen appropriately:

Theorem 3.2.3 *With the bit allocation chosen as*

$$b_k = \log_2 \left(\frac{c}{M} 2^b \left(\frac{1}{\prod_{k=1}^{M} \sigma_{h,k}^2} \right)^{\frac{1}{M}} \right) - \log_2(c_k) + \log_2([\mathbf{R}]_{kk}^2), \tag{3.26}$$

for $1 \leq k \leq M$, the system in Fig. 3.2 with $\mathbf{F}$ as in (3.20), $\mathbf{G}$ as in (3.21), and $\mathbf{B}$ as in (3.23), achieves the minimized power for the specified $\{P_e(k)\}$ and bit rate constraint. ♢

Proof: Observe first that (3.26) satisfies the total bit constraint because

$$\begin{aligned} \sum_{k=1}^{M} b_k &= \log_2 \left(\left(\frac{c}{M} \right)^M \frac{2^{Mb}}{\prod_{k=1}^{M} \sigma_{h,k}^2} \right) - \log_2 \prod_{k=1}^{M} c_k + \log_2 \left(\prod_{k=1}^{M} [\mathbf{R}]_{kk}^2 \right) \\ &= Mb - \log_2 \left(\frac{1}{\prod_{k=1}^{M} \sigma_{h,k}^2} \right) + \log_2 \left(\prod_{k=1}^{M} [\mathbf{R}]_{kk}^2 \right) \\ &= Mb, \end{aligned}$$

using (3.19) and $c = M(\prod_{k=1}^{M} c_k)^{\frac{1}{M}}$. Next, (3.26) implies

$$\frac{c_k 2^{b_k}}{[\mathbf{R}]_{kk}^2} = \frac{c}{M} 2^b \left(\frac{1}{\prod_{k=1}^{M} \sigma_{h,k}^2} \right)^{\frac{1}{M}}. \tag{3.27}$$

Substituting into (3.25) we get

$$P_{trans} = \sum_{k=1}^{M} \frac{c_k 2^{b_k}}{[\mathbf{R}]_{kk}^2} = M \times \frac{c}{M} 2^b \Big(\frac{1}{\prod_{k=1}^{M} \sigma_{h,k}^2} \Big)^{\frac{1}{M}}. \tag{3.28}$$

Since this is the minimum achievable power P_{min} (see discussion leading to Eq. (3.18)), the proof is complete. □

The extra flexibility in designing the transceivers, offered by this GTD-based DFE system, must be carefully understood. Recall that the bit loading formula for the linear transceiver to achieve the minimum transmitted power is [48]

$$b_k = D - \log_2 c_k + \log_2(\sigma_{h,k}^2), \tag{3.29}$$

where $\sigma_{h,k}$ are fixed numbers given to us by the channel. The values computed from (3.29) are not guaranteed to be integers, or even nonnegative. For the GTD-based DFE system, the bit loading scheme (3.26) can be written as

$$b_k = D - \log_2 c_k + \log_2([\mathbf{R}]_{kk}^2). \tag{3.30}$$

The freedom of the GTD-based system allows us to reshape the value of $[\mathbf{R}]_{kk}$ as long as the multiplicative majorization property (2.3) is satisfied. This flexibility may be used, for example, to ensure that the bit loading scheme in (3.30) is realizable. So, even though the linear transceiver with bit allocation (3.29) can achieve the same minimum power (3.28) as any optimal DFE transceiver, the bit allocation formula in the GTD-based DFE opens up more freedom.

We now make an interesting observation about the powers P_k in the optimal system. Substituting (3.24) into (3.7) and using the definition of c_k in (3.11) we find

$$P_k = \frac{2^{b_k} c_k}{[\mathbf{R}]_{kk}^2}. \tag{3.31}$$

Substituting from (3.30) it then follows that $P_k = 2^D$ for all k. Thus in the optimal system which has orthonormal columns for the precoder $\mathbf{F}$, the powers P_k are identical for all k. Since $P_{trans} = \sum_k P_k$ from (3.31) and(3.25), we therefore have $P_k = P_{trans}/M$ for all k.

In Chapter 2 we mentioned many examples of the GTD, such as SVD, Schur decomposition, GMD, and so on. Some of these have already appeared in the literature in different contexts. Each of

these serves as a specific realization of the optimal DFE transceiver acheiving minimum transmitted power, provided the bits are allocated as in Eq. (3.26). Each realization has a different choice of r_k ($= [\mathbf{R}]_{kk}$) satisfying the majorization condition (2.3), and in all cases, we restrict the precoder $\mathbf{F}$ to be the orthonormal choice (3.20). $\mathbf{G}$ is chosen as in (3.21), and $\mathbf{B}$ as in (3.23). We now elaborate on these different realizations arising from different GTD forms of $\mathbf{H} = \mathbf{QRP}^\dagger$.

1. *SVD Transceiver - the Linear Transceiver*. The singular value decomposition (SVD) of the channel matrix can be written as $\mathbf{H} = \mathbf{U\Sigma V}^\dagger$, where $\mathbf{U}$ and $\mathbf{V}$ are unitary and Σ is a diagonal matrix. Since $\mathbf{R} = \mathbf{\Sigma}$ is diagonal, the feedback matrix $\mathbf{B} = \mathbf{0}$ from (3.23), and the system reduces to a linear transceiver as in Fig. 3.3. This optimal solution for linear transceivers was proposed in [48].

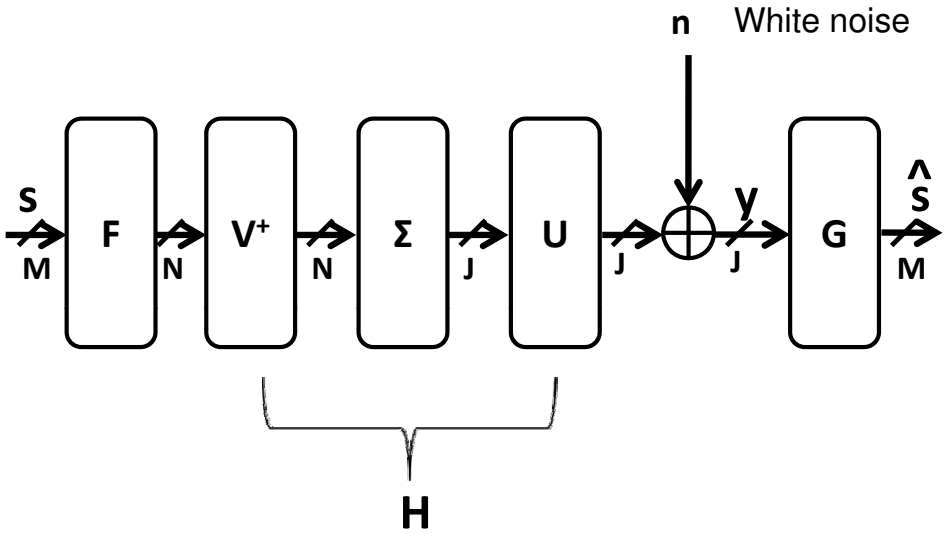


Figure 3.3: The SVD system, which represents a linear transceiver.

2. *GMD Transceiver*. The geometric mean decomposition (GMD) was introduced in [36]. The GMD of the channel $\mathbf{H}$ has the form $\mathbf{H} = \mathbf{QRP}^\dagger$, where $\mathbf{Q}$ and $\mathbf{P}$ have orthonormal columns, and $\mathbf{R}$ is an upper triangular matrix. Furthermore *the first M diagonal elements of $\mathbf{R}$ are identical*, and equal to the geometric mean of the M dominant channel singular values. For the case where the specified error probabilities $P_e(k)$ (hence c_k) are identical for all k, it follows from (3.26) that there is no need for bit allocation, that is, $b_k = b$ for all k. Unlike other special cases of the GTD such as the SVD, the question of b_k becoming unrealizable (i.e., taking noninteger or negative values) therefore does not arise.

3. *QR Transceiver - ZF-VBLAST System*. The QR decomposition of the channel matrix can be written as $\mathbf{H} = \mathbf{QR}$, where $\mathbf{Q}$ has orthonormal columns, and $\mathbf{R}$ is upper triangular. This yields a special case of the GTD transceiver, where the precoder is $\mathbf{F} = \begin{pmatrix} \mathbf{I}_M \\ \mathbf{0} \end{pmatrix}$, and can be implemented at no cost. See Fig. 3.4. This system leads to the ZF-VBLAST system, widely used in MIMO wireless communication [135].

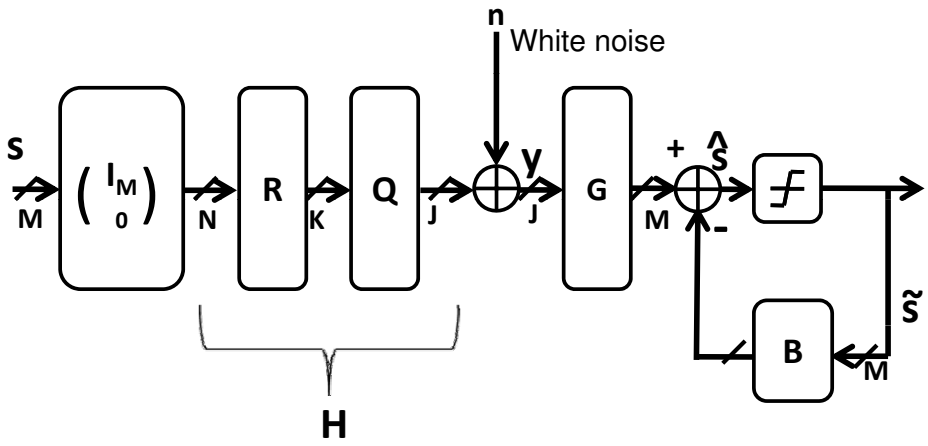


Figure 3.4: The QR transceiver, which has the lazy precoder. This is identical to the ZF-VBLAST system.

The optimal transceiver design usually assumes that $\mathbf{H}$ is known at the transmitter side. This assumption is not generally true. The more practical scheme would be the so called *limited feedback scheme*, in which the receiver uses a low rate feedback channel to tell the transmitter to use one of the precoders in a pre-determined codebook of precoders [56].

The QR based transceiver with bit loading is very suitable in limited feedback systems because the precoder matrix is identity, and only the bit loading vector $[b_1 \ \ldots \ b_M]$ needs to be known.[2] The receiver can compute $\{b_k\}$ from (3.15), quantize it to the bit loading vector nearest to the vectors in a predetermined codebook, and feed back the index of that vector to the transmitter. The design of this codebook is an interesting problem, but is beyond the scope of this thesis. Intuitively, this scheme will perform better than limited feedback schemes using the Grassmann codebook [91, 56], since the Grassmann codebook aims to cover the Grassmann manifold [92] while the bit loading codebook only tries to cover M-vectors with integer valued entries. This intuition is supported by Monte Carlo simulations in Sec. 3.2.6.

[2] In the scheme described in [14], the power allocation P_k also should be fed back, but in the GTD based optimal system $P_k = P_{trans}/M$ for all k as shown at the end of Sec. 3.2.3.

It is reassuring to know that since all GTD-based systems are optimal when the bit loading formula is realizable, this QR based special case has no loss of optimality even though it offers a simple precoder and a simple way to perform limited feedback.

4. *Bidiagonal Transceiver*. It is well-known [25] that any $J \times P$ matrix $\mathbf{H}$ can be factored as $\mathbf{H} = \mathbf{QRP}^\dagger$, where $\mathbf{Q}$ and $\mathbf{P}$ have orthonormal columns, and $\mathbf{R}$ has the bidiagonal form

$$\mathbf{R} = \begin{pmatrix} d_1 & f_1 & 0 & \cdots & 0 \\ 0 & d_2 & f_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & d_{P-1} & f_{P-1} \\ 0 & \cdots & \cdots & 0 & d_P \\ & & \mathbf{0} & & \end{pmatrix}.$$

With the channel represented in this bi-diagonal form, the feedback matrix given in (3.23) becomes

$$\mathbf{B} = \begin{pmatrix} 0 & f_1 & 0 & \cdots & 0 \\ 0 & 0 & f_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & f_{M-1} \\ 0 & \cdots & \cdots & 0 & 0 \end{pmatrix}.$$

Therefore the implementation of the DFE will be very simple since we need only to feed back *one* previous decision for detecting the current symbol. Also, the computation of the bidiagonal decomposition is inexpensive [25]. To the best of our knowledge, this kind of system has not previously been reported in transceiver literature.

Summarizing, any of the above four GTD-based systems achieves optimality. However, each one of them has some special features, which might be useful in different situations. Also, it is possible that other GTD-based systems exist with potential benefits in specific situations.

3.2.4 Other Transceiver Problems Solved by GTD-Based Transceiver

We now consider two variations of the transceiver optimization problem. The first problem we consider is the quality of service problem, in which we want to minimized the transmitted power

subjects to the individual BER and bit rate constraints in each subchannel. The second problem we consider is the bit rate maximization problem, in which we maximize the bit rate subject to the transmitted power and BER constraints. We will show that both of these two problems have solutions based on the GTD.

Quality of Service (QoS) Problem

The quality of service problem in MIMO communication has been considered by a number of authors [77, 39, 28]. In these papers the QoS is defined in the output SINR sense, and furthermore there is no bit allocation. In fact, reference [28] addresses a special case of the problem discussed in [39], namely the case where the channel $\mathbf{H} = \mathbf{I}$. Here we consider a different situation where the error probability $P_e(k)$ (equivalently the constants c_k in Eq. (3.11)) and bit rate b_k of each substream are specified to be the QoS parameters. We will show that under some multiplicative majorization condition, we can customize the GTD-based transceiver to obtain an optimal solution that minimizes power subject to the QoS specifications $\{c_k, b_k\}$. More precisely, the problem considered here is

$$\begin{aligned} \min_{\mathbf{F},\mathbf{G},\mathbf{B}} \quad & P_{trans} \\ \text{s.t.} \quad (a) \quad & \mathbf{GHF} = \mathbf{I} + \mathbf{B} \\ (b) \quad & \{c_k, b_k\} \text{ fixed (QoS for data stream } k). \end{aligned} \tag{3.32}$$

Having the GTD concept in mind, we know that if we are able to find a matrix $\mathbf{F}_0$ such that

$$\{c_k 2^{b_k}\}_{k=1}^{M} \prec_{\times} \{\sigma_k(\mathbf{HF}_0)\}_{k=1}^{M},$$

we are able to find some semi-unitary matrix $\mathbf{Q}$, such that if the precoder is chosen as $\mathbf{F} = \mathbf{F}_0\mathbf{Q}$, the QoS constraint is satisfied exactly. With such precoder, the transmitted power is proportional to $\text{Tr}(\mathbf{FF}^\dagger) = \text{Tr}(\mathbf{F}_0\mathbf{QQ}^\dagger\mathbf{F}_0^\dagger) = \text{Tr}(\mathbf{F}_0\mathbf{F}_0^\dagger)$. Therefore, we can transform the QoS problem (3.32) in the following form:

$$\begin{aligned} \min_{\mathbf{F}_0} \quad & \text{Tr}(\mathbf{F}_0\mathbf{F}_0^\dagger) \\ \text{s.t.} \quad & \{c_k 2^{b_k}\}_{k=1}^{M} \prec_{\times} \{\sigma_k^2(\mathbf{HF}_0)\}_{k=1}^{M} \end{aligned} \tag{3.33}$$

Let ϕ_k denote the square of the k-th largest singular value of $\mathbf{F}_0$. By Theorem H.1 of [65], we can further transform problem (3.33) to be

$$\begin{aligned} \min_{\phi_k} \quad & \sum_{k=1}^{M} \phi_k \\ \text{s.t.} \quad & \{c_k 2^{b_k}\}_{k=1}^{M} \prec_{\times} \{\sigma_{h,k}^2 \phi_k\}_{k=1}^{M} \end{aligned} \tag{3.34}$$

To solve problem (3.34), we can further perform some parameter transformations. Let $\alpha_k = \ln \phi_k$, and $\beta_k = \ln \frac{c_k 2_k^b}{\sigma_{h,k}^2}$. Substituting these into (3.34), we obtain

$$\begin{aligned} \min_{\alpha_k} \quad & \sum_{k=1}^{M} e^{\alpha_k} \\ \text{s.t.} \quad & \beta_1 \leq \alpha_1 \\ & \beta_1 + \beta_2 \leq \alpha_1 + \alpha_2 \\ & \quad \vdots \\ & \sum_{k=1}^{M} \beta_k = \sum_{k=1}^{M} \alpha_k. \end{aligned} \tag{3.35}$$

The objective function in problem (3.35) is convex in α_k, and the constraints are all affine, thus it is a convex optimization problem and can be numerically solved efficiently. However, in the following we provide a theorem based on the majorization theory to obtain the solution without performing numeric convex optimization programs under some conditions.

Theorem 3.2.4 *For the QoS problem (3.32), the following are true: (a) The minimum required power to achieve the specification will be at least as large as*

$$P_{min} = c 2^b \left(\frac{1}{\prod_{k=1}^{M} \sigma_{h,k}^2} \right)^{\frac{1}{M}},$$

where $c = M(\prod_{k=1}^{M} c_k)^{\frac{1}{M}}$ *and* $b = \sum_{k=1}^{M} b_k / M$.

(b) This P_{min} *is achievable if*

$$\frac{\{c_1 2^{b_1}, \cdots, c_M 2^{b_M}\}}{c 2^b / M} \prec_{\times} \frac{\{\sigma_{h,1}^2, \cdots, \sigma_{h,M}^2\}}{(\prod_{k=1}^{M} \sigma_{h,k}^2)^{\frac{1}{M}}}, \tag{3.36}$$

that is, if the vector on the left which is determined by the QoS constraints, is multiplicatively majorized by

the vector on the right which is determined by the channel. ♢

Proof: See Appendix. □

This system, which achieves $P_{trans} = P_{min}$ under condition (3.36), will be referred to as the custom GTD-based system, since the value of the precoder and equalizer are not computed solely depending on $\mathbf{H}$, but also depending on the given QoS $\{c_k, b_k\}$. This example shows that the GTD-based system has much more flexibility than the linear transceiver system.

It should be pointed out here that when the QoS specification $\{c_k, b_k\}$ is identical for all k, the custom GTD reduces to the GMD. This is because the multiplicative majorization relation (3.36) always holds in this case.

Bit Rate Maximization Problem

The bit rate maximization problem subject to transmitted power constraint is the dual of the problem described in Eq. (3.12). It will be shown that the GTD transceiver gives the optimal solution. For the special case of linear transceivers this problem was considered in [46]. Consider again the system with the zero-forcing condition. Under the high bit rate assumption, b_k can be rearranged as

$$b_k \approx \log_2 \left(\frac{P_k}{\sigma_{ek}^2 d_k} \right), \tag{3.37}$$

where d_k represents the bit error rate via (3.9). Therefore, the problem of maximizing the average bit rate for fixed set of bit error rates $\{d_k\}$ and total power can be written as

$$\begin{aligned} \max_{\mathbf{F},\mathbf{G},\mathbf{B},\{P_k\}} \quad & b = \frac{1}{M} \sum_{k=1}^{M} \log_2 \left(\frac{P_k}{\sigma_{ek}^2 d_k} \right) \\ \text{s.t.} \quad (a) \quad & \text{Tr}(\mathbf{F}\mathbf{\Lambda}_s\mathbf{F}^\dagger) \leq P_{total} \\ (b) \quad & \mathbf{GHF} = \mathbf{I} + \mathbf{B} \text{ (zero-forcing)}, \end{aligned} \tag{3.38}$$

where $\mathbf{\Lambda}_s = \text{diag}(P_1, P_2, \cdots, P_M)$. The power constraint can be rewritten as

$$\sum_{k=1}^{M} P_k [\mathbf{F}^\dagger \mathbf{F}]_{kk} \leq P_{total}.$$

We solve the above optimization problem in two stages. First we find the optimal power P_k for given $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{B}$, under the power constraint. We then derive the optimal transceiver matrices. Suppose $\{P_k^*\}$ are optimal for problem (3.38), then the KKT condition [8] states that there exists α

such that

$$\alpha \leq 0, \tag{3.39}$$

$$\frac{\partial}{\partial P_k} \left(\frac{1}{M} \sum_{k=1}^{M} \log_2(\frac{P_k}{\sigma_{ek}^2 d_k}) + \alpha\Big(\sum_{k=1}^{M} P_k [\mathbf{F}^\dagger \mathbf{F}]_{kk} - P_{total}\Big)\right)\Bigg|_{P_k = P_k^*} = 0,$$

and

$$\alpha\Big(\sum_{k=1}^{M} P_k [\mathbf{F}^\dagger \mathbf{F}]_{kk} - P_{total}\Big)\Big|_{P_k = P_k^*} = 0. \tag{3.40}$$

By solving these equations, we get $\alpha = \frac{-M}{P_{total} \log_e 2}$, and the optimal power allocation

$$P_k^* = \frac{P_{total}}{M[\mathbf{F}^\dagger \mathbf{F}]_{kk}}. \tag{3.41}$$

Observe that when the triplet $\{\mathbf{F}, \mathbf{G}, \mathbf{B}\}$ is fixed, Eq. (3.38) is concave in the vector $[P_1 \ ... \ P_M]$. So the preceding solution represents a maximum (rather than minimum) of Eq. (3.38). The derivation of (3.41) is similar to the one in [46] for linear transceivers. Using (3.41) in (3.38) and simplifying, we have

$$b \quad = \quad \log_2 \left(\prod_{k=1}^{M} \frac{P_{total}}{M c_k [\mathbf{F}^\dagger \mathbf{F}]_{kk} [\mathbf{G}\mathbf{G}^\dagger]_{kk}} \right)^{\frac{1}{M}} \tag{3.42}$$

Thus, the problem of maximizing the bit rate is reduced to maximizing (3.42) subject to zero forcing. But maximizing (3.42) is equivalent to minimizing (3.16). The latter minimization can be achieved with the GTD and results in $P_{trans} = P_{min}$ given by (3.18). So it follows that the optimal solution is such that

$$\prod_{k=1}^{M} [\mathbf{F}^\dagger \mathbf{F}]_{kk} [\mathbf{G}\mathbf{G}^\dagger]_{kk} = 1 / \prod_{k=1}^{M} \sigma_{h,k}^2 \tag{3.43}$$

Substituting this into (3.42) the maximized bit rate becomes:

$$b_{max} = \log_2 \left(\frac{P_{total}}{c} (\prod_{k=1}^{M} \sigma_{h,k}^2)^{\frac{1}{M}} \right) \tag{3.44}$$

This is exactly the maximum bit rate that has been achieved with linear transceivers, as shown in

[46]. Thus, whenever bit allocation is permitted, the DFE transceiver offers no advantage over the linear transceiver, as far as maximizing the bit rate is concerned.

For completeness recall that the GTD based optimal solution has matrices $\mathbf{F} = [\mathbf{P}]_{P\times M}$, $\mathbf{G}_0 = [\mathbf{Q}^\dagger]_{M\times J}$, $\mathbf{G}$ as in (3.21), and $\mathbf{B}$ as in (3.23). Since this achieves the maximum bit rate, all the special cases discussed in Sec. 3.2.3 maximize bit rate. Jiang *et. al.* considered a different problem in [36] where they showed that the GMD system achieves maximum channel throughput (defined in terms of mutual information) with uniform bit allocation, for the case of large SNR. This result is consistent with our result in this section for the actual bit rate, which holds for any GTD.

3.2.5 Simulation Results with Perfect CSI

Here we present simulations for the case where the channel is known to the transmitter and the receiver. We consider a number of methods in the comparison. These include the linear transceiver based on SVD, and DFE-based transceivers based on GMD, QR decomposition, and bi-diagonal decomposition BID. For these methods, whenever the bit loading formula (3.15) is not realizable due to finite constellation granularity, we replace it with the optimal bit loading algorithm in [9], and take the precoder and equalizer matrices to be the optimum ones determined by the theory.

In addition, we introduce a new procedure that allows us to achieve the minimum power P_{min} of Eq. (3.18) with integer bit allocation for the special case of equal c_k (equal error probabilities for all k). This procedure will be denoted as the GB method (generalized bit allocation method) in all simulations. It exploits the freedom offered by the GTD in choosing the diagonal elements $\mathbf{R}_{kk}$ of the lower triangular matrix $\mathbf{R}$. Since the method is somewhat involved we first describe it briefly before proceeding with the simulation examples.

Assume c_k is identical for all k. The method that we refer to as generalized bit allocation (GB) proceeds as follows. First we compute b_k using (3.29), and truncate it to the nearest even integer to get a square QAM constellation (replacing b_k with zero if it turns out to be negative). We then check if the bit rate constraint $\sum_k b_k = Mb$ is satisfied with equality. If this is not the case, then we adjust b_k in one of two possible ways depending on the situation. For convenience assume b_k is renumbered such that $b_k \geq b_{k+1}$.

1. If $\sum_k b_k < Mb$ we replace b_M with $b_M + 2$ until either $\sum_k b_k = Mb$ or $b_{M-1} = b_M$. In the former event we stop. If the latter is true but $\sum_k b_k < Mb$ still prevails, we replace b_{M-1} with $b_{M-1} + 2$, and continue the process. If we reach a point where $b_{M-n-1} > b_{M-n} =$

$\ldots = b_{M-1} = b_M$ for some $n > 1$, with $\sum_k b_k < Mb$ still prevailing, then we replace b_{M-n} with $b_{M-n} + 2$. Repeated application of this procedure leads to a bit allocation that satisfies $\sum_k b_k = Mb$.

2. If $\sum_k b_k > Mb$ we modify the preceding in the obvious way: We replace b_1 with $b_1 - 2$ until either $\sum_k b_k = Mb$ or $b_1 = b_2$. In the former event we stop. If the latter event is true but $\sum_k b_k > Mb$ still prevails, we replace b_2 with $b_2 - 2$, and continue the process. If we reach a point where $b_1 = b_2 = \ldots = b_n > b_{n+1}$ with $\sum_k b_k > Mb$ still prevailing, then we replace b_n with $b_n - 2$. Repeated application of this procedure leads to a bit allocation that satisfies $\sum_k b_k = Mb$.

Let $\{b_1^i, b_2^i \ldots b_M^i\}$ denote the final bit allocation resulting from this algorithm (superscript i is for "integer") and let $\{b_1, b_2 \ldots b_M\}$ denote the initial allocation from (3.29). We have $\sum_k b_k = \sum_k b_k^i$ by construction. Furthermore, if $\{\sigma_{h,k}\}$ has a wide distribution, then the final bit allocation satisfies

$$\begin{bmatrix} b_1^i & b_2^i & \ldots & b_M^i \end{bmatrix} \prec_+ \begin{bmatrix} b_1 & b_2 & \ldots & b_M \end{bmatrix}. \tag{3.45}$$

The notation $\prec_+$ means that the vector on the left is additively majorized by that on the right [65, 32]. The next step depends upon whether this happens or not. Suppose (3.45) indeed holds (which is often the case as seen through simulations). If $c_i = c/M$ for all i then by using (3.29) we verify that this is equivalent to the multiplicative majorization condition (3.36). Now, with $[\mathbf{R}]_{kk}$ defined as in (3.30) or more precisely

$$b_k^i = D - \log_2 c_k - \log_2([\mathbf{R}]_{kk}^2), \tag{3.46}$$

Eqn. (3.36) (hence (3.45)) is equivalent to the condition (2.3) demanded by the existence of the specific GTD. This means that there exists a GTD for the channel $\mathbf{H}$ such that both (3.36) and the integer bit allocation (3.46) hold simultaneously.

According to Theorem 3.2.3, this design therefore achieves minimum power while at the same time satisfies the integer bit rate constraint for the case where $c_i = c/M$ for all i. This is precisely the beauty of the GTD. We have successfully exploited the flexibility in bit allocation offered by the freedom to choose the diagonal elements $[\mathbf{R}]_{kk}$ in the GTD.

There remains one more case to be considered, namely the situation where the majorization relation (3.45) does not hold. In this case we have observed that the SVD transceiver (linear transceiver)

with integer bit allocation (3.46) typically yields a smaller BER than all the other GTD methods. So we simply use the SVD system whenever the second situation prevails.

In the following we assume $M = N = 4$ and $J = 5$. So the channel matrix $\mathbf{H}$ is of size 5×4; each of its entries $[\mathbf{H}]_{km}$ is drawn from a iid Gaussian distribution with zero mean and unit variance. For each realization of this random $\mathbf{H}$ we compute the BER, and average it over 1000 such realizations. The additive noise is complex circular Gaussian with average power normalized to 0 dB. Gray encoded bits are adopted. The results are given in terms of bit error rate versus transmitted power. Here we compare the uncoded bit error rate. Since in all our designs the MSE matrix is diagonal, this makes the overall systems act like a set of parallel AWGN channels. Channel coding may be further added to provide coding gain independent of the transceiver designs discussed in the thesis. Decision feedback is operative in all the systems being compared, except in the special case of the SVD system.

Example 1: High bit rate case: In this example we consider GTD transceivers with bit allocation approximating (3.26). We assume $c_k = c/M$ (identical error probabilities $P_e(k)$) for all k. The GTD system minimizes the required power to the value P_{min} given in (3.18). Fig. 3.5 shows the simulated BER plots for the case where $\sum_k b_k = 40$, that is, there are 40 bits to be allocated into the four signal substreams. It can be observed that all systems perform about the same. This is consistent with Theorems 1 and 2 under the assumption of high bit rate. Notice in particular that the SVD system without DFE is almost as good as the systems with DFE. For the GB method, integer bit allocation is handled as described in this subsection. For all other methods, whenever the bit loading formula (3.15) is not realizable due to finite constellation granularity, we replace it with the optimal bit loading algorithm in [9], and take the precoder and equalizer matrices to be the optimum ones determined by the theory. Forcing b_k to be integers usually results in $P_e(k)$ being only approximately equal; the plots are based on BER values averaged over all k. As explained at the end of Sec. 3.2.3, the powers P_k are identical for all k.

Example 2: Low bit rate case: For the methods compared above, the theory in this section predicts identical performance under the "high bit rate" assumption. This was essentially confirmed in the preceding example. In the present example we will see that the performances are quite different from each other in the low bit rate case. This example is similar to Example 1 with the difference that $\sum_k b_k = 14$. Fig. 3.6 shows the BER plots. In this case, oftentimes the SVD will drop the substreams for which the corresponding singular values are too small (by not allocating any bits

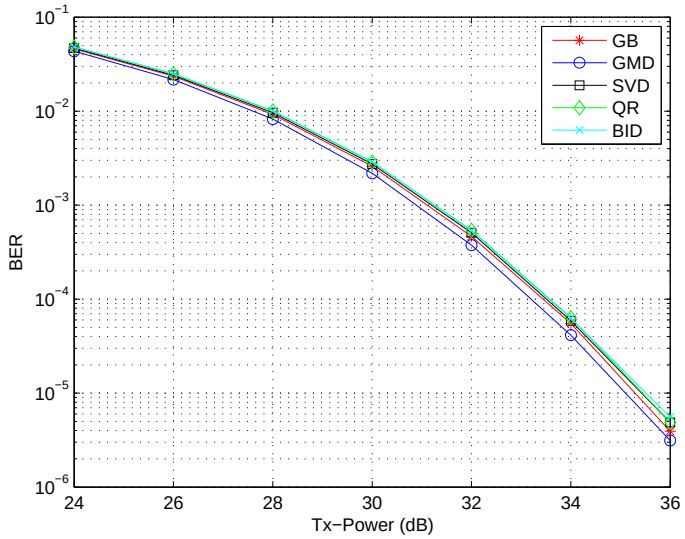


Figure 3.5: Example 1. BER versus Tx-Power for $\sum_k b_k = 32$.

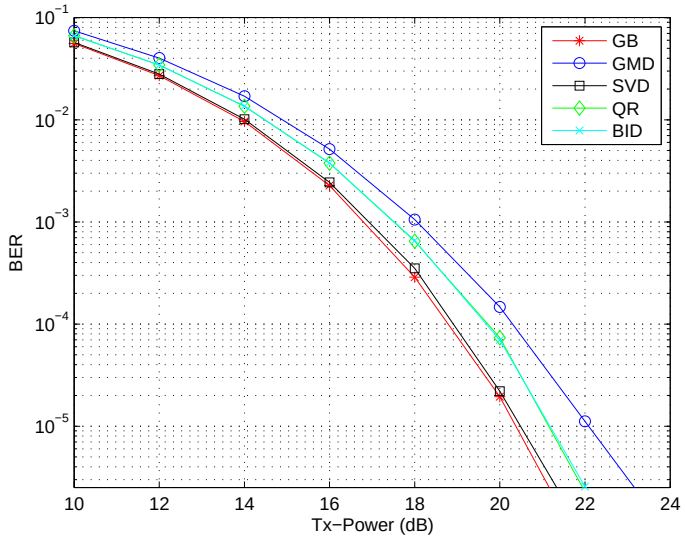


Figure 3.6: Example 2. BER versus Tx-Power for $\sum_k b_k = 14$.

for them). However, the GMD system will never drop any substream; instead it will force each of the substreams to have equal error variance and allocate about the same number of bits. If a substream is very bad (noisy), this strategy will seriously degrade the performance. But the SVD system simply drops the bad sub-channels, therefore retaining good performance. Note that this behavior of GMD is due to the zero-forcing constraint enforced throughout the section. For the MMSE receiver without the zero-forcing constraint, this effect may disappear. For the "GB" method we drop the bad substreams as in SVD. This is why both the "GB" and the SVD systems outperform other methods when there are some very bad sub-channels. Note that this effect is not so noticeable in the high bit rate case. Also the "GB" method does not have the non-integer bit allocation problem that all other methods suffer from (unless (3.45) fails in which case we replace it with SVD as explained before). This is why our GB method performs the best among all the systems.

Example 3. Fixed, identical constellations: In this example we fix $b_k = 6$ bits for each k (64-QAM streams), and all c_k (i.e., error probabilities $P_e(k)$) are identical. The term "custom" stands for the custom-GTD system with $[\mathbf{R}]_{kk}$ obtained from (3.30). In this example since $P_e(k)$ and b_k are identical for all k, the custom GTD system reduces to the "GMD" system, which is known to be optimal in terms of BER [36]. Fig. 3.7 shows the performances of various GTD systems. Clearly GMD and custom GTD outperform other GTDs.

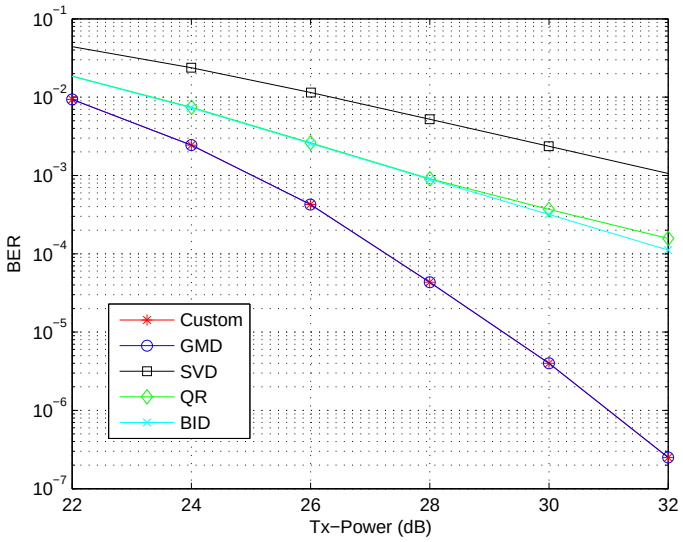


Figure 3.7: Example 3. BER versus Tx-Power when $b_k = 6$ for all k.

Example 4. Fixed, non-identical constellations: This is similar to Ex. 3 with the difference that the fixed constellations have non-identical bits: $[8,\ 8,\ 6,\ 6]$ (i.e., 256-QAM, 256-QAM, 64-QAM, and 64-QAM). Fig. 3.8 shows the BER plots. Again, "custom" denotes the custom GTD system with $[\mathbf{R}]_{kk}$ obtained from (3.30), and so it has minimum power for fixed BER. It can be observed from the plots that the custom GTD significantly outperforms all other methods including the GMD. This clearly demonstrates the advantage offered by the flexibility of the GTD. However, among the other four methods, there is no theory as to which one performs better.

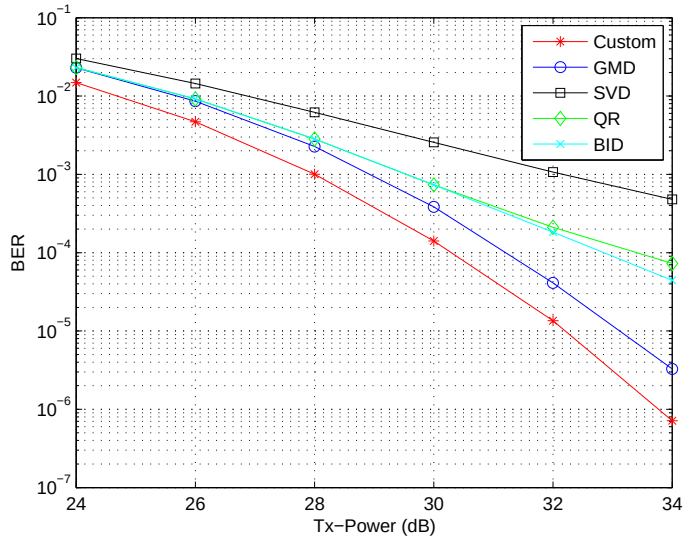


Figure 3.8: Example 4. BER versus Tx-Power when bit vector is fixed as $[8,\ 8,\ 6,\ 6]$.

3.2.6 Simulation Results with Limited Feedback

We now consider the limited feedback scheme. As in earlier sections, zero-forcing is assumed, and c_k (equivalently error probabilities $P_e(k)$) are identical for all k. We assume $M = 4$, and $N = J = 5$ so that the 5×4 orthonormal precoders in the Grassman codebook published in [57], [92] can be used. It is assumed that feedback from the receiver to the transmitter is error free. As in the previous section, each of the channel entries $[\mathbf{H}]_{km}$ is drawn from an iid Gaussian distribution with zero mean and unit variance. For each realization of this random $\mathbf{H}$ we compute the BER, and average it over 1000 such realizations. The schemes considered are as follows:

1. The scheme proposed in [56] based on the so-called projection 2-norm criterion. This is a linear transceiver with an orthonormal precoder, with no bit allocation or power allocation.

It uses a 8-bit Grassmann codebook [57, 92] to represent a set of 256 precoder matrices. The receiver feeds back the 8 bits to the transmitter to tell which precoder ought to be used. This system is referred to as "Lin-limited-FB."

2. The minimun-BER DFE design proposed in [91] which uses a 8-bit Grassmann codebook in conjunction with GMD. This is referred to as "GMD-limited-FB." It has $b_k = b$ for all k.

3. The QR based DFE design. The precoder is identity, and the receiver feeds back only bit loading information $\{b_k\}$ as described in Sec. 3.2.3. This will be referred to as "QR-limited-FB."

4. We also show the BER plots for the optimal DFE system based on GMD with perfect CSI at the transmitter. This ideal system has the smallest BER, which is shown for reference. This system is referred to as "GMD-perfect-CSI."

Like the first method, the last three methods also have identical powers P_k for all k, but for a different reason as described at the end of Sec. 3.2.3. We present BER plots for two cases: the case where $\sum_k b_k = 32$ (Fig. 3.9) and where $\sum_k b_k = 24$ (Fig. 3.10). From the plots we see that the proposed "QR-limited-FB" scheme performs significantly better than the state-of-the-art limited feedback schemes [56, 91], and comes close to the optimal "GMD-perfect-CSI" scheme. Note that the Grassmann codebook aims to cover the Grassmann manifold of orthonormal precoder matrices [92, 5] while the bit loading codebook in the "QR-limited-FB" scheme only has to cover integer valued vectors $\begin{bmatrix} b_1 & b_2 & \dots & b_M \end{bmatrix}$.

We now discuss some details about the "QR-limited-FB" scheme. As described before, the codebook here is a set of integer vectors which specifies to the transmitter what b_k are. After the receiver calculates b_k from (3.15), it quantizes the vector $\begin{bmatrix} b_1 & b_2 & \dots & b_M \end{bmatrix}$ to the nearest vector in the codebook. In the simulation we also restrict the codebook to have vectors with each b_k no less than 4 for Fig. 3.9 (and 2 for Fig. 3.10). Also, we use square QAM, so the possible number of bits in each substream will be even. The size of the codebook is therefore $(11 \times 10 \times 9)/(3 \times 2) = 165$. This requires less than 8 bits of feedback from receiver to transmitter. Even with such limited feed back, the proposed "QR-limited-FB" scheme performs very well indeed.

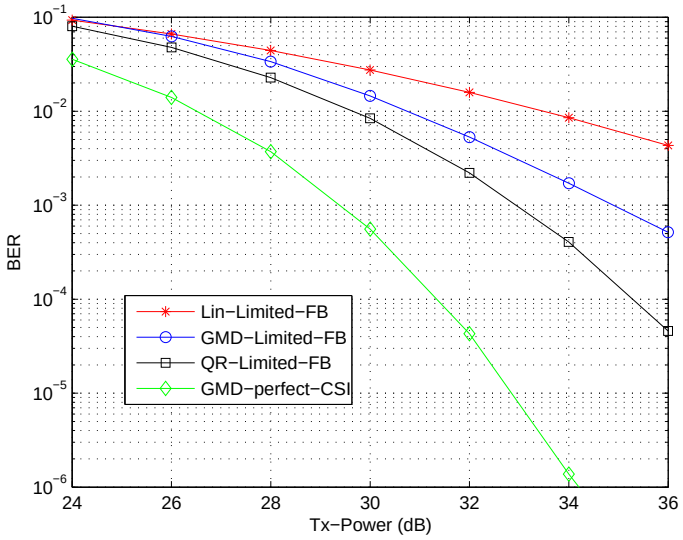


Figure 3.9: BER versus Tx-Power with limited feedback (8 feedback bits per block, and 32 bits transmitted per block).

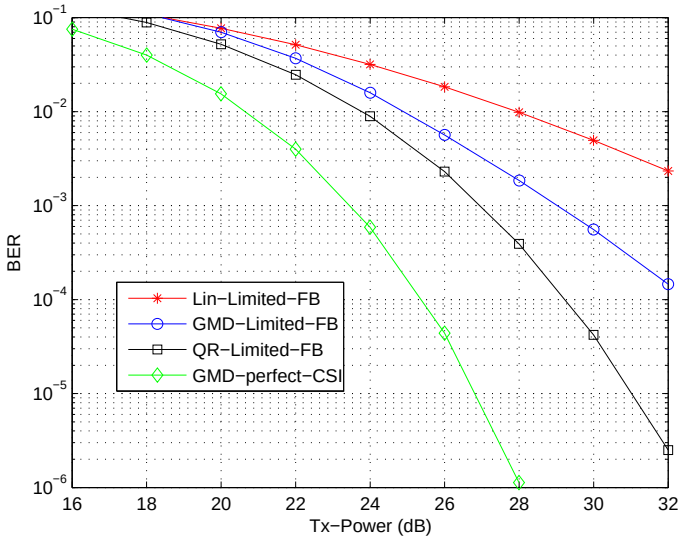


Figure 3.10: BER versus Tx-Power with limited feedback (8 feedback bits per block, and 24 bits transmitted per block).

3.2.7 Concluding Remarks

We have presented a method for the joint optimization of the matrices $\{\mathbf{F}, \mathbf{G}, \mathbf{B}\}$ and the bits $\{b_k\}$ in a transceiver with DFE. It was formally shown that when the bit allocation, precoder, and equalizer are jointly optimized, linear transceivers and transceivers with DFE have identical performance in the sense that the transmitted power is identical for a given bit rate and error probability. We also proved that any GTD-based system achieves the optimal performance. The GTD family also yields optimum solutions for the QoS problem and the bit rate maximization problem. Many existing systems are identified to be special cases of the GTD-based system, and some new GTD-based transceivers were also indicated. The QR-based GTD has the advantage of offering a simple way to do limited feedback by sending the bit allocation information from the receiver to transmitter.

3.3 MIMO Transceivers with Linear Constraints on Transmit Covariance Matrix

In this section we revisit the optimization of multiple-input multiple-output (MIMO) communication systems. Instead of only the total power constraint, in this section we also consider the more realistic per-antenna power constraints on the transmitter. In this MIMO system, the transmitter has M antennas sending independent information to the receiver equipped with N antennas. The signal vector consisting of M substreams is assumed to be linearly transformed by the channel matrix $\mathbf{H}$, and corrupted by the additive Gaussian noise.

3.3.1 Signal Model and Problem Formulation

The transceiver model is similar to the case discussed in Sec. 3.2 and can be represented as in Fig. 3.11. $\mathbf{R}_n$ is the covariance matrix of the additive Gaussian noise; $\mathbf{H}$ is the channel matrix; $\mathbf{F}$ is the precoder; $\mathbf{G}$ is the receiving filter; $\mathbf{B}$ is $\mathbf{0}$ for linear transceiver case, and strictly lower triangular [112] for the system with linear precoding and DFE. The difference between the current model and the one considered earlier is that the channel matrix $\mathbf{H}$ here is a $M \times M$ square matrix. This constraint will later be relaxed and discussed. The per-antenna power constraints can be formulated as

$$(E[\mathbf{F}\mathbf{s}\mathbf{s}^\dagger\mathbf{F}^\dagger])_{ii} = (\mathbf{F}\mathbf{F}^\dagger)_{ii} \leq P_i,\ \forall i = 1, 2, \cdots, M \tag{3.47}$$

The total power constraint can be written as

$$\mathrm{Tr}(\mathbf{F}\mathbf{F}^{\dagger}) \leq P_{\text{total}}. \tag{3.48}$$

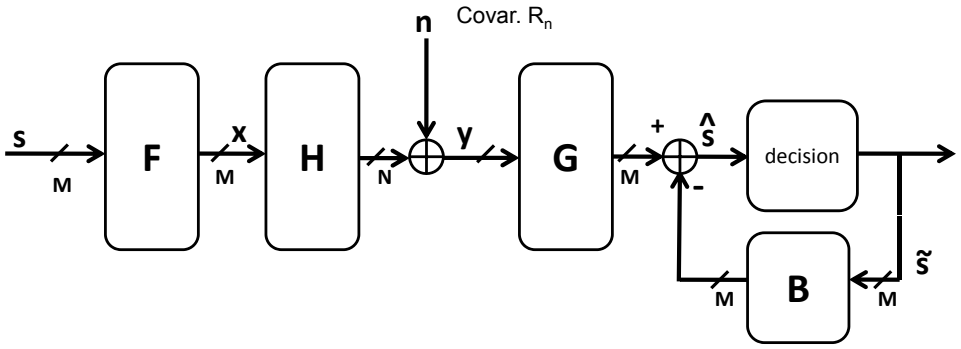


Figure 3.11: The system with linear precoding and DFE.

Here we want to minimize the uncoded bit error rate subject to the individual and total power constraints. This problem formulation is slightly different from those considered by previous papers [90, 73, 45], which considered only either the total power constraint or the individual power constraints. It will be elaborated later that because of the introduction of the constraints (3.47), the nice quasi-analytic form of the optimal solutions obtained in [73] is gone. However, since the convexity of the problem still holds true, the optimal solution can be obtained by numerical search as we show in the following sections. In this section, we will establish the link between the optimal transceiver design problems and the SDPs. Also, the majorization theory is used to further simplify the considered problems.

It can be observed that the optimal receiver structures do not change due to the individual power constraints when $\mathbf{F}$ is given in both cases. In the following we briefly review the optimal receiver structures [73, 90, 75]. In the linear transceiver case, it turns out that with no loss of optimality [75] we can use the Wiener filter to be the receiver, i.e.,

$$\mathbf{G}_{opt} = \mathbf{F}^{\dagger}\mathbf{H}^{\dagger}(\mathbf{H}\mathbf{F}\mathbf{F}^{\dagger}\mathbf{H}^{\dagger} + \mathbf{R}_n)^{-1}.$$

If the Wiener filter $\mathbf{G}_{opt}$ is used, the resulting error covariance matrix (MSE matrix) can be written as

$$\mathbf{E} := E[\mathbf{e}\mathbf{e}^{\dagger}] = \mathbf{I} - \mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{W}\mathbf{H}\mathbf{F},$$

where

$$\mathbf{W} = (\mathbf{HFF}^{\dagger}\mathbf{H}^{\dagger} + \mathbf{R}_n)^{-1}.$$

This can be rewritten in the following form by using matrix inversion lemma [73]:

$$\mathbf{E} = (\mathbf{I} + \mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}_n^{-1}\mathbf{HF})^{-1}.$$

Note that $\mathbf{G}_{opt}$ is both optimal in the sense of maximizing SINR in each substream as well as minimizing the mean square error. In this case, the SINR can be related to the MSE as [73]:

$$\text{SINR}_i = \frac{1}{\text{MSE}_i} - 1. \tag{3.49}$$

The optimum decision feedback equalization with successive decoding for MIMO channels is considered in [112]. First, the feedforward filter is $\mathbf{G}_{opt} = \mathbf{CF}^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}_y^{-1}$, and the resulting MSE matrix can be written as

$$\mathbf{E} := E[\mathbf{ee}^{\dagger}] = \mathbf{C}(\mathbf{I} + \mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}_n^{-1}\mathbf{HF})^{-1}\mathbf{C}^{\dagger} = \mathbf{CMC}^{\dagger},$$

where $\mathbf{M}$ is defined as

$$\mathbf{M} := (\mathbf{I} + \mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}_n^{-1}\mathbf{HF})^{-1},$$

and $\mathbf{C} := \mathbf{I}+\mathbf{B}$. It can also be shown that the optimal $\mathbf{C}$ can be chosen as [90] $\mathbf{C} = \text{diag}([\mathbf{L}_{11}, \cdots, \mathbf{L}_{MM}]^T)\mathbf{L}^{-1}$, where $\mathbf{L}$ is the lower triangular Cholesky factor of $\mathbf{M}$, i.e., $\mathbf{M} = \mathbf{LL}^{\dagger}$. The resulting MSE matrix will be $\mathbf{E} = \text{diag}([\mathbf{L}_{11}^2, \cdots, \mathbf{L}_{MM}^2]^T)$. Under these choices, the SINR and MSE in each substream also have a nice relation as in (3.49) as shown in [112].

3.3.2 Linear Transceivers

In this section we will focus on solving the problem of minimizing BER subject to individual and total power constraints for the linear transceiver case. We use the two-step approach. In the first step we will minimize the AM-MSE (arithmetic mean of mean square error) of the system. This is done by reformulating the problem as a semi-definite program (SDP) as we shall see. In the second step, we will argue that there is a set of systems in the minimum AM-MSE family that minimizes

the average error probability among all linear transceivers $\{\mathbf{F}, \mathbf{G}\}$. That is, the average BER is the smallest possible. An approach to find one of such optimal transceivers will also be given.

The minimum AM-MSE problem with per-antenna and total power constraints can be cast as the minimization of $\text{Tr}(\mathbf{E})$, where $\mathbf{E}$ is the MSE matrix as discussed before. In the following we will adopt the trick in [59] to formulate the current problem to be a SDP. By similar derivation as in [59] we have

$$\text{Tr}(\mathbf{E}) = M - N + \text{Tr}(\mathbf{W}\mathbf{R}_n).$$

Since M and N are constants and $\mathbf{R}_n$ is known, the AM-MSE depends only on $\mathbf{W}$, which is a function of $\mathbf{F}$. Furthermore, if we define $\mathbf{U} := \mathbf{F}\mathbf{F}^\dagger$, we can write $\mathbf{W} = (\mathbf{H}\mathbf{U}\mathbf{H}^\dagger + \mathbf{R}_n)^{-1}$.

This equation can be replaced with $\mathbf{W}_0 \succeq (\mathbf{H}\mathbf{U}\mathbf{H}^\dagger + \mathbf{R}_n)^{-1}$ (as discussed in [59]). Also it holds true if and only if the following linear matrix inequality holds (p. 472 in [31])

$$\begin{pmatrix} \mathbf{H}\mathbf{U}^\dagger\mathbf{H}^\dagger + \mathbf{R}_n & \mathbf{I} \\ \mathbf{I} & \mathbf{W}_0 \end{pmatrix} \succeq \mathbf{0}. \tag{3.50}$$

Therefore, the final form of problem formulation can be written as

$$\begin{aligned} \min_{\mathbf{U},\mathbf{W}_0} \quad & \text{Tr}(\mathbf{W}_0\mathbf{R}_n) \\ \text{s.t.} \quad & \text{(a)} \quad (\mathbf{U})_{ii} \le P_i, \ \ \forall i = 1, 2, \cdots, M \\ & \text{(b)} \quad \text{Tr}(\mathbf{U}) \le P_{\text{total}} \\ & \text{(c)} \quad \mathbf{U} \succeq \mathbf{0} \\ & \text{(d)} \quad \begin{pmatrix} \mathbf{H}\mathbf{U}\mathbf{H}^\dagger + \mathbf{R}_n & \mathbf{I} \\ \mathbf{I} & \mathbf{W}_0 \end{pmatrix} \succeq \mathbf{0}. \end{aligned} \tag{3.51}$$

In (3.51) the objective function is linear, and the constraints are either linear or positive semi-definite. Therefore, the problem (3.51) is an SDP problem [105]. This ensures that the global minimum of (3.51) can be found in polynomial time, when the precision of the solution is specified.

Now consider any given precoder $\mathbf{F}$, where a unitary matrix $\mathbf{\Psi}$ is further inserted in front of $\mathbf{F}$. We notice that this substitution does not change the individual power in each antenna nor the AM-MSE [10]. In the high SNR region, the average BER is an increasing Schur-convex function [73] in the vector diag($\mathbf{E}$). Therefore we have

$$P_{br} = \frac{1}{M}\sum_{i=1}^{M} \alpha \mathrm{Q}\left(\sqrt{\beta \cdot (\frac{1}{\mathbf{E}_{ii}} - 1)}\right) \geq \alpha \mathrm{Q}\left(\sqrt{\beta \cdot (\frac{1}{\frac{1}{M}\mathrm{Tr}(\mathbf{E})} - 1)}\right), \tag{3.52}$$

where α and β are constants depending on the QAM constellation. It is now clear that the lower bound is minimized by minimum AM-MSE. The equality is achieved by choosing the matrix $\mathbf{\Psi}$ to equalize the MSE in each substream.

Now we provide an approach to obtain one of the optimal minimum BER solutions. Taking any solution of $\mathbf{U}_{AM}$ to the problem (3.51), the optimal minimum AM-MSE solution $\mathbf{F}_{AM}$ can be taken as any Cholesky factor of $\mathbf{U}_{AM}$. Let $\mathbf{V}$ denote the unitary matrix that diagonalizes $\mathbf{F}_{AM}^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}_n^{-1}\mathbf{H}\mathbf{F}_{AM}$: $\mathbf{F}_{AM}^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}_n^{-1}\mathbf{H}\mathbf{F}_{AM} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^{\dagger}$. The optimal precoder can be taken as $\mathbf{F}_{opt} = \mathbf{F}_{AM}\mathbf{V}\mathbf{\Phi}$, where $\mathbf{\Phi}$ denotes the unitary matrix such that the MSE matrix has the identical diagonal elements. The existence of such unitary matrix $\mathbf{\Phi}$ is given by [65]. $\mathbf{\Phi}$ can be taken as a matrix with constant magnitude in each of its entries. Examples of such $\mathbf{\Phi}$ are the Hadamard matrix and the discrete Fourier transform (DFT) matrix [73]. Note that such $\mathbf{\Phi}$ are not unique, which means the minimum BER system is not unique.

3.3.3 DFE Transceivers

In this section we will focus on solving the problem of minimizing the BER subject to individual and total power constraints for the system with DFE and linear precoding. We will take the two-step approach. In the first step we will minimize the GM-MSE (geometric mean of mean square error) of the system. This is done by formulating the problem as a semi-definite program (SDP), as we shall see. In the second step, we will argue there is a set of systems in the minimum GM-MSE family, which yields the minimum average BER among all transceivers $\{\mathbf{F}, \mathbf{G}, \mathbf{B}\}$. A method to find one such optimal transceiver is also discussed in this section.

Since $\mathbf{E} = \mathbf{C}\mathbf{M}\mathbf{C}^{\dagger}$, and $\mathbf{C}$ is a lower triangular matrix with diagonal terms equal to the identity, we have the relation $\det(\mathbf{E}) = \det(\mathbf{C}\mathbf{M}\mathbf{C}^{\dagger}) = \det(\mathbf{M}) = \prod_{i=1}^{M} \mathbf{L}_{ii}^2$, which is the product of the MSE in each substream. Therefore, the minimization of the geometric mean of the MSEs is equivalent to the minimization of the determinant of $\mathbf{M}$.

Since $\mathbf{M}$ has the form as in Sec. 3.3.1, we have $\det(\mathbf{M}) = \det(\mathbf{I}_M + \mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}_n^{-1}\mathbf{H}\mathbf{F})$. Note that for any $m \times n$ matrix $\mathbf{X}$, we have the equality $\det(\mathbf{I}_m + \mathbf{X}\mathbf{X}^{\dagger}) = \det(\mathbf{I}_n + \mathbf{X}^{\dagger}\mathbf{X})$. Therefore we have $\det(\mathbf{I}_M + \mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}_n^{-1}\mathbf{H}\mathbf{F}) = \det(\mathbf{I}_N + \mathbf{R}_n^{-\frac{1}{2}}\mathbf{H}\mathbf{F}\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}_n^{-\frac{1}{2}})$, where $\mathbf{R}_n^{-\frac{1}{2}}$ is the Cholesky factor of the

noise covariance matrix $\mathbf{R}_n$. By setting $\mathbf{U} := \mathbf{F}\mathbf{F}^\dagger$, we can rewrite the problem of minimizing the GM-MSE in the following form:

$$\begin{aligned} \max_{\mathbf{U}} \quad & \log\det(\mathbf{I}_N + \mathbf{R}_n^{-\frac{1}{2}}\mathbf{H}\mathbf{U}\mathbf{H}^\dagger\mathbf{R}_n^{-\frac{1}{2}}) \\ \text{s.t.} \quad & \text{(a)} \quad (\mathbf{U})_{ii} \le P_i, \ \ \forall i = 1, 2, \cdots, M \\ & \text{(b)} \quad \text{Tr}(\mathbf{U}) \le P_{\text{total}} \\ & \text{(c)} \quad \mathbf{U} \succeq \mathbf{0}. \end{aligned} \tag{3.53}$$

This reformulation holds true because the log(·) function is a monotone function when the argument is positive. Problem (3.53) has been considered by several authors [105], [104]. It is an SDP-representable problem, and can be solved numerically by the interior point method efficiently [104]. See [104] and the references therein for more detailed discussions about the determinant maximization problem. To summarize, the minimum GM-MSE problem can be solved numerically efficiently, to a specified precision by the typical SDP solver.

First we observe that substituting any $\mathbf{F}$ with $\mathbf{F}_{new} := \mathbf{F}\mathbf{\Psi}$ for some unitary matrix $\mathbf{\Psi}$ does not change the GM-MSE nor the individual power in each antenna. In the high SNR region P_{br} is a Schur-convex increasing function in the vector $\mathbf{g} = [\log(\mathbf{E}_{11})\ \log(\mathbf{E}_{22})\ \cdots \log(\mathbf{E}_{MM})]$ [90]. Based on those observations, we have

$$\begin{aligned} P_{br} &= \frac{1}{M}\sum_{i=1}^{M} \alpha\text{Q}\left(\sqrt{\beta\cdot(\frac{1}{\mathbf{E}_{ii}} - 1)}\right) \\ &\ge \alpha\text{Q}\left(\sqrt{\beta\cdot(\frac{1}{\sqrt[M]{\prod_{i=1}^{M}\mathbf{E}_{ii}}} - 1)}\right) = \alpha\text{Q}\left(\sqrt{\beta\cdot(\frac{1}{\sqrt[M]{\det(\mathbf{E})}} - 1)}\right). \end{aligned} \tag{3.54}$$

It is now clear that the lower bound is minimized by minimum GM-MSE. The equality is achieved by choosing the matrix $\mathbf{\Psi}$ to equalize the MSE in each substream.

Now we provide a way to compute one solution for the optimal precoders. Suppose we already found the solution $\mathbf{U}_{GM}$ to the problem (3.53). The minimum GM-MSE precoder $\mathbf{F}_{GM}$ can be taken as any Cholesky factor of $\mathbf{U}_{GM}$. Suppose $\mathbf{V}$ is the unitary matrix diagonalizing $\mathbf{F}_{GM}^\dagger\mathbf{H}^\dagger\mathbf{R}_n^{-1}\mathbf{H}\mathbf{F}_{GM}$:

$$\mathbf{F}_{GM}^\dagger\mathbf{H}^\dagger\mathbf{R}_n^{-1}\mathbf{H}\mathbf{F}_{GM} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^\dagger.$$

Recall that the MSE matrix will be as in Sec. 3.3.1. From GTD theory, it can be shown that there exist unitary matrices $\mathbf{Q}$ and $\mathbf{\Phi}$ such that $(\mathbf{I}+\mathbf{\Sigma})^{-\frac{1}{2}} = \mathbf{Q}\mathbf{R}\mathbf{\Phi}^{\dagger}$, where $\mathbf{R}$ is an upper triangular matrix with diagonal terms all equal to the geometric mean of the diagonal terms of $(\mathbf{I}+\mathbf{\Sigma})^{-\frac{1}{2}}$. The optimal $\mathbf{F}_{opt}$ can be taken as $\mathbf{F}_{opt} = \mathbf{F}_{GM}\mathbf{V}\mathbf{\Phi}^{\dagger}$, where $\mathbf{\Phi}$ is the unitary matrix obtained by the triangular decomposition discussed above. Note that such $\mathbf{\Phi}$ are not unique, which means the minimum BER system is not unique.

3.3.4 Numerical Simulations

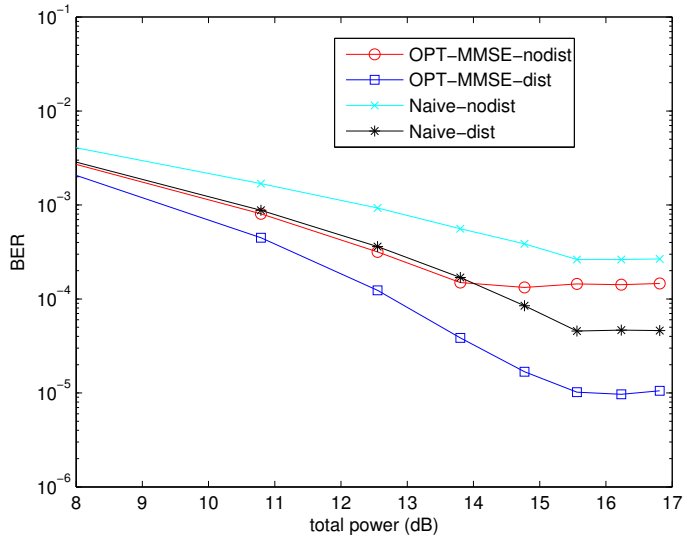


Figure 3.12: Comparing four transceivers for 100 channel realizations, with each antenna power ≤ 9. The x-axis represents the total power constraint.

We choose $M = 4$, $N = 5$, and per-antenna power constraints to be $[P_1, P_2, P_3, P_4] = [9, 9, 9, 9]$. The total power is varied in the simulation. The constellations are all QPSK. The noise is additive white Gaussian, with covariance matrix $\mathbf{R}_n = \mathbf{I}$. "OPT-MMSE-nodist" denotes the optimal MMSE design but without distributing the MSE in each substream. "OPT-MMSE-dist" denotes the optimal MMSE design with the MSE in each substream identical, which is the method proposed in this section. "Naive-nodist" denotes the case where the power constraints are satisfied by using the simple choice $\mathbf{F}_{naive} = \text{diag}([P_1, P_2, \cdots, P_M])$. if the total power constraint is not violated. If the

choice violates the total power constraint, then we take

$$\mathbf{F}_{naive} = \frac{P_{total}}{\sum_{i=1}^{M} P_i} \times \text{diag}([P_1, P_2, \cdots, P_M]).$$

"Naive-dist" corresponds to the case where the precoder matrix is $\mathbf{F} = \mathbf{F}_{naive}\mathbf{V}\mathbf{\Phi}$, where $\mathbf{\Phi}$ is to force the MSE matrix to have identical diagonal elements. Note that this method is exactly the one that was proposed in [76]. In Fig. 3.12 we provide the simulation of the linear transceivers for BER averaged over 100 channel realizations. The channel entries are drawn from an i.i.d. Gaussian distribution. It can be seen that the typical performance of the proposed method is significant. When the total power constraint is more than $\sum_{i=1}^{M} P_i = 36$ (or15.56dB), the total power constraint is actually inactivated. Therefore we can see the performance saturate after this point.

3.3.5 Concluding Remarks

In this subsection we gives several remarks on the extension of the framework developed here, the relation to literature, and the case where there is rectangular precoder.

Schur-Convex Objective Functions and Additional Linear Constraints

It can be observed that the discussion given above relies only on the fact that the average BER is a Schur-convex and increasing function of MSEs. Therefore the same concept can also be applied to other objective functions that have these two properties. Many examples of such optimization problems are provided in [73] for the linear transceiver case, and in [38, 90] for the DFE case. For all such objective functions, the systems discussed in the previous subsections are optimal.

It can also be seen that in the framework developed in this section, any finite number of linear constraints on the covariance matrix of the transmitted signals can be further added with no difficulty. This is because when the problem (3.51) or the problem (3.53) has one more constraint added: $f(\mathbf{F}\mathbf{F}^\dagger) \geq 0$, where $f(\mathbf{F}\mathbf{F}^\dagger)$ is a linear function in the elements of covariance matrix $\mathbf{F}\mathbf{F}^\dagger$ of the transmitted signal, it still remains an SDP.

Several examples of such linear constraints were addressed in [76], such as spectral masks in cable systems to control the crosstalk among DSL users, and limiting the power transmitted along some directions in wireless systems. Here we elaborate further about spatial masks constraints in the wireless systems. Suppose $\mathbf{a}$ is some spatial steering vector of interest, then the power along the direction is proportional to $\mathbf{a}^\dagger\mathbf{F}\mathbf{F}^\dagger\mathbf{a}$. Suppose we want to limit the power transmitted along this direction of interest, the constraint on the transmitted signal covariance becomes $\mathbf{a}^\dagger\mathbf{F}\mathbf{F}^\dagger\mathbf{a} \leq \alpha$, for

some constant α. This equation can be rewritten as $\mathbf{a}^\dagger\mathbf{F}\mathbf{F}^\dagger\mathbf{a} = \text{Tr}(\mathbf{F}\mathbf{F}^\dagger\mathbf{a}\mathbf{a}^\dagger) \leq \alpha$, which is a linear constraint in the transmitted covariance matrix $\mathbf{F}\mathbf{F}^\dagger$. Therefore the framework here can be easily modified to include this kind of constraints, both for the linear transceiver and the DFE with linear precoding.

Relation to Literature

An idea similar to the one discussed here was proposed in [76] where the author considered shaping constraints on the linear transceivers. In [76], the constraint on the covariance matrix of the transmitted signals is $\mathbf{F}\mathbf{F}^\dagger \preceq \mathbf{S}$, which means the matrix $\mathbf{S} - \mathbf{F}\mathbf{F}^\dagger$ is positive semi-definite. However, there is a difference between our work and [76], i.e., our constraint is componentwise while the constraint in [76] is the positive definiteness constraint. This is why our approach needs to be more involved (reformulating the problem to be a SDP for solving the minimum-AM MSE). Our approach has some advantages over that in [76].

1) Our work as well as some work in the literature, for example [139] and [45], precisely capture the individual power constraints. As acknowledged by the author in [76], the individual power constraints (3.47) are replaced with the tighter constraint (as Eq.(7) in [76]):

$$\mathbf{F}\mathbf{F}^\dagger \preceq \text{diag}([P_1, P_2, \cdots, P_M]).$$

This artificial replacement yields a solution in which the nondiagonal elements of $\mathbf{F}\mathbf{F}^\dagger$ are zero. We will see that the optimized solutions to the individual power constraint problem need not have zero nondiagonal elements for $\mathbf{F}\mathbf{F}^\dagger$. Therefore, the solution obtained in [76] is a sub-optimal solution to the individual power constraint problem. This fact will be amply shown in the numerical simulations.

2) As we argued in earlier, the power constraint along a direction should be like what is discussed in this section. It is shown here that this problem can be optimally solved under our framework. In [76] the author needs to find a shaping upper bound for the covariance matrix. However, the procedure of finding the tight upper bound was not trivial [76].

3) Actually, our framework can incorporate the problem discussed in [76]. This can be seen from the fact that the constraint in [76], $\mathbf{U} \preceq \mathbf{S}$, can be added in our framework, and the problem remains SDP. In our work, we also consider the optimization of the transceivers with DFE and linear precoder, while in [76] only linear transceivers are considered.

However, there are some disadvantages of our formulations compared to [76] when dealing

with the shaping constraint problems.

1) Our framework can only deal with square precoding matrix.

2) Given the SDP formulation, the optimal signaling direction cannot be characterized, whereas [76] gives a nice interpretation.

3) The computational complexity of our approach is much higher than [76] because of the need to solve the SDP. Solving an SDP requires $O(M^6)$ flop counts for each iteration.

Rectangular Precoder

If the precoder matrix $\mathbf{F}$ is not square, for example, when the channel matrix $\mathbf{H}$ is $N \times P$ and $P > M$, the rank of $\mathbf{U}$ should be no greater than M. This rank constraint should be further added into the problem formulation (3.51) and (3.53), which will destroy the convexity of the problem. Generally speaking, the rank-constrained problem is difficult to solve optimally. Therefore we propose a heuristic way to take care of this issue.

When $P > M$, suppose we first relax the rank constraint, then the rank-relaxed covariance matrix $\mathbf{U}$ is solved by the SDP solver as before. Now we denote $\mathbf{U} = \sum_{i=1}^{P} \lambda_i \mathbf{u}_i \mathbf{u}_i^\dagger$, where λ_i are the eigenvalues of $\mathbf{U}$ with non-increasing order and $\mathbf{u}_i$ are the corresponding eigenvectors with dimension $P \times 1$. Then we can take $\mathbf{F} = [\sqrt{\lambda_1}\mathbf{u}_1 \ \cdots \ \sqrt{\lambda_M}\mathbf{u}_M]\mathbf{\Phi}$, where $\mathbf{\Phi}$ is a $M \times M$ unitary matrix. This will result in $\mathbf{F}\mathbf{F}^\dagger = \sum_{i=1}^{M} \lambda_i \mathbf{u}_i \mathbf{u}_i^\dagger$, which is a rank-$M$ approximation of $\mathbf{U}$. Matrix $\mathbf{\Phi}$ can be obtained later as before, to distribute the MSE equally in each substream. It can be easily shown that with this approximation, the individual and total power constraint are still satisfied provided the original $\mathbf{U}$ is also in the feasible set of the problem formulation (3.51) and (3.53). The remaining eigenvalues may be scaled at this point to improve the system performance while maintaining the power constraints.

3.4 Conclusions

In this chapter we have presented several transceiver design applications where the concept of GTD and majorization is very useful. In Sec. 3.2 we have shown that under the zero-forcing condition, the minimum power required to achieve the specified probability of error by jointly designing the DFE transceivers and the bit loading scheme is the same as the one achieved by linear transceiver with bit loading. However, the presence of DFE gives many freedoms in the design. We showed that any instances in the GTD transceiver family are optimal solutions. In fact, many existing works in the literature can be incorporated into this GTD framework, while many new optimal

designs were also proposed. In Sec. 3.3 we have discussed the transceiver design problem with more realistic constraints, i.e., individual power constraints on the transmitted antennas or any linear constraints on the transmit covariance matrix. We have shown that the semi-definite programming (SDP) technique gives a nice framework to unify the design for linear transceivers and DFE transceivers. Furthermore, the theory of majorization was useful to obtain the minimum BER solutions.

3.5 Appendix

3.5.1 Proofs of Lemma 3.2.1

Proof: First note that the zero-forcing constraint is satisfied by $\mathbf{G}_{opt}$:

$$\mathbf{G}_{opt}\mathbf{HF} - \mathbf{B} = (\mathbf{I} + \mathbf{B})(\mathbf{HF})^{\sharp}\mathbf{HF} - \mathbf{B} = \mathbf{I}.$$

Suppose there is another $\mathbf{G}'$ satisfying the zero-forcing constraint with the given $\mathbf{F}$ and $\mathbf{B}$, i.e., $\mathbf{G}'\mathbf{HF} = \mathbf{I} + \mathbf{B}$. Define $\mathbf{\Delta} = \mathbf{G}_{opt} - \mathbf{G}'$. Since both $\mathbf{G}_{opt}$ and $\mathbf{G}'$ satisfy the zero-forcing constraint, it follows that

$$\begin{aligned}
\mathbf{\Delta}\mathbf{G}_{opt}^{\dagger} &= \mathbf{\Delta}\mathbf{HF}(\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{HF})^{-\dagger}(\mathbf{I} + \mathbf{B})^{\dagger} \\
&= (\mathbf{G}_{opt}\mathbf{HF} - \mathbf{G}'\mathbf{HF})(\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{HF})^{-\dagger}(\mathbf{I} + \mathbf{B})^{\dagger} \\
&= \mathbf{0}.
\end{aligned}$$

Therefore

$$\begin{aligned}
[\mathbf{G}'\mathbf{G}'^{\dagger}]_{kk} &= [(\mathbf{G}_{opt} - \mathbf{\Delta})(\mathbf{G}_{opt} - \mathbf{\Delta})^{\dagger}]_{kk} \\
&= [(\mathbf{G}_{opt}\mathbf{G}_{opt}^{\dagger} + \mathbf{\Delta}\mathbf{\Delta}^{\dagger}]_{kk} \\
&\geq [\mathbf{G}_{opt}\mathbf{G}_{opt}^{\dagger}]_{kk},
\end{aligned}$$

where we have used $\mathbf{\Delta}\mathbf{G}_{opt}^{\dagger} = \mathbf{0}$ in these inequalities. Therefore we have smaller sub-channel noise variances if we replace $\mathbf{G}'$ with $\mathbf{G}_{opt}$, hence with given bit rate and probabilities of error, a lower transmitted power can be achieved. □

3.5.2 Proofs of Theorem 3.2.4

Proof: Part (a) is true because the problem (3.12) discussed in previous sections is a relaxed version of the current problem (3.32). We prove part (b) by constructing a system that achieves P_{min} when (3.36) holds. If (3.36) holds then the majorization condition (2.3) can be satisfied by choosing $[\mathbf{R}]_{kk}$ to be positive square roots of

$$[\mathbf{R}]_{kk}^2 = \begin{cases} \frac{Mc_k 2^{b_k} (\prod_{k=1}^{M} \sigma_{h,k}^2)^{\frac{1}{M}}}{c2^b}, \text{for } k = 1, 2, \cdots, M. \\ \sigma_{h,k}^2 \qquad\qquad\qquad\qquad\quad , \text{for } k = M+1, \cdots K, \end{cases} \tag{3.55}$$

where K is the rank of $\mathbf{H}$. Then by the existence of GTD, there exists a $K \times K$ upper triangular matrix $\mathbf{R}$, such that the decomposition $\mathbf{H} = \mathbf{QRP}^\dagger$ is true, where $\mathbf{Q}$ and $\mathbf{P}$ have orthonormal columns. Now choose the transceiver matrices $\mathbf{F}, \mathbf{G}$, and $\mathbf{B}$ as in (3.20), (3.21), and (3.23). Then P_{trans} is as in (3.25). Substituting from (3.55) we get $P_{trans} = P_{min}$ indeed.

□

Chapter 4

Transceivers Designs for MIMO Frequency Selective Channels

In high rate digital communication systems, multiple-input-multiple-output (MIMO) frequency selective (FS) channels complicate the transceiver design process because of the inter-block-interference (IBI) effect. However, by applying the zero-padding precoding technique, we can eliminate the IBI and convert the FS channel into an equivalent MIMO block channel [12, 89]. With MIMO block channels, many researchers have developed transceiver designs to match the channel characteristics and to mitigate the noise interference [12, 73, 75, 89, 90]. One of the approaches is to focus on linear precoding and decision feedback equalization (DFE).

If the channel state information (CSI) is available both at the transmitter and the receiver sides, in terms of minimizing the average BER under the transmitted power constraint, the optimal system with linear precoding and zero-forcing DFE (ZF-DFE) [90], and the optimal system with linear precoding and minimum-mean-square-error DFE (MMSE-DFE) [37], can both be derived from the equivalent block channel matrix. For ZF-DFE, the optimal linear precoder matrix has orthonormal columns; for MMSE-DFE, the optimal linear precoder no longer has orthonormal columns, instead it has suitable power loading on the channel eigenmodes. However, precoding matrix with orthonormal columns is usually desired for simplicity reasons [142, 56, 99, 141]. Under the unitary precoder constraints, the optimal systems[1] for both receiver types can be derived. Nevertheless, it is known that the derived optimal systems suffer from two drawbacks. First, they require a large number of bits from the receiver to encode the full precoding matrix and feed it back to the transmitter [90]. Second, the full precoding matrix multiplication is computationally complex. For the

[1]The optimal systems is referred to the optimal designs within the class of systems using unitary precoders and DFEs (ZF or MMSE). These systems will be called the optimal systems through out this chapter.

block channel derived from a zero-padded MIMO FS channel, these disadvantages become more apparent when the block size is large.

The block diagonal GMD (BD-GMD) is proposed in [47] to design memoryless transceivers for MIMO broadcast channels. In this chapter, we consider applying the BD-GMD technique to design the *zero-padded* MIMO FS transceiver that solves the two mentioned drawbacks. Two novel systems (which we call ZP-BD-GMD systems) are proposed: the ZF-BD-GMD system, which uses block diagonal unitary precoder and ZF-DFE receiver, and the MMSE-BD-GMD system, which uses block diagonal unitary precoder and MMSE-DFE receiver. We will show the following properties of the proposed ZP-BD-GMD systems:

1. Because of the block diagonal structure of the precoder matrix, the proposed ZP-BD-GMD systems solve the two implementation drawbacks of the optimal systems. It is also shown that the receiver structures are simpler than those of the optimal systems. Therefore, the ZP-BD-GMD systems have a much smaller implementation cost than the optimal systems.

2. For finite block sizes, it can be seen that any block diagonal unitary precoder system solves the above drawbacks. In particular, ZP-BD-GMD systems are minimizers of the average BER within the family of systems that use block diagonal unitary precoder. That is, the ZP-BD-GMD systems have optimality for any block size. As block size gets larger and approaches infinity, the average BER of the ZP-BD-GMD systems also approaches that of the optimal unitary precoded systems. In other words, the ZP-BD-GMD systems are asymptotically optimal within the systems that use unitary precoder, as the bandwidth efficiency approaches unity.

3. In all four unitary precoded systems (ZF-Optimal, ZF-BD-GMD, MMSE-Optimal, and MMSE-BD-GMD), there is a tradeoff between the bandwidth efficiency and the average BER performance. This suggests that one has to carefully design the block length to maintain the target BER for both the ZP-BD-GMD systems and the optimal systems. In [72], a similar tradeoff in single-carrier zero-padded SISO FS channels with linear equalization was reported. Thus, this aspect of our work can be seen as an extension of [72].

4. In the case of the SISO channel, ZP-BD-GMD systems have the same performance as the lazy precoder systems, i.e., systems with identity precoding matrix. Therefore, in SISO channels the lazy precoder systems inherit the benefits of the ZP-BD-GMD systems, making the lazy precoder transceivers asymptotically optimal in the class of systems with unitary precoder

and DFE. While having the same performance, lazy precoders are more desirable than the ZP-BD-GMD systems in terms of implementation cost. This is due to the fact that lazy precoders can transmit data without CSI at the transmitter and precoding matrix multiplication.

These properties make the proposed ZP-BD-GMD systems more favorable designs in practical implementation than the optimal systems.

The content of this chapter is mainly drawn from [130], and portions of it have been presented in [127].

4.1 Outline

This chapter is structured as follows: In Sec. 4.2, we will introduce the communication model and some preliminaries. Sec. 4.3 describes the proposed ZF-BD-GMD transceiver structure, which uses a block diagonal unitary precoder and a ZF-DFE design based on BD-GMD of the effective channel matrix. Several properties of the ZF-BD-GMD transceiver are discussed. The implementation cost is also analyzed. Sec. 4.4 extends the idea to the MMSE-DFE case. The proposed MMSE-BD-GMD system is discussed. Most of the results will be similar to the ZF case, so this section will be brief. Sec. 4.5 explains that for the two ZF transceivers (ZF-Optimal and ZF-BD-GMD) and the two MMSE transceivers (MMSE-Optimal and MMSE-BD-GMD), there exists a tradeoff between the bandwidth efficiency and the BER performance. Sec. 4.6 discusses the SISO channel case, in which the lazy precoder system is discussed. Sec. 4.7 presents the numerical simulation results related to the topics in the chapter.

4.2 Signal Model

We consider a point-to-point communication system with N_T transmitting antennas and N_R receiving antennas. The input-output relation of the frequency selective MIMO channel can be expressed as

$$\mathbf{y}_i = \sum_{k=0}^{L} \mathbf{H}_k \mathbf{x}_{i-k} + \mathbf{n}_i, \tag{4.1}$$

where $\mathbf{x}_i$ is the $N_T \times 1$ transmitted signal, $\mathbf{H}(z) = \mathbf{H}_0 + \mathbf{H}_1 z^{-1} + \cdots + \mathbf{H}_L z^{-L}$ is the Lth order $N_R \times N_T$ frequency selective FIR MIMO channel, $\mathbf{n}_i$ is the additive channel noise, and $\mathbf{y}_i$ is the $N_R \times 1$

received vector. The noise covariance matrix is assumed to be $\mathbf{R}_n = \sigma_n^2\mathbf{I}$. The zero-padded system transmits N_P zero vectors after every K symbol vectors. That is, in $K + N_P$ symbol durations, the following is transmitted: $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_K, \mathbf{0}, \cdots, \mathbf{0}\}$. In order to prevent contamination from previous blocks, one must choose $N_P \geq L$. The bandwidth efficiency is defined as

$$\epsilon = \frac{K}{K + N_P}. \tag{4.2}$$

Therefore it is desirable to choose $N_P = L$, so that the BW efficiency is maximized, and equal to $K/(K + L)$. Throughout the chapter, we assume $N_P = L$. The I/O relation of the zero-padded MIMO frequency selective system can be expressed as an equivalent block channel:

$$\underbrace{\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{K+L} \end{bmatrix}}_{\mathbf{y}_{ZP,K}} = \mathbf{H}_{ZP,K} \underbrace{\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_K \end{bmatrix}}_{\mathbf{x}_{ZP,K}} + \underbrace{\begin{bmatrix} \mathbf{n}_1 \\ \mathbf{n}_2 \\ \vdots \\ \mathbf{n}_{K+L} \end{bmatrix}}_{\mathbf{n}_{ZP,K}}, \tag{4.3}$$

where

$$\mathbf{H}_{ZP,K} = \begin{bmatrix} \mathbf{H}_0 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{H}_1 & \mathbf{H}_0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{H}_L & \vdots & \ddots & \mathbf{H}_0 \\ \mathbf{0} & \mathbf{H}_L & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{H}_L \end{bmatrix}, \tag{4.4}$$

and K in the subscript denotes that $\mathbf{H}_{ZP,K}$ has KN_T columns. Note that Eq. (4.3) holds for any $N_P \geq L$. We assume $(K + L)N_R \geq KN_T$, so that the zero-forcing condition can be satisfied.

We consider the system where the transmitted vector is linear precoded by a $N_TK \times N_TK$ matrix $\mathbf{P}$:

$$\mathbf{x}_{ZP,K} = \mathbf{P}\mathbf{s},$$

where $\mathbf{s} = \{\mathbf{s}_1^T, \mathbf{s}_2^T, \cdots, \mathbf{s}_K^T\}^T$, and $\mathbf{s}_i$ is the $N_T \times 1$ transmitted symbol vector. We use the usual

assumption that the transmitted signal is zero-mean, white, and uncorrelated with the noise, i.e., $E[\mathbf{s}_i\mathbf{s}_j^H] = \delta(i-j)\sigma_s^2\mathbf{I}$ and $E[\mathbf{s}_i\mathbf{n}_j] = \mathbf{0}$. Here we define a constant ζ, which stands for the noise to symbol power ratio:

$$\zeta \doteq \sigma_n^2/\sigma_s^2. \tag{4.5}$$

The average power of the transmitted vector $\mathbf{x}_{ZP,K}$ is restricted to be less than $KN_T\sigma_s^2$. Note that the power constraint is proportional to K because K is the number of symbol vectors transmitted in one block. Since $E[\mathbf{s}\mathbf{s}^H] = \sigma_s^2\mathbf{I}$, the power constraint can be written as

$$\frac{1}{\sigma_s^2}\mathrm{Tr}(E[\mathbf{P}\mathbf{s}\mathbf{s}^H\mathbf{P}^H]) = \mathrm{Tr}(\mathbf{P}\mathbf{P}^H) \leq KN_T, \tag{4.6}$$

which is a constraint expressed solely in terms of the precoder matrix. We assume each symbol is selected from the same QAM constellation, i.e., no bit allocation is applied. In this case, the BER will be the function of SINR of the input to the decision device (see Eq. (12) in [73]), i.e.,

$$\mathrm{BER}(\mathrm{SINR}_k) = \alpha Q(\beta\sqrt{\mathrm{SINR}_k}), \tag{4.7}$$

where α and β are constants which depend on the constellation, and $Q(\cdot)$ is the Q-function defined as $Q(x) = (1/\sqrt{2\pi})\int_x^\infty e^{-\lambda^2/2}d\lambda$. We are interested in the high SNR regime so that the BER function is a convex function of the logarithm of the SINR [90].[2]

The optimization problem we are interested in is to minimize the average BER by designing a linear precoder and a zero-forcing or MMSE decision feedback equalizer jointly under the power constraint (4.6). We can treat the I/O relation (4.3) as an effective block channel communication system. For the ZF-DFE case, the optimal solution is suggested by the Theorem 1 in [91]. The optimal precoder is with no loss of generality a unitary matrix. The optimal receiver is the corresponding ZF-DFE solution suggested in Sec. III of [91].

If the receiver is MMSE-DFE instead of ZF-DFE, the optimal precoder will no longer be unitary [90, 37]. Instead, a suitable water-filling power loading to the channel eigenmodes is needed to achieve the optimal performance [37]. However, unitary precoding is usually desired for simplicity

[2]The property we need for the discussion in this chapter is that the average BER is a Schur-convex function of the logarithm of the effective subchannel gains (see Appendix A.(f) in [90] for details). Therefore, it should be noted that the theorems developed in this chapter (Thm 4.3.4, Thm 4.3.5, Thm 4.5.1, and the corresponding properties of the MMSE-BD-GMD systems) are not restricted to the average BER, but are also true for a broader class of metric.

reasons [142, 56, 99, 141]. Therefore in this chapter, we restrict our interest only to the unitary precoder case. In this case, the optimal system for MMSE-DFE can also be obtained.

Generally, the optimal precoders for both optimal ZF and MMSE systems are full matrices. Therefore, the implementation of the optimal systems suffers from two disadvantages:

1. In the limited feedback scheme [55, 91], the channel state information is estimated by the receiver, and the optimal precoder information is quantized (or quantized to some predetermined codebook) and fed back to the transmitter. Since the optimal precoder $\mathbf{P}$ is an $N_T K \times N_T K$ unitary matrix, it requires a large codebook for quantizing it to cover the whole space of the $N_T K \times N_T K$ unitary matrices.

2. Computation of the transmitted signal $\mathbf{x}_{ZP,K} = \mathbf{P}\mathbf{s}$ is expensive for the full matrix multiplication, and takes $O(N_T^2 K^2)$ operations.

These two disadvantages are more severe when K is large, i.e., when the bandwidth efficiency (4.2) approaches unity. To overcome these, we propose using the BD-GMD technique, which was introduced in [47], to design the transceiver. We restrict the precoder to be block-diagonal, i.e., $\mathbf{P} = \text{diag}(\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_K)$, where $\mathbf{P}_i$ is an $N_T \times N_T$ unitary matrix. The block diagonal constraint clearly simplifies the implementation:

1. Since there are only $N_T^2 K$ out of $N_T^2 K^2$ elements that are nonzero, the required size of the codebook is much smaller for covering the precoder matrix space.

2. To form the transmitted vector $\mathbf{x}_{ZP,K} = \mathbf{P}\mathbf{s}$, we only need to form $\mathbf{x}_i = \mathbf{P}_i \mathbf{s}_i$ for $i = 1, \cdots, K$ and concatenate them. The complexity is only $O(N_T^2 K)$ instead of $O(N_T^2 K^2)$ as in the optimal precoder case without block diagonal constraint.

Although the block diagonal precoder gives these benefits, it is natural to ask how much the performance degrades due to this constraint. In the following sections, we will discuss several important properties of the proposed ZP-BD-GMD transceivers. It will be shown that the ZP-BD-GMD system has similar performance as the optimal systems when the bandwidth efficiency approaches unity. In addition, the receiver structure of the ZP-BD-GMD systems is computationally simple. Also, we will prove that the ZP-BD-GMD systems are optimal within the family of transceivers with DFEs and block diagonal unitary precoders.

4.3 Transceivers with Zero-Forcing DFEs

The block-diagonal geometric mean decomposition (BD-GMD) technique was introduced in [47] to design precoders with dirty paper coding for MIMO broadcast channels. The schemes in [47] convert each user's MIMO channel into parallel subchannels with identical SINRs, thus equal-rate coding can be applied across the subchannels of each user.

In this chapter we use the BD-GMD idea for a new application – transceiver design for zero-padded MIMO frequency selective channels. We introduce the proposed zero-forcing BD-GMD (ZF-BD-GMD) system, which uses a block diagonal unitary linear precoder and a zero-forcing DFE. Let us consider the BD-GMD of $\mathbf{H}_{ZP,K}^H$:

$$\begin{aligned}\mathbf{H}_{ZP,K}^H &= \begin{bmatrix} \mathbf{H}_0^H & \mathbf{H}_1^H & \cdots & \mathbf{H}_L^H & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_0^H & \ddots & \ddots & \mathbf{H}_L^H & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{H}_0^H & \cdots & \cdots & \mathbf{H}_L^H \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} \mathbf{P}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{P}_K \end{bmatrix}}_{\mathbf{P}} \underbrace{\begin{bmatrix} \mathbf{L}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \times & \mathbf{L}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \times & \cdots & \times & \mathbf{L}_K \end{bmatrix}}_{\mathbf{L}} \mathbf{Q}^H,\end{aligned}$$

where $\mathbf{P}_i$'s are $N_T \times N_T$ unitary matrices, $\mathbf{Q}$ is a $(K+L)N_R \times KN_T$ matrix with orthonormal columns, each $N_T \times N_T$ matrix $\mathbf{L}_i$ is lower triangular with equal diagonal elements. Also, "$\times$" refers to possibly nonzero entries. Note that in the ZF-DFE case, unitarity of precoder is not a loss of generality [91], and so, this constraint will not be mentioned explicitly again.

The proposed ZF-BD-GMD transceiver is based on this decomposition. The block diagonal precoder is chosen as the block diagonal matrix $\mathbf{P}$, and the receiving feedforward filter is chosen as $\mathbf{Q}^H$. Since $\mathbf{Q}$ has orthonormal columns, the channel noise after $\mathbf{Q}^H$ is still white with variance σ_n^2. Since $\mathbf{Q}^H\mathbf{H}_{ZP,K}\mathbf{P} = \mathbf{L}^H$, the effective channel from the input of precoder $\mathbf{P}$ to the output of the feedforward filter $\mathbf{Q}^H$ at the receiver is the upper triangular matrix $\mathbf{L}^H$. The corresponding effective channel noise is still white. This triangular structure facilitates simple decision feedback equalization [38, 90]. Fig. 4.1 shows the transceiver structure of the ZF-BD-GMD system.

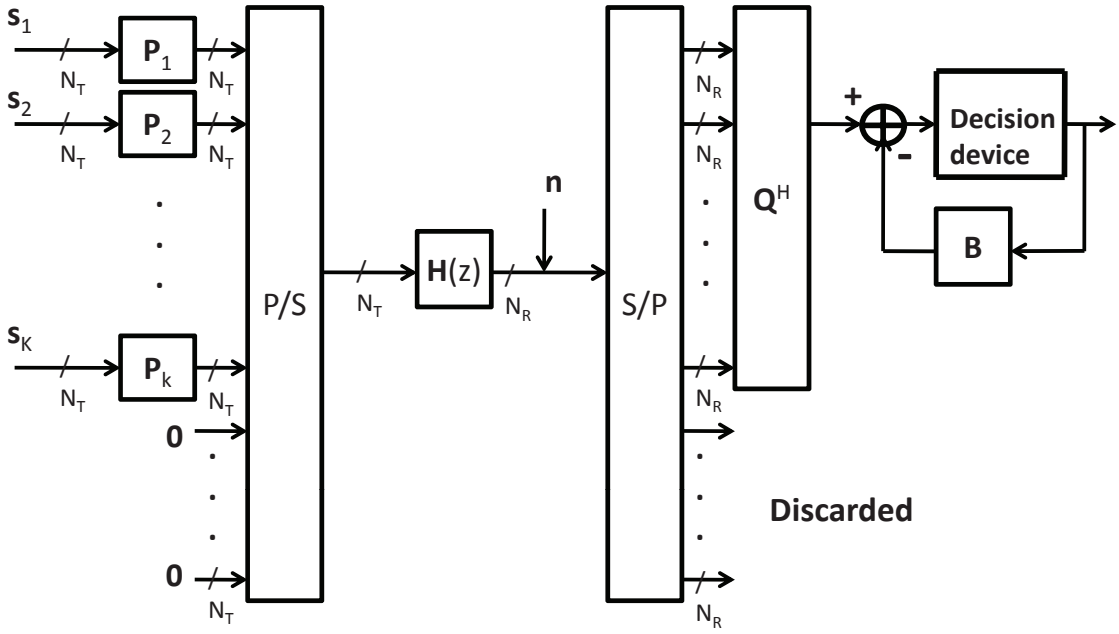


Figure 4.1: The ZP-BD-GMD transceiver. The signal vector $\mathbf{s}_i$ is first linear precoded by the unitary matrix $\mathbf{P}_i$. The precoded symbol vectors and N_P zero vectors are then passed through a parallel-to-serial converter before transmitting to channel $\mathbf{H}(z)$. The receiver discards the contaminated signals, and passes the clean signal through DFE. $\mathbf{Q}^H$ is the feedforward filter, and $\mathbf{B}$ is the feedback filter, whose coefficients are obtained from the entries in $\mathbf{L}$.

As in many analyses of DFE systems [37, 90], we assume that there is no error propagation in the feedback loop. Based on this assumption, after the decision feedback process, the overall ZF-BD-GMD system behaves like a system with KN_T independent parallel SISO AWGN subchannels. Each subchannel has identical noise variance σ_n^2, and the mth subchannel has subchannel gain $[\mathbf{L}]_{mm}$, which is the mth diagonal entry of $\mathbf{L}$. Since the transmitted symbol vector has energy σ_s^2 per component, the SINR in mth stream before the detection device is

$$\text{SINR}_m = |[\mathbf{L}]_{mm}|^2\sigma_s^2/\sigma_n^2 = |[\mathbf{L}]_{mm}|^2/\zeta. \tag{4.8}$$

The BER of mth stream will be the function of SINR as in (4.7), i.e.,

$$\text{BER}(\text{SINR}_m) = \alpha Q(\beta\sqrt{\text{SINR}_m}) = \alpha Q(\beta\sqrt{|[\mathbf{L}]_{mm}|^2/\zeta}) \tag{4.9}$$

Therefore, to analyze the performance of the BD-GMD transceiver, we have to study the diagonal entries of $\mathbf{L}$.

Since the $N_T \times N_T$ lower triangular $\mathbf{L}_i$ has identical diagonal entries, we shall denote it as r_i. This is the gain of the ith effective subchannel, i.e.,

$$r_i \equiv [\mathbf{L}_i]_{kk}, \text{for } k = 1, 2, \cdots, N_T.$$

From the property of BD-GMD (Eq. (21) in [47]), we have

$$r_i = \left(\frac{\det(\mathbf{H}_{ZP,i}^H\mathbf{H}_{ZP,i})}{\det(\mathbf{H}_{ZP,i-1}^H\mathbf{H}_{ZP,i-1})}\right)^{\frac{1}{2N_T}} \tag{4.10}$$

Let us define

$$\tilde{\mathbf{H}}_m \doteq \begin{bmatrix}\mathbf{H}_m^H & \mathbf{H}_{m+1}^H & \cdots & \mathbf{H}_L^H\end{bmatrix}\begin{bmatrix}\mathbf{H}_0\\ \mathbf{H}_1\\ \vdots\\ \mathbf{H}_{L-m}\end{bmatrix} = \sum_{k=m}^{L}\mathbf{H}_k^H\mathbf{H}_{k-m} \tag{4.11}$$

for $m = 0, 1, \cdots, L$, and $\tilde{\mathbf{H}}_m = \mathbf{0}$ for $m > L$. Let

$$\tilde{\mathbf{H}}_{-k} = \tilde{\mathbf{H}}_k^H.$$

In particular, $\tilde{\mathbf{H}}_0$ is Hermitian. From the above definition (4.11), we have the following equation

$$\mathbf{H}_{ZP,m+1}^H \mathbf{H}_{ZP,m+1} = \begin{bmatrix} \tilde{\mathbf{H}}_0 & \tilde{\mathbf{H}}_1 & \cdots & \tilde{\mathbf{H}}_m \\ \tilde{\mathbf{H}}_{-1} & \tilde{\mathbf{H}}_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \tilde{\mathbf{H}}_{-m} & \cdots & \cdots & \tilde{\mathbf{H}}_0 \end{bmatrix}, \tag{4.12}$$

which is a Hermitian block Toeplitz positive semi-definite matrix. From (4.11), it is also noted that

$$\left(\sum_{k=0}^{L} z^k \mathbf{H}_k^H\right)\left(\sum_{k=0}^{L} z^{-k} \mathbf{H}_k\right) = \sum_{m=-L}^{L} \tilde{\mathbf{H}}_m z^{-m}.$$

Thus

$$\tilde{\mathbf{H}}(e^{j\omega}) \doteq \sum_{m=-L}^{L} \tilde{\mathbf{H}}_m e^{-jm\omega} = \left(\mathbf{H}(e^{j\omega})\right)^H \mathbf{H}(e^{j\omega}), \tag{4.13}$$

where $\mathbf{H}(e^{j\omega}) = \sum_{k=0}^{L} \mathbf{H}_k e^{-j\omega k}$. This shows that the matrix $\tilde{\mathbf{H}}(e^{j\omega})$ is always positive semi-definite, and can be seen as the power spectrum density matrix of a fictitious wide-sense-stationary (WSS) process $\mathbf{z}(n)$ formed by passing a unit-variance white input vector $\mathbf{x}(n)$ through an LTI system with transfer function $\mathbf{H}(z)$. These quantities will be useful in characterizing the proposed system performances.

In the following, we will provide several properties of r_i, which characterize the performance of ZF-BD-GMD systems. The following lemma is useful for deriving the properties.

Lemma 4.3.1 *Suppose that the Hermitian matrices* $\mathbf{D}$ *and* $\begin{bmatrix} \mathbf{A} & \mathbf{B}^H \\ \mathbf{B} & \mathbf{D} \end{bmatrix}$ *are both positive definite, where* $\mathbf{D}$ *has size* $n \times n$*,* $\mathbf{A}$ *has size* $m \times m$ *and* $\mathbf{B}$ *has size* $n \times m$*. Then, for any pair of matrices* $\mathbf{P}$ *and* $\mathbf{Q}$*, where* $\mathbf{P}$ *has size* $k \times m$*, and* $\mathbf{Q}$ *has size* $k \times n$*, we have*

$$\begin{bmatrix} \mathbf{P} & \mathbf{Q} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{B}^H \\ \mathbf{B} & \mathbf{D} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{P}^H \\ \mathbf{Q}^H \end{bmatrix} \succeq \mathbf{Q}\mathbf{D}^{-1}\mathbf{Q}^H. \tag{4.14}$$

Proof: See Appendix. □

Now, we are ready to introduce our first theorem.

Theorem 4.3.2 *r_m is non-increasing. That is, for $m \geq 1$, $r_{m+1} \leq r_m$.* ♢

Proof: Based on the Block Toeplitz structure of $\mathbf{H}_{ZP,m}$, we can write

$$\mathbf{H}_{ZP,m+1}^H \mathbf{H}_{ZP,m+1} = \begin{bmatrix} \tilde{\mathbf{H}}_0 & \mathbf{C}_m \\ \mathbf{C}_m^H & \mathbf{H}_{ZP,m}^H \mathbf{H}_{ZP,m} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{ZP,m}^H \mathbf{H}_{ZP,m} & \mathbf{B}_m^H \\ \mathbf{B}_m & \tilde{\mathbf{H}}_0 \end{bmatrix}, \tag{4.15}$$

where $\mathbf{B}_m = \begin{bmatrix} \tilde{\mathbf{H}}_{-m} & \tilde{\mathbf{H}}_{-(m-1)} & \cdots & \tilde{\mathbf{H}}_{-1} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{H}}_{-m} & \mathbf{B}_{m-1} \end{bmatrix}$, and $\mathbf{C}_m = [\tilde{\mathbf{H}}_1 \quad \tilde{\mathbf{H}}_2 \quad \cdots \quad \tilde{\mathbf{H}}_m] = [\mathbf{C}_{m-1} \quad \tilde{\mathbf{H}}_m]$. By taking determinant on both sides of (4.15), we get

$$\det(\mathbf{H}_{ZP,m+1}^H \mathbf{H}_{ZP,m+1}) = \det(\mathbf{H}_{ZP,m}^H \mathbf{H}_{ZP,m}) \det(\tilde{\mathbf{H}}_0 - \mathbf{B}_m (\mathbf{H}_{ZP,m}^H \mathbf{H}_{ZP,m})^{-1} \mathbf{B}_m^H),$$

where we have used the expression for determinants of partitioned matrices (p. 472 in [31]). Using this in (4.10), r_m can be written as

$$r_m = \det(\tilde{\mathbf{H}}_0 - \mathbf{B}_{m-1} (\mathbf{H}_{ZP,m-1}^H \mathbf{H}_{ZP,m-1})^{-1} \mathbf{B}_{m-1}^H)^{\frac{1}{2N_T}}.$$

Therefore we have

$$\begin{aligned} & \mathbf{B}_m (\mathbf{H}_{ZP,m}^H \mathbf{H}_{ZP,m})^{-1} \mathbf{B}_m^H \\ = \; & [\tilde{\mathbf{H}}_{-m} \quad \mathbf{B}_{m-1}] \begin{bmatrix} \tilde{\mathbf{H}}_0 & \mathbf{C}_{m-1} \\ \mathbf{C}_{m-1}^H & \mathbf{H}_{ZP,m-1}^H \mathbf{H}_{ZP,m-1} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{\mathbf{H}}_{-m}^H \\ \mathbf{B}_{m-1}^H \end{bmatrix} \\ \succeq \; & \mathbf{B}_{m-1} (\mathbf{H}_{ZP,m-1}^H \mathbf{H}_{ZP,m-1})^{-1} \mathbf{B}_{m-1}^H, \end{aligned}$$

where the last inequality follows from Lemma 1. Subtracting the Hermitian matrix $\tilde{\mathbf{H}}_0$ from both sides, we get

$$\tilde{\mathbf{H}}_0 - \mathbf{B}_m (\mathbf{H}_{ZP,m}^H \mathbf{H}_{ZP,m})^{-1} \mathbf{B}_m^H \preceq \tilde{\mathbf{H}}_0 - \mathbf{B}_{m-1} (\mathbf{H}_{ZP,m-1}^H \mathbf{H}_{ZP,m-1})^{-1} \mathbf{B}_{m-1}^H.$$

Taking the determinant, we arrive at $r_{m+1} \leq r_m$. □

Theorem 6.4.1 states that the ZF-BD-GMD system creates unequal subchannel gains for the effective parallel SISO subchannels, and the subchannel gains are in a non-increasing order. For a given block size K, r_K is the worst subchannel gain (because $r_{m+1} \leq r_m$). This r_K is non-increasing and has a limit:

Theorem 4.3.3 *The limit of r_K as $K \to \infty$ is:*

$$\lim_{K\to\infty} r_K \doteq r = \exp\left(\frac{1}{4N_T\pi}\int_{-\pi}^{\pi} \log\det \tilde{\mathbf{H}}(e^{j\omega})d\omega\right), \tag{4.16}$$

where $\tilde{\mathbf{H}}(e^{j\omega})$ is defined in (4.13). ♢

Proof: By Eq. (4.10), $r_{m+1}^{2N_T}$ is the ratio of the determinant of the block Toeplitz matrices with order $m+1$ and m. Applying Eq. (1.12) in [117], we have

$$\lim_{M\to\infty} r_M^{2N_T} = \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\det \tilde{\mathbf{H}}(e^{j\omega})d\omega\right), \tag{4.17}$$

where we have also used Eq. (1.10) in [117]. Taking the $2N_T$-th root of (4.17), the theorem follows. □

The preceding result holds under certain conditions that ensure the integrand above is well behaved (see condition Eq. (1.8) in [117]).

It can be seen that if the matrix $\tilde{\mathbf{H}}(e^{j\omega})$ is close to singular in some frequency band, the value of the above integral will be small. In this case, the asymptotic subchannel gain will be small, which implies a poor BER performance. This is consistent with intuition.

To gain more understanding of these theorems, we refer the reader to Sec. 4.7 where we show the channel gain behavior of a ZF-BD-GMD system for a MIMO FS channel for different block sizes K. The optimal system has unitary precoder and identical subchannel gains for all the subchannels [90]. It can be shown that for a given K, the value of the subchannel gains of the ZF-Optimal system, which is denoted as g_K, will be equal to the geometric mean of $\{r_1, r_2, \cdots, r_K\}$. That is,

$$g_K = \left(\prod_{k=1}^{K} r_k\right)^{\frac{1}{K}}. \tag{4.18}$$

Since from Theorem 6.4.1 we know that r_k is non-increasing, g_k is also non-increasing. It can also be shown that $g_K \to r$ as $K \to \infty$, where r is defined in (4.16). Thus, the subchannel gains of the ZF-Optimal system approach r as the block size increases. This is the intuition behind the third theorem, which states the asymptotic optimality of the ZF-BD-GMD transceiver. In what follows, $P(x)$ is defined as

$$P(x) \doteq \alpha Q(\beta\sqrt{x^2/\zeta}), \tag{4.19}$$

which denotes the BER as a function of the subchannel gain in the ZF-DFE systems with noise to symbol power ratio ζ defined in (4.5).

Theorem 4.3.4 *The average BER of the ZF-BD-GMD transceiver approaches the average BER of the optimal system when the bandwidth efficiency approaches unity:*

$$\lim_{K\to\infty} BER_{ZFBDGMD}(K) = \lim_{K\to\infty} BER_{ZFoptimal}(K) = P(r),$$

where K denotes the block size. $\diamondsuit$

Proof: Let us first focus on the BER of the ZF-BD-GMD system. From (4.7) and (4.9), we see that the BER in the kth subchannel is only a function of channel gain $[\mathbf{L}]_{kk}$. It is clear that $P(x)$ is a continuous and decreasing function of its argument x. By Theorem 6.4.3, the channel gain r_i is non-increasing. Since $P(\cdot)$ is decreasing with its argument, we have $P(r_{i+1}) \geq P(r_i)$.

Let β_K denote the average BER:

$$\beta_K \doteq \text{BER}_{ZFBDGMD}(K) = \frac{1}{K}\sum_{i=1}^{K} P(r_i). \tag{4.20}$$

It can be seen that β_K is a non-decreasing sequence, so it converges (since it is upper bounded by unity). It can shown that the limit will be $P(r)$, where r is the limit of r_i in Theorem 6.4.3. Thus we have proved

$$\lim_{K\to\infty} \text{BER}_{ZFBDGMD}(K) = \lim_{K\to\infty} \frac{1}{K}\sum_{i=1}^{K} P(r_i) = P(r).$$

Now let us look at the average BER of the ZF-Optimal system. The optimal system has unitary precoder and identical subchannel gains for all the subchannels [90]. Since the product of the subchannel gains is always $\det(\mathbf{H}_{ZP,K}^H \mathbf{H}_{ZP,K})^{\frac{1}{2}}$, the subchannel gain can be calculated by taking the KN_T-th root. The average BER becomes

$$\text{BER}_{ZFoptimal}(K) = P\left(\det(\mathbf{H}_{ZP,K}^H \mathbf{H}_{ZP,K})^{\frac{1}{2KN_T}}\right).$$

The limit of $\det(\mathbf{H}_{ZP,K}^H \mathbf{H}_{ZP,K})^{\frac{1}{K}}$ is given by Eq. (1.5) in [117]:

$$\lim_{K\to\infty} \det(\mathbf{H}_{ZP,K}^H \mathbf{H}_{ZP,K})^{\frac{1}{K}} = \lim_{K\to\infty} \sqrt[K]{C}\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\det \tilde{\mathbf{H}}(e^{j\omega})d\omega\right) = r^{2N_T},$$

where r is defined in (4.16), C is a constant, and we have used the fact that $\sqrt[K]{C}$ approaches 1 when K approaches infinity. Therefore,

$$\lim_{K\to\infty} \text{BER}_{ZFoptimal}(K) = \lim_{K\to\infty} P\left(\det(\mathbf{H}_{ZP,K}^H \mathbf{H}_{ZP,K})^{\frac{1}{2KN_T}}\right) = P(r).$$

This completes the proof. □

This theorem shows the asymptotic optimality of the ZF-BD-GMD transceiver. However, the block size needs to be chosen according to the channel coherence time and cannot be too large in a fast fading channel scenario [89]. Thus, it is important to characterize the performance of the ZF-BD-GMD system for finite block sizes. The following theorem suggests the optimality of the ZF-BD-GMD system for any finite block size.

Theorem 4.3.5 *Consider the family of systems for zero-padded MIMO frequency selective channels with fixed block size K that use block diagonal unitary precoders and ZF-DFEs. Within this family, the ZF-BD-GMD system is one of the minimizers for the average BER.* ♢

Proof: See Appendix. □

We now discuss the implementation cost of transmitter and receiver in the ZF-BD-GMD system. As mentioned previously, the block diagonal structure of $\mathbf{P}$ is much simpler computationally than the optimal precoder. We will see that the receiver structure of the ZF-BD-GMD system is also much simpler than that of the optimal system.

For the transmitter side to form the transmitted vector $\mathbf{x}_{ZP,K}$, we need multiplications of an $N_T \times N_T$ matrix with an N_T vector K times, which has complexity $O(KN_T^2)$. Compared to $O(K^2N_T^2)$ in the ZF-Optimal system, there will be K times saving.

Now let us look at the receiver side. The lower triangular matrix $\mathbf{L}$ consists of K^2 blocks and each block is an $N_T \times N_T$ matrix. The matrix $\mathbf{Q}$ consists of $(K+L)K$ blocks and each block is an $N_R \times N_T$ matrix. The theorem below shows that both $\mathbf{L}$ and $\mathbf{Q}$ contain many zero entries. The consequence is that compared to the ZF-Optimal system, the saving in the implementation of the feedback paths ($\mathbf{L}$ matrix) is in the order of $O(K)$, and the saving in the implementation of the feedforward filter ($\mathbf{Q}$ matrix) is about a factor of two.

Theorem 4.3.6 *In the ZF-BD-GMD system, $\mathbf{L}$ and $\mathbf{Q}$ both have lower block bandwidth*[3] *L, where L is*

[3]The block bandwidth for a block matrix is defined similarly to the bandwidth defined in p. 152 of [25] original for matrix with scalar entries.

the order of the frequency selective channel. That is, whenever $i > j + L$, the (i, j)th block in $\mathbf{L}$ is a $\mathbf{0}_{N_T \times N_T}$ matrix, and the (i, j)th block in $\mathbf{Q}$ is a $\mathbf{0}_{N_R \times N_T}$ matrix. When K is large, there will be about $K\left((L + 1/2)N_T^2 + N_T/2\right)$ possibly nonzero entries in $\mathbf{L}$, and about $N_T N_R K^2/2$ possibly nonzero entries in $\mathbf{Q}$. ♢

Proof: See Appendix. □

Here we provide a BD-GMD example for $K = 4$ and $L = 1$. We can see that both $\mathbf{L}$ and $\mathbf{Q}$ are block banded matrices with lower block bandwidth L.

$$\underbrace{\begin{bmatrix} \times & \times & & & \\ & \times & \times & & \\ & & \times & \times & \\ & & & \times & \times \end{bmatrix}}_{\mathbf{H}_{ZP,K}^H} = \underbrace{\begin{bmatrix} \times & & & \\ & \times & & \\ & & \times & \\ & & & \times \end{bmatrix}}_{\mathbf{P}} \underbrace{\begin{bmatrix} \times & & & \\ \times & \times & & \\ & \times & \times & \\ & & \times & \times \end{bmatrix}}_{\mathbf{L}} \underbrace{\begin{bmatrix} \times & \times & & & \\ \times & \times & \times & & \\ \times & \times & \times & \times & \\ \times & \times & \times & \times & \times \end{bmatrix}}_{\mathbf{Q}^H}$$

From this theorem we can see the implementation of the ZF-BD-GMD system is much simpler than the ZF-Optimal system. The number of nonzero entries in $\mathbf{L}$ equals the number of feedback paths in the DFE. Therefore, in contrast to the ZF-Optimal case, in which the number of nonzero entries in the feedback matrix is $(K^2 N_T^2 + K N_T)/2$, the saving of the ZF-BD-GMD system is in the order $O(K)$. The number of nonzero entries in $\mathbf{Q}$ equals the number of operations when the signal is passed through the feedforward filter. In contrast to the ZF-Optimal case, in which the number of nonzero entries in $\mathbf{Q}$ is about $N_T N_R K^2$, the ZF-BD-GMD feedforward part saves about half of the operations.

4.4 Transceivers with MMSE DFEs

Suppose some unitary precoder matrix $\mathbf{P}_0$ is used for the channel in (4.4), the MMSE-DFE receiver [12] can be obtained by the QR decomposition of :

$$\begin{bmatrix} \mathbf{H}_{ZP,K}\mathbf{P}_0 \\ \sqrt{\zeta}\mathbf{I}_{KN_T} \end{bmatrix} = \mathbf{Q}_0\mathbf{R}_0, \tag{4.21}$$

where the $((K + L)N_R + KN_T) \times KN_T$ matrix $\mathbf{Q}_0$ has orthogonal columns, and $\mathbf{R}_0$ is a $KN_T \times KN_T$ upper triangular matrix. The feed-forward filter $\mathbf{Q}^H$ is chosen as the first $(K + L)N_R$ columns

of $\mathbf{Q}_0^H$. Under the no-error-propagation assumption, the MMSE-DFE system behaves like KN_T parallel uncorrelated SISO subchannels, and the SINR ρ_k in the kth SISO subchannel is [37, 12]:

$$\rho_k = \frac{1}{\zeta}[\mathbf{R}_0]_{k,k} - 1, \tag{4.22}$$

which depends only on the diagonal terms of $\mathbf{R}_0$. Furthermore, the product of the diagonal terms in $\mathbf{R}_0$ is a constant when the entries of $\mathbf{H}(z)$ are given, i.e.,

$$\begin{aligned} & \prod_{k=1}^{KN_T} [\mathbf{R}_0]_{kk}^2 \\ = \;& \det(\mathbf{R}_0^H \mathbf{R}_0) = \det(\mathbf{R}_0^H \mathbf{Q}_0^H \mathbf{Q}_0 \mathbf{R}_0) \\ = \;& \det \begin{bmatrix} (\mathbf{H}_{ZP,K}\mathbf{P}_0)^H & \sqrt{\zeta}\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{H}_{ZP,K}\mathbf{P}_0 \\ \sqrt{\zeta}\mathbf{I} \end{bmatrix} \\ = \;& \det(\mathbf{H}_{ZP,K}^H \mathbf{H}_{ZP,K} + \zeta \mathbf{I}). \end{aligned}$$

By the Schur-convexity argument [90] of BER function, the optimal $\mathbf{P}_0$ is the one that forces $\mathbf{R}_0$ to have equal diagonal entries, i.e.,

$$[\mathbf{R}_0]_{kk} = \left(\det(\mathbf{H}_{ZP,K}^H \mathbf{H}_{ZP,K} + \zeta \mathbf{I})\right)^{\frac{1}{2KN_T}}$$

for all $k = 1, \cdots, NK_T$. The optimal system can be computed as follows: Consider the GMD [36]

$$\begin{bmatrix} \mathbf{H}_{ZP,K} \\ \sqrt{\zeta}\mathbf{I}_{KN_T} \end{bmatrix} = \mathbf{Q}_g \mathbf{R}_g \mathbf{P}_g^H,$$

where $\mathbf{R}_g$ has equal diagonal terms. Taking $\mathbf{P}_0 = \mathbf{P}_g$, $\mathbf{R}_0 = \mathbf{R}_g$, and $\mathbf{Q}_0 = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_0^H \end{bmatrix} \mathbf{Q}_g$, we are able to arrive at the form (4.21) where $\mathbf{R}_0$ has equal diagonal entries.

Since in general $\mathbf{P}_0$ is a full matrix, the optimal unitary precoder system suffers from the same two implementation problems as in the zero forcing case. Similar to the ZF case, in the following we derive the BD-GMD system for the MMSE counterpart.

Consider the following BD-GMD

$$\begin{bmatrix} \mathbf{H}_{ZP,K}^H & \sqrt{\zeta}\mathbf{I} \end{bmatrix} = \mathbf{P}_{ms} \underbrace{\begin{bmatrix} \mathbf{L}_{ms,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \times & \mathbf{L}_{ms,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \times & \cdots & \times & \mathbf{L}_{ms,K} \end{bmatrix}}_{\mathbf{L}_{ms}} \mathbf{Q}_{ms}^H, \tag{4.23}$$

where $\mathbf{P}_{ms}$ is a block diagonal unitary matrix, $\mathbf{L}_{ms,i}$ is lower triangular matrix with equal diagonal elements, and $\mathbf{Q}_{ms}$ has orthonormal columns. We shall refer to l_i as the channel gain of the effective subchannels, i.e.,

$$l_i \doteq [\mathbf{L}_{ms,i}]_{mm}, \text{ for } m = 1, \cdots, N_T.$$

Then,

$$\begin{bmatrix} \mathbf{H}_{ZP,K}\mathbf{P}_{ms} \\ \sqrt{\zeta}\mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{(K+L)N_R} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{ms}^H \end{bmatrix} \begin{bmatrix} \mathbf{H}_{ZP,K} \\ \sqrt{\zeta}\mathbf{I} \end{bmatrix} \mathbf{P}_{ms} = \underbrace{\begin{bmatrix} \mathbf{I}_{(K+L)N_R} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{ms}^H \end{bmatrix} \mathbf{Q}_{ms}}_{\mathbf{Q}_{new}} \mathbf{L}_{ms}^H,$$

where the second equality is obtained by substituting from (4.23). Note that $\mathbf{Q}_{new}$ has orthonormal columns, and $\mathbf{Q}_{new}\mathbf{L}_{ms}^H$ can be viewed as the QR decomposition of the left hand side.

The proposed MMSE-BD-GMD system is based on this decomposition. The block diagonal unitary matrix $\mathbf{P}_{ms}$ is used as the precoder. At the receiver, the feedforward filter is chosen as the first $(K+L)N_R$ columns of $\mathbf{Q}_{new}^H$, and the feedback filter can be obtained from $\mathbf{L}_{ms}$. The resulting system, which we call the MMSE-BD-GMD system, has effectively KN_T SISO channels, and the SINR ρ_k in the kth subchannel is $\frac{1}{\zeta}[\mathbf{L}_{ms}]_{kk} - 1$. Similar to the ZF-BD-GMD case, the MMSE-BD-GMD system has several good properties, which we will discuss in the following.

First, analogous to Theorem 6.4.1, we can show that the SINR ρ_k in the kth SISO channel of the MMSE-BD-GMD system, is non-increasing. Since $\rho_k = \frac{1}{\zeta}[\mathbf{L}_{ms}]_{kk} - 1$, it is sufficient to show that $[\mathbf{L}_{ms}]_{kk}$ is non-increasing. Suppose the diagonal term in $\mathbf{L}_{ms,m}$ is denoted as l_m, then proving this is equivalent to proving that l_m is non-increasing. Since l_m is obtained from the BD-GMD of

$\left[\mathbf{H}_{ZP,K}^{H} \quad \sqrt{\zeta}\mathbf{I}\right]$, a relation similar to (4.10) can be obtained:

$$l_m = \left(\frac{\det(\mathbf{H}_{ZP,m}^{H}\mathbf{H}_{ZP,m} + \zeta\mathbf{I})}{\det(\mathbf{H}_{ZP,m-1}^{H}\mathbf{H}_{ZP,m-1} + \zeta\mathbf{I})} \right)^{\frac{1}{2N_T}}. \tag{4.24}$$

Since $\mathbf{H}_{ZP,m+1}^{H}\mathbf{H}_{ZP,m+1} + \zeta\mathbf{I}$ can be written as in (4.12) with $\tilde{\mathbf{H}}_0$ replaced by $\tilde{\mathbf{H}}_0 + \zeta\mathbf{I}$, a similar approach to the proof of Theorem 6.4.1 can be used here to show that l_m is non-increasing, i.e.,

$$l_m \geq l_{m+1}.$$

Second, we observe that since $\left(\mathbf{H}_{ZP,m+1}^{H}\mathbf{H}_{ZP,m+1} + \zeta\mathbf{I}\right)$ can be written as in (4.12) with $\tilde{\mathbf{H}}_0$ replaced by $\tilde{\mathbf{H}}_0 + \zeta\mathbf{I}$, we can replace $\tilde{\mathbf{H}}(e^{j\omega})$ with $\left(\tilde{\mathbf{H}}(e^{j\omega}) + \zeta\mathbf{I}\right)$ as well, then apply similar technique as in the proof of Theorem 6.4.3. Analogous to Theorem 6.4.3, by the above approach we can obtain the limit of the subchannel gain. The worst subchannel gain converges to:

$$l \doteq \lim_{K\to\infty} l_K = \exp\left(\frac{1}{4N_T\pi} \int_{-\pi}^{\pi} \log\det(\tilde{\mathbf{H}}(e^{j\omega}) + \zeta\mathbf{I}) d\omega \right). \tag{4.25}$$

Note that this formula should be compared with (4.16). From (4.25), since $\tilde{\mathbf{H}}(e^{j\omega})$ is positive semi-definite, we have

$$\det(\tilde{\mathbf{H}}(e^{j\omega}) + \zeta\mathbf{I}) > \det(\zeta\mathbf{I}) = \zeta^{N_T} \text{ for all } \omega.$$

Thus, it can be shown that $l > \sqrt{\zeta}$. This means that, in the MMSE case even when the channel is close to singular in some frequency band, the effective channel gain will still be greater than $\sqrt{\zeta}$ due to the additional term $\zeta\mathbf{I}$. This is not the case in the zero-forcing systems, which can be clearly seen from (4.16).

Third, for the BER performance of the MMSE-BD-GMD system and the MMSE-Optimal[4] system, we can also show that the asymptotic performance will be the same. This can be seen by the following argument: Because of (4.7) and (4.22), we define

$$P_{ms}(x) = \alpha Q\left(\beta\sqrt{\frac{x^2}{\zeta} - 1}\right), \tag{4.26}$$

[4]Here the MMSE-Optimal system denotes the joint optimal transceiver with MMSE-DFE receiver and linear precoder under the unitary constraint.

which denotes the BER function of the channel gain x for MMSE-DFE systems [12] with noise to symbol power ratio ζ defined in (4.5). Since both systems convert the original system to NK_T independent SISO channels, the average BER depends on the effective subchannel gains. It can be seen that $P_{ms}(x)$ is a continuous and decreasing function of x. Therefore, similar arguments as in the proof of Theorem 4.3.4 can be used here to establish the following equation:

$$\lim_{K\to\infty} \text{BER}_{MSBDGMD}(K) \lim_{K\to\infty} \text{BER}_{MSoptimal}(K) = P_{ms}(l),$$

where l is defined in (4.25). This implies that the MMSE-BD-GMD transceiver is asymptotic optimal when $K \to \infty$.

Fourth, consider the family of systems for zero-padded MIMO frequency selective channels with fixed block size K that use block diagonal unitary precoders and MMSE-DFEs. We can prove that within this family, the MMSE-BD-GMD system is one of the minimizers for the average BER. The proof is similar to that of Theorem 4.3.5 and omitted here.

Finally we consider the implementation cost of the MMSE-BD-GMD system. Since the precoder matrix in MMSE-BD-GMD system is block diagonal, the transmitter implementation cost is the same as the ZF-BD-GMD case. For the receiver part, we can prove that receiver implementation cost is the same as that of the ZF-BD-GMD system. The proof is similar to the proof for Theorem 4.3.6, and is omitted here.

To summarize, the MMSE-BD-GMD system appears to be a more favorable design than the MMSE-Optimal system in that it has much less implementation cost and similar BER performance. If compared to its ZF-BD-GMD counterpart, the MMSE-BD-GMD system has the same implementation cost but better BER performance because it has no zero-forcing constraint.

4.5 Trade-Off between BW Efficiency and Performance

In this section, we will discuss the relationship between bandwidth efficiency and the system performance for four different designs (all with unitary precoders) discussed in this section. These four schemes are: the ZF-BD-GMD system, the ZF-Optimal system, the MMSE-BD-GMD system, and the MMSE-Optimal system. The following theorem summarizes the performances of these four systems when K increases, i.e., BW efficiency increases.

Theorem 4.5.1 *For all of the four mentioned schemes, and for any channel, the average BER is a non-decreasing function in the information block size K. That is, the BER performance degrades as K increases.* $\diamondsuit$

Proof: We will prove the theorem for the two zero-forcing systems. For MMSE case, similar approaches can be applied to the proof and the details are left to the reader.

For the ZF-BD-GMD system, BER is determined by the SISO subchannel gains r_i. In the proof of Theorem 4.3.4, we have shown that β_k in (4.20), which denotes the BER of the ZF-BD-GMD system when block size is K, is a non-decreasing sequence. This proves the theorem for the ZF-BD-GMD part.

For the ZF-Optimal system with information block size K, the effective SISO subchannels have the same channel gain $g_K = (\prod_{i=1}^{K} r_i)^{1/K}$ as in (4.18). Since r_i is non-increasing, we have

$$g_K = (\prod_{i=1}^{K} r_i)^{\frac{1}{K}} \geq (\prod_{i=1}^{K+1} r_i)^{\frac{1}{K+1}} = g_{K+1}.$$

By the fact that $P(\cdot)$ is a decreasing function, we arrive at

$$\text{BER}_{ZFoptimal}(K+1) = P\left((\prod_{i=1}^{K+1} r_i)^{\frac{1}{K+1}}\right) \geq P\left((\prod_{i=1}^{K} r_i)^{\frac{1}{K}}\right) = \text{BER}_{ZFoptimal}(K),$$

which proves the ZF-Optimal case. □

The bandwidth efficiency of a transmission can be defined as in (4.2). Thus, to increase K implies to increase BW efficiency. Therefore, *there exists a clear tradeoff between the bandwidth efficiency and the BER performance for these four systems*. This effect has been observed in the literature (see p.635 and p.636 in [111] for the SISO case). Because of the tradeoff between BER and BW efficiency, one has to choose K carefully to achieve the target BER. An adaptive rate control transmission scheme similar to what is mentioned in [72] can be used here: if the target BER performance is not attained with some K, the receiver asks the transmitter to reduce K to obtain a better BER, but at the expense of losing the bandwidth efficiency.

The best BER is realized when $K = 1$. In this case, the ZF-BD-GMD system becomes equivalent to the ZF-Optimal system (since the block diagonal constraint is no longer active), and the MMSE-BD-GMD system is equivalent to the MMSE-Optimal system. The BER can be expressed as $P\left(\det(\tilde{\mathbf{H}}_0)^{\frac{1}{2N_T}}\right)$ for the ZF systems and $P_{ms}\left(\det(\tilde{\mathbf{H}}_0 + \zeta\mathbf{I})^{\frac{1}{2N_T}}\right)$ for the MMSE systems.

4.6 ZP for SISO Frequency Selective Channel

In this section we discuss the situation when $N_T = N_R = 1$, i.e., the case of the SISO frequency selective channel. The identity precoding matrix, in which the symbols and the padded zeros are directly vectorized without any precoding, will be referred to as the "lazy precoder." A system with lazy precoder and DFE for the ZP-SISO-FS channel is shown in Fig. 4.2.

For SISO frequency selective channels, the zero-padded system with linear precoder and receiver has been discussed and optimally solved [89]. For the lazy precoder (where the precoding matrix is just an identity matrix) and linear ZF or MMSE equalizer case, the tradeoff between bandwidth efficiency and the BER performance has been established in [72]. In the following we will explain that a similar tradeoff occurs when we use a DFE receiver. For $N_T = N_R = 1$ in the ZF-BD-GMD system, each small block $\mathbf{P}_i$ in the block diagonal matrix $\mathbf{P}$ is a 1×1 unitary matrix, which can be expressed as $e^{j\theta_i}$. The BD-GMD can be written as

$$\begin{aligned}
\mathbf{H}_{ZP,K}^{H} &= \begin{bmatrix} e^{j\theta_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & e^{j\theta_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & e^{j\theta_K} \end{bmatrix} \begin{bmatrix} r_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \times & r_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \times & \cdots & \times & r_k \end{bmatrix} \mathbf{Q}^H \\
&= \begin{bmatrix} 1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} r_1 e^{j\theta_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \times & r_2 e^{j\theta_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \times & \cdots & \times & r_k e^{j\theta_K} \end{bmatrix} \mathbf{Q}^H.
\end{aligned}$$

If we use a lazy precoder, the ZF-DFE structure will result in effective K SISO subchannels with channel gains $r_i e^{j\theta_i}$. In this case, the SINR is $|r_i e^{j\theta_i}|^2/\zeta = r_i^2/\zeta$. This is the same as that of the ZF-BD-GMD system. Therefore, all the discussions about the ZF-BD-GMD system performance in Sec. 4.3 continue to hold. Directly following Theorem 4.5.1, we can show that for the system with lazy precoder and ZF-DFE, the same tradeoff also exists. In [72], the author proved the tradeoff exists between the BW efficiency and the BER performance for the lazy precoder with ZF or MMSE linear receiver. Therefore, the point made here can be seen as an extension of [72] to the DFE receiver case.

From Theorem 4.3.4, we know that the ZF-BD-GMD system is asymptotically optimal for ZF-

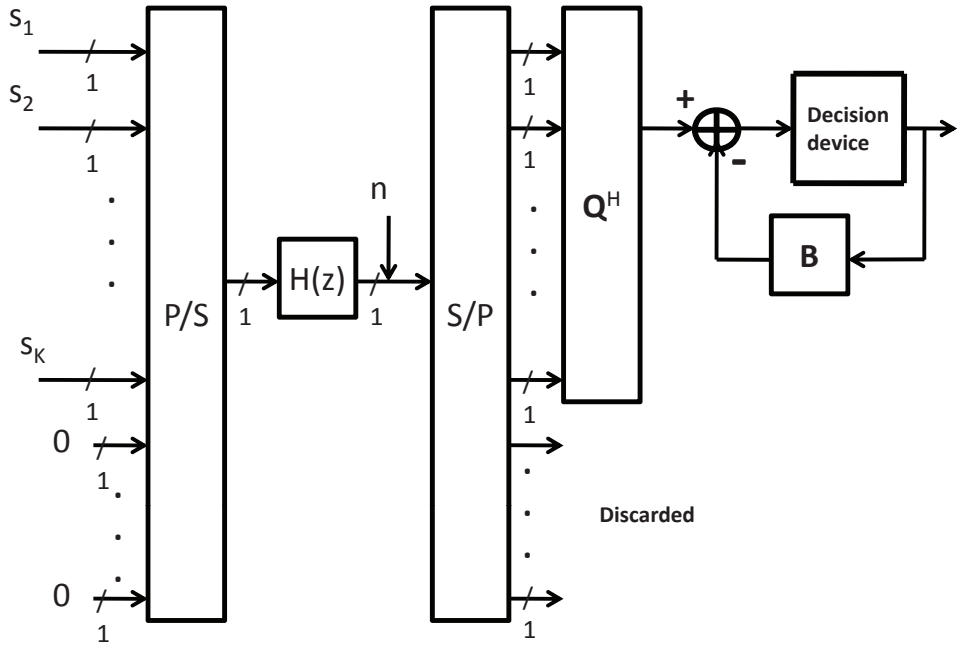


Figure 4.2: The ZP-BD-GMD transceiver for SISO channels. The lazy precoder is used. The vector with signal symbols s_i appended with N_P zeros is passed through a parallel-to-serial converter before transmitting to the channel $H(z)$. The receiver discards the contaminated signal, and passes the clean signal through DFE. $\mathbf{Q}^H$ is the feedforward filter, and $\mathbf{B}$ is the feedback filter.

DFE receiver. Thus, for the SISO frequency selective channel case, *the systems with lazy precoder and ZF-DFE are asymptotically optimal in the class of systems with linear precoders and ZF-DFE receivers*. This property is especially attractive, since the lazy precoder has the practical advantage that the transmitter does not require CSI. Also, it saves the multiplication and addition operations because there is no linear precoding.

In the MMSE-DFE case, where we operate the BD-GMD on $\left[\mathbf{H}_{ZP.K}^H \quad \sqrt{\zeta}\mathbf{I}\right]$, we can also argue similarly that the lazy precoder performance is the same as the MMSE-BD-GMD system. Thus, all the discussions in this Section apply to the MMSE case as well. In particular, *the systems with lazy precoder and MMSE-DFE are asymptotically optimal in the class of systems with linear unitary precoder and MMSE-DFE receiver*.

4.7 Numerical Simulations

In this section, we provide numerical simulations to verify the theoretical results developed in this chapter.

The first example is to plot the subchannel gains of the ZF-BD-GMD and the ZF-Optimal systems for a MIMO ($N_T = 2$, and $N_R = 3$) frequency selective channel $\mathbf{H}_a(z)$ when $K = 5$ and $K = 10$. Here $\mathbf{H}_a(z) = \mathbf{H}_0 + \mathbf{H}_1 z^{-1} + \mathbf{H}_2 z^{-2} + \mathbf{H}_3 z^{-3}$, and the coefficients of these matrices are shown below:

$$\mathbf{H}_0 = \begin{bmatrix} 0.0476 - 0.4556i & -1.1298 + 1.2318i \\ 0.5694 + 0.9440i & 0.4338 - 0.9422i \\ -1.0402 + 0.2657i & 0.7277 + 0.2035i \end{bmatrix}$$

$$\mathbf{H}_1 = \begin{bmatrix} 0.7978 + 1.2002i & 0.2938 - 0.2975i \\ -0.7598 + 0.4878i & -1.0624 + 0.2195i \\ -0.1242 + 0.9503i & 0.6558 - 1.0261i \end{bmatrix}$$

$$\mathbf{H}_2 = \begin{bmatrix} -0.4325 + 0.6604i & -0.9536 + 1.0979i \\ 0.0105 - 0.2053i & -0.2412 - 1.0218i \\ 0.2921 - 0.2946i & 0.8198 + 0.9613i \end{bmatrix}$$

$$\mathbf{H}_3 = \begin{bmatrix} -0.0030 + 0.0512i & -0.2331 + 0.3128i \\ -0.7642 - 0.2375i & -1.1634 - 1.5280i \\ 1.0184 + 0.2687i & -0.2281 + 0.8638i \end{bmatrix}$$

Fig. 4.3 shows the subchannel gains. For the ZF-BD-GMD systems, the first 10 subchannel gains when $K = 5$ are the same as that when $K = 10$. This can be seen from (4.10). The 11th to the 20th subchannel gains when $K = 10$ are smaller than the first 10 subchannel gains. Also, the subchannel gains are non-increasing with the subchannel index. This is consistent with Theorem 6.4.1. For ZF-Optimal systems, all the subchannel gains are identical and equal (see Eq.(4.18)) to the geometric mean of the subchannel gains in the ZF-BD-GMD system. For example, all the subchannel gains equal 5.08 dB in $K = 5$ case. It can be seen that the subchannel gains for $K = 10$ equal 4.92 dB, which is less than that for $K = 5$. This is consistent with Theorem 4.5.1, where the tradeoff exists in the system BER performance and the BW efficiency. We also calculated the value of r defined in

(4.16) by numeric integration. The result is $r = 4.787$dB and is shown in Fig. 4.3. From this figure, we can also see the trend of the subchannel gains as K increases. As K increases, the subchannel gains of the ZF-Optimal system will be lower and closer to the value of r. For $K \to$ infinity, these channel gains should converge to r. This fact was also predicted by Theorem 4.3.4, which states that both the ZF-BD-GMD systems and the ZF-Optimal systems will have BER close to $P(r)$ for very large K.

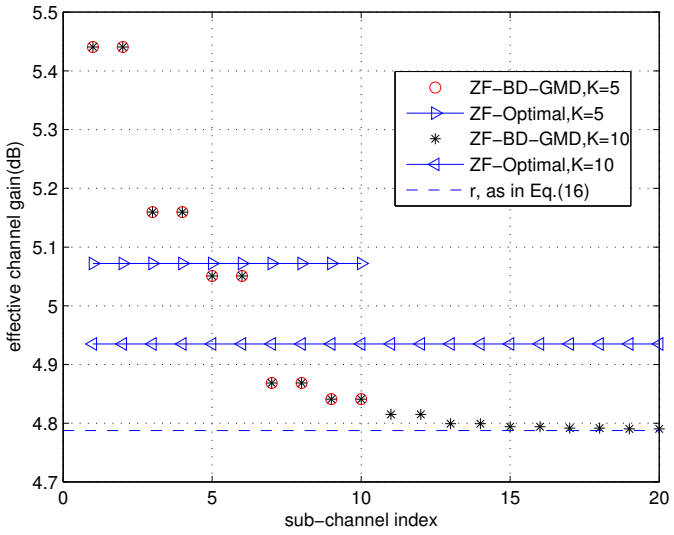


Figure 4.3: The effective channel gain of ZF-BD-GMD and ZF-Optimal transceivers for channel $\mathbf{H}_a(z)$.

In the following simulations, symbols are generated using gray encoded QPSK constellations with symbol power σ_s^2. In each case, 10^3 different channels are used for the Monte Carlo simulations. These channels have the entries coming from i.i.d. complex zero-mean Gaussian distributions with unit variance. The additive channel noise has covariance matrix $\mathbf{R}_n = \mathbf{I}$.

In Fig. 4.4 and Fig. 4.5 we show the average BER simulation results with respect to different values of σ_s^2 for the MIMO systems with $N_T = N_R = 2$, for ZF-DFE and MMSE-DFE case, respectively. The MIMO channels are with order $L = 2$. The zero-forcing system performances for $K = 3$, $K = 10$, and $K = 20$ are shown in Fig. 4.4. The ZF-Optimal system appears to have the best performance for all K. The ZF-DFE system with lazy precoder has about 2 dB loss at BER$= 10^{-5}$ when $K = 3$ compared to ZF-Optimal. However, the ZF-BD-GMD only has about 0.3 dB loss. We can see that the ZF-BD-GMD system indeed has similar performance as the ZF-Optimal system. For larger K, the performance difference between the two systems becomes even smaller.

The MMSE system performance with the same channel setting is shown in Fig. 4.5. Similar conclusions as the ZF case can be made here for MMSE case. From these results, we see that the BD-GMD systems have nearly optimal uncoded BER performance for both ZF and MMSE case, and are much better than the lazy precoder systems. Compared to these two figures, we can see that by using the MMSE-DFE systems, the average BER performance can be a little better than using the ZF-DFE systems. It is important to note that the MMSE-BD-GMD systems have the same implementation cost as the ZF-BD-GMD systems, thus the MMSE-BD-GMD systems is an even better candidate in practice.

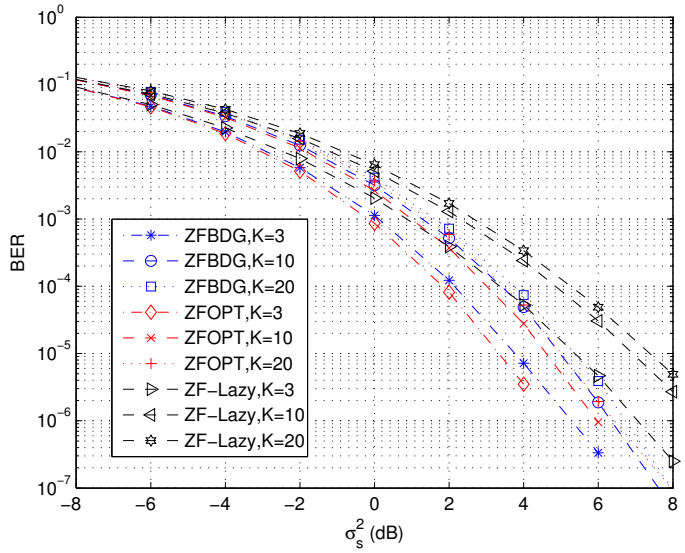


Figure 4.4: The BER performance of the zero-forcing systems for MIMO ($N_T = N_R = 2$) Rayleigh channels of order 3, with $K = 3$, $K = 10$, and $K = 20$. "ZFBDG" represents the ZF-BD-GMD system; "ZFOPT" represents the ZF-Optimal system; and "ZF-Lazy" represents the lazy precoder with zero-forcing DFE.

In Fig. 4.6 and Fig. 4.7 we show the simulation results for case of single transmitting antenna and single receiving antenna. The SISO channels have $L = 2$. The zero-forcing system performances for $K = 3$, $K = 10$, and $K = 20$ are shown in Fig. 4.6. The MMSE system performances for the same channel settings is shown in Fig. 4.7. For large K, lazy precoder case has BER performance almost identical to that of the optimal systems. The simulation results confirm the discussions in 4.6. Also, when K is larger, the BER performance is worse. This confirms the tradeoff between BW efficiency and BER.

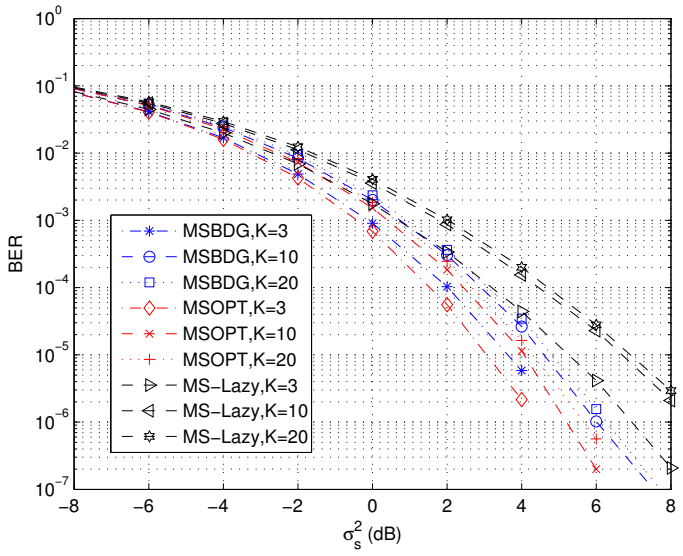


Figure 4.5: The BER performance of the MMSE systems for MIMO ($N_T = N_R = 2$) Rayleigh channels of order 3, with $K = 3$, $K = 10$, and $K = 20$. "MSBDG" represents the MMSE-BD-GMD system; "MSOPT" represents the MMSE-Optimal system; and "MS-Lazy" represents the lazy precoder with MMSE-DFE.

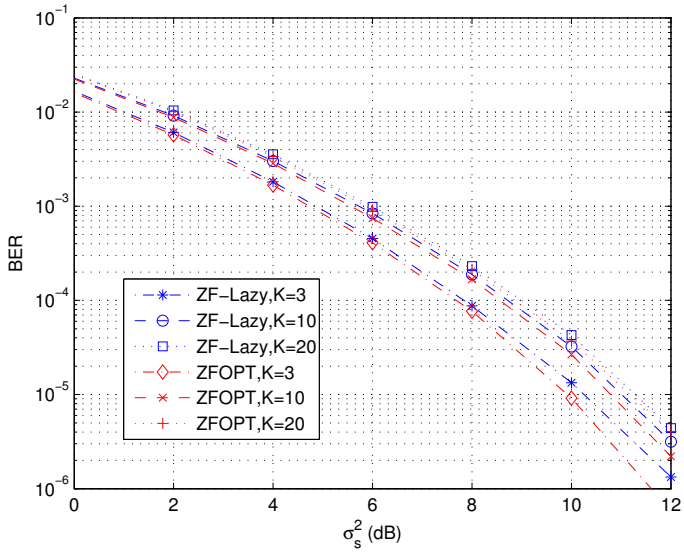


Figure 4.6: The BER performance of the zero-forcing systems for SISO ($N_T = N_R = 1$) Rayleigh channels of order 3, with $K = 3$, $K = 10$, and $K = 20$. "ZFOPT" represents the ZF-Optimal system; and "ZF-Lazy" represents the lazy precoder with zero-forcing DFE.

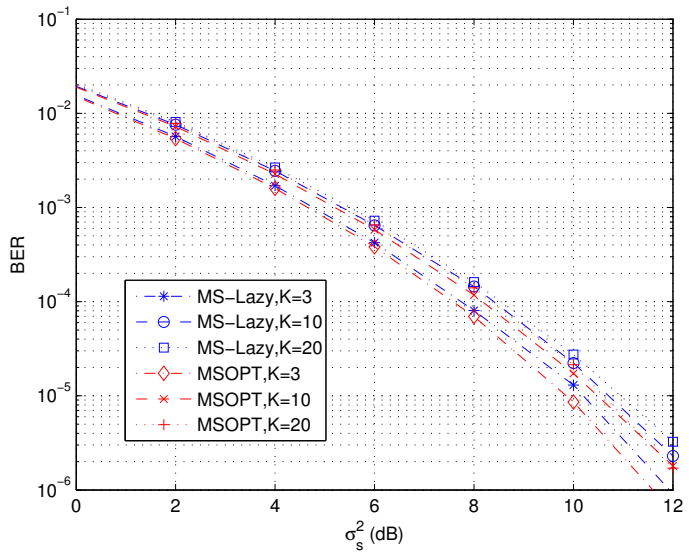


Figure 4.7: The BER performance of the MMSE systems for SISO ($N_T = N_R = 1$) Rayleigh channels of order 3, with $K = 3$, $K = 10$, and $K = 20$. "MSOPT" represents the MMSE-Optimal system; and "MS-Lazy" represents the lazy precoder with MMSE-DFE.

4.8 Concluding Remarks

ZP-BD-GMD transceivers for zero-padded linearly precoded MIMO frequency selective channels have been discussed in this chapter. We proposed the ZF-BD-GMD and the MMSE-BD-GMD systems. Both systems have block diagonal linear precoder matrices and thus simplify the implementation. We showed that the ZP-BD-GMD transceivers have performance similar to the optimal systems when the bandwidth efficiency approaches unity. Thus, both proposed systems appear to be more favorable candidates for practical implementations. We also discussed the tradeoff between the BW efficiency and the BER performance for the ZP-BD-GMD transceivers and the optimal systems. The lazy precoder in SISO channel was also discussed, and it was shown to be asymptotically optimal.

An alternative way to eliminate the IBI in a MIMO FS channel, instead of using zero-padding, is to use cyclic-prefix precoding. This will lead to a MIMO-OFDM system [94]. In this case, if we transmit inverse-FFT-transformed signals and also perform FFT at the receiver, the effective channel becomes a block-diagonal matrix. Thus, if we use block-diagonal precoder for this block-diagonal effective channel, the optimal system design problem reduces to finding the optimal linear precoder and ZF-DFE for the channel coefficients of each carrier, which can be solved by perform-

ing the conventional algorithm (e.g. [90]) on each of them.

4.9 Appendix

4.9.1 Proof of Lemma 4.3.1

Proof: The proof is based on decomposing the matrix as

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^H \\ \mathbf{B} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{B}^H\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{\Delta}_D & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{B} & \mathbf{I} \end{bmatrix},$$

where $\mathbf{\Delta}_D = \mathbf{A} - \mathbf{B}^H\mathbf{D}^{-1}\mathbf{B}$. By theorem 7.7.6 in [31], $\mathbf{\Delta}_D$ is positive definite. We can rewrite the left hand side of (4.14) as

$$\begin{aligned}
& \text{LHS} \\
= \; & \begin{bmatrix} \mathbf{P} & \mathbf{Q} \end{bmatrix} \left(\begin{bmatrix} \mathbf{I} & \mathbf{B}^H\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{\Delta}_D & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{B} & \mathbf{I} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{P}^H \\ \mathbf{Q}^H \end{bmatrix} \\
= \; & \begin{bmatrix} \mathbf{P} & \mathbf{Q} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{D}^{-1}\mathbf{B} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{\Delta}_D^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{B}^H\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{P}^H \\ \mathbf{Q}^H \end{bmatrix} \\
= \; & \begin{bmatrix} \mathbf{P} - \mathbf{Q}\mathbf{D}^{-1}\mathbf{B} & \mathbf{Q} \end{bmatrix} \begin{bmatrix} \mathbf{\Delta}_D^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{P}^H - \mathbf{B}^H\mathbf{D}^{-1}\mathbf{Q}^H \\ \mathbf{Q}^H \end{bmatrix} \\
= \; & (\mathbf{P} - \mathbf{Q}\mathbf{D}^{-1}\mathbf{B})\mathbf{\Delta}_D^{-1}(\mathbf{P} - \mathbf{Q}\mathbf{D}^{-1}\mathbf{B})^H + \mathbf{Q}\mathbf{D}^{-1}\mathbf{Q}^H \\
\succeq \; & \mathbf{Q}\mathbf{D}^{-1}\mathbf{Q}^H,
\end{aligned}$$

where in the last equality we have used the fact that $\mathbf{\Delta}_D$ is positive definite.

□

4.9.2 Proof of Theorem 4.3.5

Proof: Consider a system with block size K that uses the block diagonal unitary precoder $\mathbf{P}' = \text{diag}(\mathbf{P}'_1, \cdots, \mathbf{P}'_K)$. We call this system the $\mathbf{P}'$-BD system. We will prove that the ZF-BD-GMD system has average BER smaller than or equal to that of the $\mathbf{P}'$-BD system.

For the $\mathbf{P}'$-BD system, the corresponding optimal ZF-DFE can be obtained from the QR decom-

position: $\mathbf{H}_{ZP,K}\mathbf{P}' = \mathbf{Q}'\mathbf{L}'^H$, where $\mathbf{L}'$ is a lower triangular matrix. This can be rewritten as

$$\mathbf{H}_{ZP,K}^H = \underbrace{\begin{bmatrix} \mathbf{P}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2' & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{P}_K' \end{bmatrix}}_{\mathbf{P}'} \underbrace{\begin{bmatrix} \mathbf{L}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \times & \mathbf{L}_2' & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \times & \cdots & \times & \mathbf{L}_K' \end{bmatrix}}_{\mathbf{L}'} \mathbf{Q}'^H,$$

where each $\mathbf{L}_k'$ is a lower triangular matrix, and the diagonal entries of $\mathbf{L}_k'$: $[\mathbf{L}_k']_{mm}$ for $k = 1, \cdots, K$, and $m = 1, \cdots, N_T$, are the subchannel gains of the $\mathbf{P}'$-BD system. Similar to Eq. (20) in [47], from the above equation we have

$$\prod_{m=1}^{N_T} |[\mathbf{L}_k']_{mm}|^2 = \det(\mathbf{L}_k'\mathbf{L}_k'^H) = \left(\frac{\det(\mathbf{H}_{ZP,k}^H \mathbf{H}_{ZP,k})}{\det(\mathbf{H}_{ZP,k-1}^H \mathbf{H}_{ZP,k-1})} \right) = r_k^{2N_T}, \tag{4.27}$$

where r_k is defined in (4.10), which stands for the subchannel gain of some subchannels in the ZF-BD-GMD system.

First we note that the SINRs in the subchannels for both the ZF-BD-GMD system and the $\mathbf{P}'$-BD system relate to the subchannel gains by (4.8). Second, the Schur-convex function preserves the partial ordering of majorization [65]. Therefore, to prove this theorem, what we need to prove is that the vector consisting of the absolute values of the subchannel gains of the $\mathbf{P}'$-BD system multiplicatively majorizes the vector consisting of the subchannel gains of the ZF-BD-GMD system. The following is devoted to prove this relation.

Let a_p be defined as the product of the largest p absolute values of the subchannel gains of the $\mathbf{P}'$-BD system, and let b_p be defined as the product of the largest p subchannel gains of the ZF-BD-GMD system. What is left to be proved is the majorization relation:

$$a_p \geq b_p, \text{ for } p = 1, \cdots, KN_T \tag{4.28}$$

with equality when $p = KN_T$.

Suppose we express integer p as $p = \tau N_T + \sigma$, where τ and σ are the nonnegative integers and $\sigma < N_T$. Since Theorem 6.4.1 tells us that r_i is non-increasing, we have

$$b_p = \left(\prod_{s=1}^{\tau} r_s^{N_T} \right) r_{\tau+1}^{\sigma}. \tag{4.29}$$

Let us form a permutation π' of $\{|[\mathbf{L}'_k]_{mm}|\}$ (which may not be in a non-increasing order) to be:

$$\pi' = [\mathbf{L}_1{}'_{[1]}, \cdots, \mathbf{L}_1{}'_{[N_T]}, \mathbf{L}_2{}'_{[1]}, \cdots, \mathbf{L}_2{}'_{[N_T]}, \cdots],$$

where $\mathbf{L}_k{}'_{[m]}$ denote the mth largest component in $[|[\mathbf{L}'_k]_{11}|, \cdots, |[\mathbf{L}'_k]_{N_T N_T}|]$. Let $\pi'(q)$ denote the qth element in π'. Since a_p is the product taken from the largest p elements in π', we have the following inequality

$$a_p \geq \prod_{q=1}^{p} \pi'(q). \tag{4.30}$$

We note that $r_{\tau+1}$ is the geometric mean of $[\mathbf{L}_{\tau+1}{}'_{[1]}, \cdots, \mathbf{L}_{\tau+1}{}'_{[N_T]}]$, and therefore $[r_{\tau+1}, \cdots, r_{\tau+1}]$ is always multiplicatively majorized [65, 90] by $[\mathbf{L}_{\tau+1}{}'_{[1]}, \cdots, \mathbf{L}_{\tau+1}{}'_{[N_T]}]$. The consequence is that

$$\left(\prod_{t=1}^{\sigma} \mathbf{L}_{\tau+1}{}'_{[t]} \right) \geq r_{\tau+1}^{\sigma}. \tag{4.31}$$

Observe that

$$\prod_{q=1}^{p} \pi'(q) = \left(\prod_{s=1}^{\tau} r_s^{N_T} \right) \left(\prod_{t=1}^{\sigma} \mathbf{L}_{\tau+1}{}'_{[t]} \right) \geq \left(\prod_{s=1}^{\tau} r_s^{N_T} \right) r_{\tau+1}^{\sigma} = b_p,$$

where the first equality is from (4.27), and the second inequality is from (4.31). Combining the above equations with (4.30), we have now proved that $a_p \geq b_p$ for $p = 1, \cdots, KN_T$. When $p = KN_T$, both a_p and b_p equal $\left(\det(\mathbf{H}_{ZP,K}^H \mathbf{H}_{ZP,K})\right)^{1/2}$, thus (4.28) is proved. This completes the proof.

□

4.9.3 Proof of Theorem 4.3.6

Proof: Let us reproduce the BD-GMD equation for the zero-padded systems:

$$\mathbf{H}_{ZP,K}^H = \mathbf{PLQ}^H.$$

Now let us first prove $\mathbf{Q}$ is block banded with lower block bandwidth L. Since $\mathbf{P}$ is a unitary block diagonal matrix, and $\mathbf{L}$ is a lower triangular matrix, $\mathbf{PL}$ will be a block lower triangular matrix. Since the diagonal blocks of $\mathbf{L}$ have full rank, all the diagonal blocks in $\mathbf{PL}$ will have full rank as well.

Let us now look at the first row of $\mathbf{H}_{ZP,K}^H$. The $(1,j)$th block in $\mathbf{H}_{ZP,K}^H$ is the product of the $(1,1)$th block in $\mathbf{PL}$ and the $(1,j)$th block of $\mathbf{Q}^H$. For $j > 1+L$, the $(1,j)$th block in $\mathbf{H}_{ZP,K}^H$ is a zero matrix. In this case, since the $(1,1)$th block in $\mathbf{PL}$ has full rank, the $(1,j)$th block of $\mathbf{Q}^H$ must be a zero matrix.

Now let us look at the second row of $\mathbf{H}_{ZP,K}^H$. For $j > 1+L$, since the $(1,j)$th block of $\mathbf{Q}^H$ is zero, the $(2,j)$th block in $\mathbf{H}_{ZP,K}^H$ is the product of the $(2,2)$th block in $\mathbf{PL}$ and the $(2,j)$th block of $\mathbf{Q}^H$. For $j > 2+L$, the $(2,j)$th block in $\mathbf{H}_{ZP,K}^H$ is a zero matrix. In this case, since the $(2,2)$th block in $\mathbf{PL}$ has full rank, the $(2,j)$th block of $\mathbf{Q}^H$ must be a zero matrix. Using a similar argument, we are able to prove that $\mathbf{Q}$ has lower block bandwidth L.

Now let us prove $\mathbf{L}$ is block banded with lower block bandwidth L. The BD-GMD can be re-written as

$$\mathbf{H}_{ZP,K}^H \mathbf{Q} = \mathbf{PL},$$

where we know $\mathbf{Q}$ has lower block bandwidth L. If we look at the (i,j)th block in $\mathbf{PL}$, it is the product of the ith block-row in $\mathbf{H}_{ZP,K}^H$ and the jth block-row in $\mathbf{Q}$. Since $\mathbf{H}_{ZP,K}^H$ is block banded with upper block bandwidth L and $\mathbf{Q}$ is block banded with lower block bandwidth L, the $(i,j)th$ block of $\mathbf{PL}$ will be zero if $i > j+L$. This shows that $\mathbf{PL}$ has lower block bandwidth L. Since $\mathbf{P}$ is a block diagonal matrix, this implies that $\mathbf{L}$ has lower block bandwidth L.

Since $\mathbf{L}$ is a lower triangular matrix, this theorem implies $\mathbf{L}$ is a block banded matrix with $(L+1)$ bands (including the main block diagonal). We can calculate the number of nonzero entries in $\mathbf{L}$: for the main block diagonal, there are $K\left(\frac{N_T^2+N_T}{2}\right)$ nonzero entries; for the lower L bands, there are $((K-1)+(K-2)+\cdots+(K-L))\,N_T^2 = \left(\frac{(2K-L-1)L}{2}\right)N_T^2$ nonzero entries. Thus, the total number of nonzero entries in $\mathbf{L}$ is $K\left(\frac{N_T^2+N_T}{2}\right)+\left(\frac{(2K-L-1)L}{2}\right)N_T^2 \approx K\left((L+1/2)N_T^2+N_T/2\right)$, which grows linearly with K when K is large.

We can also calculate the number of nonzero entries in $\mathbf{Q}$. Note that zero entries are in the lower kth band if $k > L$. Thus, the number of zero entries is $(1+2+\cdots+(K-1))N_TN_R = N_TN_R(K^2-K)/2$. Therefore the number of nonzero entries can be calculated as the number of total entries minus the number of zero entries: $N_TN_R(K+L)K - N_TN_R(K^2-K)/2 \approx N_TN_RK^2/2$ when K is large.

□

Chapter 5

The Role of GTD in Transform Coding

In the first part of this chapter, a general family of optimal transform coders (TC) is introduced based on the generalized triangular decomposition. This family includes the Karhunen-Loéve transform (KLT), and the generalized version of the prediction-based lower triangular transform (PLT) introduced by Phoong and Lin [79], as special cases. The coding gain of the entire family, with optimal bit allocation, is equal to those of the KLT and the PLT. Other special cases of the GTD-TC are the GMD (geometric mean decomposition) coder and the BID (bidiagonal transform) coder. The GMD coder in particular has the property that the optimum bit allocation is a uniform allocation; this is because all its transform domain coefficients have the same variance, implying thereby that the dynamic ranges of the coefficients to be quantized are identical.

The above advantage of the GMD coder is shown to be true in the high bit rate case. However, the performance of the GMD transform coder is degraded in the low rate case. There are mainly two reasons for this degradation. First, the high bit rate quantizer model becomes invalid. Second, the quantization error is no longer negligible in the prediction process when the bit rate is low. In the second part of this section, we introduce dithered quantization to tackle the first difficulty, and then redesign the precoders and predictors in the GMD transform coders to tackle the second. We propose two dithered GMD transform coders: the GMD subtractive dithered transform coder (GMD-SD) where the decoder has access to the dither information and the GMD non-subtractive dithered transform coder (GMD-NSD) where the decoder has no knowledge about the dither. Under the uniform bit loading schemes in scalar quantizers, it is shown that the proposed dithered GMD transform coders perform significantly better than the original GMD coder in the low rate case.

The content of this chapter is mainly drawn from [122, 129], and portions of it have been presented in [121].

5.1 Outline

The chapter will be organized as follows. In Sec. 5.2 we will introduce the GTD transform coder. We will show that the GTD transform coder framework unifies the existing optimal designs, and also predicts many novel optimal structures. Since the theory developed in Sec. 5.2 depends on the validity of high bit rate assumption, we observe some performance degradation for the GMD transform coder when the bit rate is low. In Sec. 5.3 we address this problem by proposing the dithered GMD coder. Two novel designs are proposed, namely, subtractive dithered GMD (SD-GMD) and nonsubtractive dithered GMD coders. Finally, the conclusions are made in Sec. 5.4.

5.2 GTD Transform Coder for Optimizing Coding Gain

In transform coder (TC) theory, the Karhunen-Loéve transform (KLT) is known for its optimality properties [2, 35, 107]. For example it provides maximum coding gain when high bit rate scalar quantizers are used in the transform domain. The KLT essentially diagonalizes the autocorrelation matrix of the input vector $\mathbf{x}$ before quantization. The decorrelated components are typically quantized by independent scalar quantizers.

If the vector $\mathbf{x}$ being transformed is a blocked version of a scalar wide sense stationary (WSS) process $x(n)$, then the coding gain of the KLT can also be achieved by using a different kind of transform called the prediction-based lower triangular transform or PLT, which was introduced into the signal processing literature by Phoong and Lin [79]. The PLT is based on the theory of linear prediction for the scalar WSS process $x(n)$. PLT has smaller design cost because fast algorithms such as the Levinson algorithm can be used instead of matrix diagonalization. The implementation complexity for the PLT is 50% smaller than that of the KLT [79]. However, the PLT as introduced in [79] is in the context of blocked versions of scalar WSS processes only, which is not applicable for general WSS vectors processes.

This section introduces a general family for transform coding based on the generalized triangular decomposition (GTD). We will show that the GTD-TC family has the following features:

1. Unlike the original PLT,[1] the input vector $\mathbf{x}$ is not required to be a blocked version of a WSS process, but when such is the case the complexity of the new transform can be made comparable to that of the original PLT. One of the attractive features of the PLT is the existence of a structure with unit noise gain, called the MINLAB structure [79]. The GTD based family includes a PLT-like special case which also enjoys the MINLAB structure. In this sense it extends some of the features of the PLT for the case where $\mathbf{x}$ is not a blocked version of a scalar process.

2. It includes the KLT and PLT as special cases.

3. The coding gain for any member of the family is equal to that of the KLT.

4. Like the KLT and the PLT the GTD family also produces a decorrelated set of components at the inputs of the scalar quantizers. The GTD offers a great deal of freedom in the distribution of the variances of these decorrelated transform domain components.

5. Other special cases of the GTD transform coder includes the GMD (geometric mean decomposition) and the BID (bidiagonal transform).

6. The GMD in particular has the property that the optimum bit allocation is a uniform allocation. This follows from the fact that all transform coefficients have the same variance (same dynamic range from a practical view point [93]) and thus the same machine word length can be used for all coefficients. Recall here that the closed form formula for optimal bit allocation used by KLT and other transforms [35] often yields non-integer values for the bits. The approximation of these with integers would lead to suboptimality of the transform coder. Since the GMD-based method uses the same number of bits for all the transform domain coefficients without compromising optimality, this disadvantage is not present any more.

The family of GTD coders therefore provides a unified framework for a number of optimal linear transforms for high bit rate coders.

This section is organized as follows. Sec. 5.2.1 briefly reviews the KLT and the PLT. In Sec. 5.2.2 we discuss the proposed GTD-TC. Several examples of the GTD-TC, such as the GMD-TC and BID-TC are given here. The use of GTD in progressive transmission will also be described. Sec. 5.2.3

[1] The original PLT, as introduced in [79], assumes that the input vector $\mathbf{x}$ is a blocked version of a scalar WSS process. The natural extension of the PLT, will be shown to be optimal in terms of coding gain for any stationary vector process (not necessarily a blocked version of a scalar process) with well-defined covariance matrix [33]. This generalization will be referred to as the "PLT." and the restricted one in [79] as the "original PLT" throughout the section.

provides numerical simulations related to the topic discussed. In particular the theoretical claim that the GMD-TC with uniform bit allocation is as good as the KLT with optimal bit allocation is clearly demonstrated in this section.

Assumptions. All signals and transforms discussed in this section are assumed to be real-valued. We assume that the $M \times 1$ input $\mathbf{x}(n)$ is a zero mean real-valued wide-sense stationary vector process, with positive definite covariance matrix $\mathbf{R}_x$. The time argument n is dropped when redundant.

5.2.1 Preliminaries and Reviews

The transform coder is shown in Fig. 5.1. The signal $\mathbf{x}$ is first multiplied by an $M \times M$ matrix $\mathbf{T}$ so that $\mathbf{y} = [y_1\ y_2\ \cdots\ y_M]^T = \mathbf{T}\mathbf{x}$. The quantizers are scalar quantizers, and are modeled as an additive noise sources so that $\widehat{y_i} = y_i + q_i$. Suppose the ith quantizer Q_i has b_i bits, then the variance of the quantization error q_i satisfies

$$\sigma_{q_i}^2 = c2^{-2b_i}\sigma_{y_i}^2, \tag{5.1}$$

where $\sigma_{y_i}^2$ is the variance of the signal input to the ith quantizer. This result generally holds under the high bit rate assumption [35, 64, 107]. The constant c depends on the type of the quantizer and the statistics of y_i. It is assumed that all the scalar quantizers have the same c. The signal is reconstructed at the decoder by multiplying with $\mathbf{T}^{-1}$.

The problem of minimizing the arithmetic mean of MSE (AM-MSE) of the reconstructed coefficients $\mathrm{E}[||\mathbf{x} - \widehat{\mathbf{x}}||_2^2]$, under the average bit rate constraint, is solved by the KLT [107]. The KLT uses $\mathbf{T} = \mathbf{U}^T$, where $\mathbf{U}$ is any $M \times M$ orthonormal matrix such that $\mathbf{R}_x = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$, where $\mathbf{\Sigma}$ is the diagonal matrix of the eigenvalues $\{\sigma_1^2, \cdots, \sigma_M^2\}$ of $\mathbf{R}_x$ (assumed to be in non-increasing order).

Under the high bit rate assumption (5.1), the optimal bit allocation is given by the bit-loading formula [35, 107]

$$b_i = b + \frac{1}{2}\log_2 \frac{\sigma_i^2}{\det(\mathbf{R}_x)^{\frac{1}{M}}}, \tag{5.2}$$

where the average bit rate is constrained to be b bits per data stream. Note that σ_i^2 is actually the signal variance of the transform coefficient y_i. With the bit allocation choosen as in (5.2), the MSE $\sigma_{q_i}^2$ due to the ith quantizer becomes independent of i (as seen by substituting (5.2) into (5.1), with

$\sigma_{y_i} = \sigma_i$.). The resulting AM-MSE is

$$\mathcal{E}_{KLT} = c2^{-2b} \det(\mathbf{R_x})^{\frac{1}{M}}. \tag{5.3}$$

It was shown in [106] that under the high bit rate assumption, it is not a loss of generality to assume that the transform is orthonormal.[2] It should be noted that the KLT decorrelates the signal, so the components of $\mathbf{y}$ are statistically independent (under the Gaussian assumption) [24]. This is a necessary condition for optimality (minimum MSE) under the use of scalar quantizers [35] in the high bit rate case.

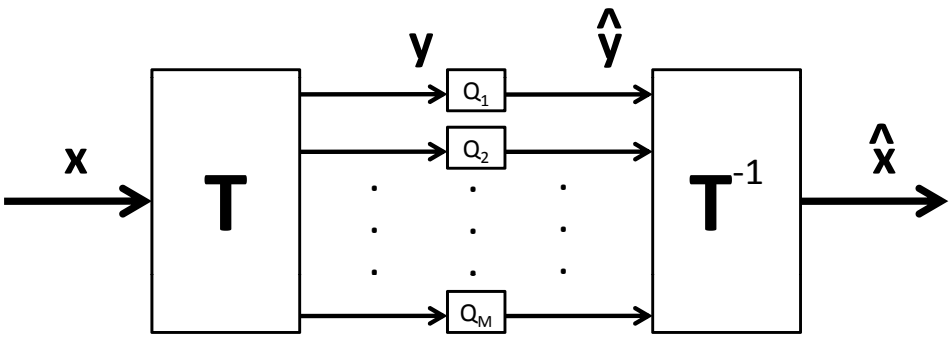


Figure 5.1: Schematic of a transform coder with scalar quantizers.

The PLT, proposed in [79], is a signal dependent non-orthonormal transform, which utilizes linear prediction theory [35, 108]. It has the same decorrelation property as the KLT, and is shown to have the same MMSE performance if the so-called *minimum noise structure* and optimal bit allocation are used [79]. In the article of Phoong and Lin [79], the original PLT is used for the vector $\mathbf{x}$ obtained by blocking a scalar WSS $x(n)$. In the following review of the PLT idea, it can be seen that the PLT can actually be used for a vector process which need not to be a blocked version of a scalar process. The development of [79] which was based on linear prediction theory does not apply in this case, but some of the main conclusions continue to be true as we shall elaborate next.

Consider the LDU decomposition [31] of the covariance matrix $\mathbf{R}_x$ given by

$$\mathbf{R}_x = \mathbf{L}\mathbf{D}\mathbf{L}^T. \tag{5.4}$$

[2]It should be noted that the KLT is optimal among memoryless transforms. If $\mathbf{T}$ is replaced with $\mathbf{T}(z)$ which has memory, then the lapped transform and its variations can be used to further improve the coding performance [1, 62]. In the lapped transform the optimal transform is no longer necessarily orthogonal but biorthogonal [62]. Such transforms are popular in modern practical transform coders [93].

Here $\mathbf{L}$ is lower triangular with diagonal elements equal to unity, and $\mathbf{D}$ is a diagonal matrix with positive diagonal elements. We can rewrite this as

$$(\mathbf{L}^{-1}\mathbf{R}_x^{\frac{1}{2}})(\mathbf{L}^{-1}\mathbf{R}_x^{\frac{1}{2}})^T = \mathbf{D}.$$

That is, $\mathbf{L}^{-1}\mathbf{x}$ has the diagonal covariance matrix $\mathbf{D}$. So premultiplying $\mathbf{x}$ with $\mathbf{L}^{-1}$ results in decorrelation. The transform coder with $\mathbf{T} = \mathbf{L}^{-1}$ will be referred to as the PLT here, and its implementation is shown in Fig. 5.2. The multipliers s_{km} in the figure are the coefficients in the matrix $\mathbf{L}$. In this implementation the quantizer noise is amplified by $\mathbf{T}^{-1} = \mathbf{L}$. A different implementation, called the minimum noise structure I (MINLAB(I)) [82] is shown in Fig. 5.3. At each step of the Minlab encoder as well as the decoder, a prediction is made based on the quantized data, whereas in the structure in Fig. 5.2 the encoder makes predictions based on the original data but the decoder makes predictions with quantized data. This structure is shown to have the unity noise gain property [79]. It minimizes the AM-MSE if the bit loading for each quantizer follows the bit loading formula:

$$b_i = b + \frac{1}{2}\log_2 \frac{\mathbf{D}_{ii}}{\det(\mathbf{R}_\mathbf{x})^{\frac{1}{M}}}. \tag{5.5}$$

Note that $\mathbf{D}_{ii}$ is actually the signal variance of the input to the ith quantizer. By choosing the bit allocation as in (5.5), the MSEs at the output of the quantizers are made identical as seen by using (5.5) in (5.1). The resulting AM-MSE will be

$$\mathcal{E}_{PLT} = c2^{-2b}\det(\mathbf{R}_\mathbf{x})^{\frac{1}{M}}, \tag{5.6}$$

which is the same as what the KLT can achieve when the optimal bit loading is applied. The reason for the name PLT is that the multipliers s_{km} are related to optimal linear predictor coefficients [79] when $\mathbf{x}(n)$ is the blocked version of a scalar WSS process $x(n)$. For simplicity we shall continue to use the term PLT even when this is not the case. The PLT achieves the same optimal performance as the KLT but with less computational complexity in the implementation. Other attractive features are mentioned in [79].

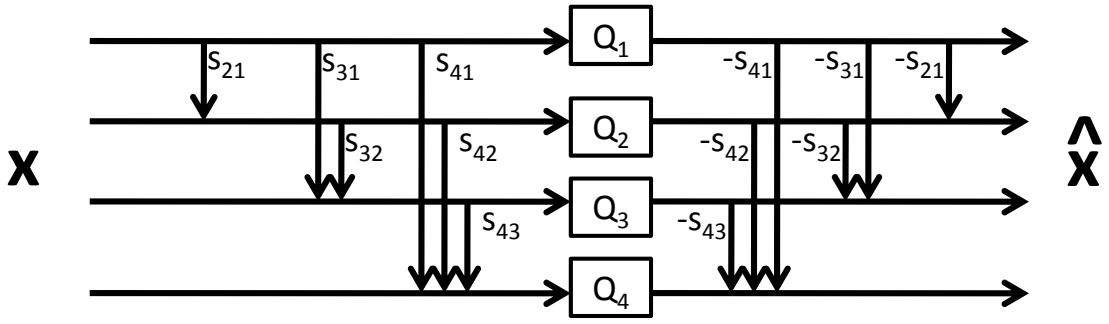


Figure 5.2: A direct implementation of the PLT.

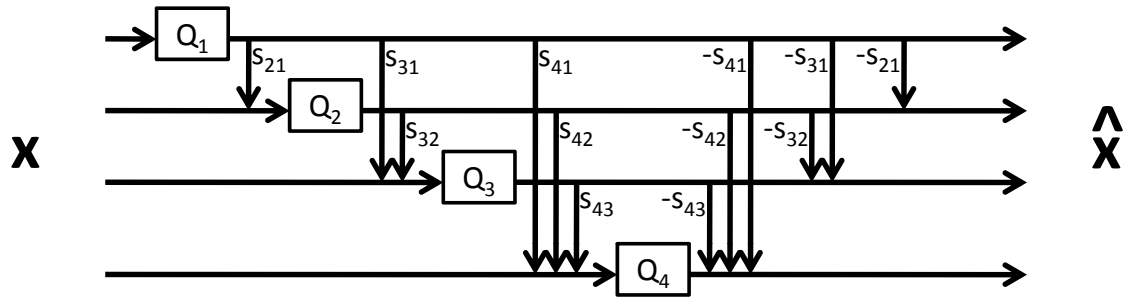


Figure 5.3: The PLT implemented using MINLAB(I) structure.

5.2.2 Generalized Triangular Decomposition Transform Coder

In this subsection we will show how to construct the GTD-TC from a given covariance matrix. We will also show that actually both the KLT and the PLT are special cases of the GTD-TC. Several other interesting instances of GTD-TC, i.e., GMD-TC, BID-TC, and the combination of GMD-TC with progressive transmission, will be discussed.

Consider the transform coding problem again. Suppose the LDU decomposition of $\mathbf{R}_x$ is $\mathbf{R}_x = \mathbf{L}\mathbf{D}\mathbf{L}^T$. Decompose $\mathbf{D}^{\frac{1}{2}}\mathbf{L}^T$ using the GTD, i.e.,

$$\mathbf{D}^{\frac{1}{2}}\mathbf{L}^T = \mathbf{Q}\mathbf{R}\mathbf{P}^T. \tag{5.7}$$

Then we can express $\mathbf{R}_x$ as

$$\begin{aligned}\mathbf{R}_x &= \mathbf{P}\mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R}\mathbf{P}^T \\ &= \mathbf{P}\mathbf{L}_1\text{diag}([\mathbf{R}_{11}^2, \mathbf{R}_{22}^2, \cdots, \mathbf{R}_{MM}^2])\mathbf{L}_1^T\mathbf{P}^T,\end{aligned}$$

where $\mathbf{L}_1$ is a unit-diagonal lower triangular matrix which satisfies

$$\mathbf{L}_1\text{diag}([\mathbf{R}_{11}, \mathbf{R}_{22}, \cdots, \mathbf{R}_{MM}]) = \mathbf{R}^T.$$

Note that because of the GTD theory, the multiplicative majorization property

$$[\mathbf{R}_{11}^2, \mathbf{R}_{22}^2, \cdots, \mathbf{R}_{MM}^2] \prec_\times [\sigma_1^2, \sigma_2^2, \cdots, \sigma_M^2] \tag{5.8}$$

holds, where $[\sigma_1^2, \sigma_2^2, \cdots, \sigma_M^2]$ are the eigenvalues of $\mathbf{R}_x$ with non-increasing order, i.e., $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_M^2$. Note that (5.8) implies the fact that the diagonal terms of $\mathbf{R}$ cannot be arbitrarily chosen, but have to satisfy the multiplicative majorization property.

If we pass the signal $\mathbf{x}$ through the orthonormal matrix $\mathbf{P}^T$ to produce $\mathbf{z}$, i.e., $\mathbf{z} = \mathbf{P}^T\mathbf{x}$, the covariance of $\mathbf{z}$ is

$$\mathbf{R}_z = \mathbf{P}^T\mathbf{R}_x\mathbf{P} = \mathbf{L}_1\text{diag}([\mathbf{R}_{11}^2, \mathbf{R}_{22}^2, \cdots, \mathbf{R}_{MM}^2])\mathbf{L}_1^T.$$

Therefore, $\mathbf{L}_1$ is the lower triangular matrix of the LDU form of $\mathbf{R}_z$. If now apply the PLT $\mathbf{L}_1^{-1}$ to the signal $\mathbf{z}$, the components of the resulting vector are decorrelated. The system is called GTD-TC,

and is demonstrated in Fig. 5.4 for $M = 4$. Here we have used the MINLAB(I) structure [79]. The multipliers s_{km} are the entries of the matrix $\mathbf{L}_1^{-1}$. For example, when $M = 4$,

$$\mathbf{L}_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ s_{21} & 1 & 0 & 0 \\ s_{31} & s_{32} & 1 & 0 \\ s_{41} & s_{42} & s_{43} & 1 \end{bmatrix}.$$

The bit loading formula becomes

$$b_i = b + \frac{1}{2}\log_2 \frac{\mathbf{R}_{ii}^2}{\det(\mathbf{R}_z)^{\frac{1}{M}}} = b + \frac{1}{2}\log_2 \frac{\mathbf{R}_{ii}^2}{\det(\mathbf{R}_x)^{\frac{1}{M}}}, \tag{5.9}$$

where we have used $\det(\mathbf{R}_z) = \det(\mathbf{P}^T\mathbf{R}_x\mathbf{P}) = \det(\mathbf{R}_x)$. Note that the signal variance of the input to the ith quantizer is $\mathbf{R}_{ii}^2$. Again, by using the bit loading formula (5.9), the MSEs of the outputs of the quantizers are identical. This is the same property that the KLT and the PLT have, as introduced in Sec. 5.2.1.

The AM-MSE is invariant to the orthonormal matrix $\mathbf{P}$ at the decoder, therefore the AM-MSE is the same as the one for the PLT part for the transform coding of $\mathbf{z}$. As in eq. (5.6), the MSE is

$$\mathcal{E}_{GTD} = c2^{-2b}\det(\mathbf{R}_\mathbf{z})^{\frac{1}{M}} = c2^{-2b}\det(\mathbf{R}_\mathbf{x})^{\frac{1}{M}}, \tag{5.10}$$

which is the same as the MSE for KLT and PLT with optimal bit allocation. Note that this result is true because of the minimum noise structure for the PLT (which has unit noise gain).

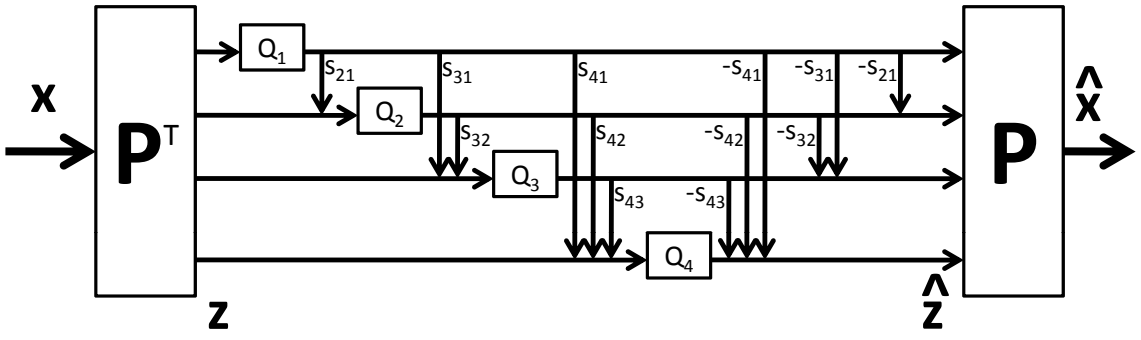


Figure 5.4: The GTD transform coder implemented using MINLAB(I) structure.

We can regard $\mathbf{P}$ and $\mathbf{P}^T$ as the *precoder* and *postcoder*, and the system in between as the PLT part as indicated in the figure. Since there are infinitely many GTD realizations [38], this framework includes many transform coders that achieve the maximized coding gain. Actually it contains both the KLT and the PLT as special cases:

1. Suppose in (5.7), the GTD $\{\mathbf{Q}, \mathbf{R}, \mathbf{P}\}$ is taken as the SVD of $\mathbf{D}^{\frac{1}{2}}\mathbf{L}^T$:

$$\mathbf{D}^{\frac{1}{2}}\mathbf{L} = \mathbf{V}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{U}^T.$$

   In this case, we actually have $\mathbf{R}_x = \mathbf{L}\mathbf{D}\mathbf{L}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$, thus $\mathbf{P} = \mathbf{U}$, which consists of the eigenvectors of the input covariance matrix. We also have $\mathbf{R}_z = \mathbf{U}^T\mathbf{R}_x\mathbf{U} = \mathbf{\Sigma}$. In this case, *the GTD-TC is reduced to the KLT*. The PLT part in Fig. 5.4 is simply a series of scalar quantizers, and the optimal bit loading is according to the formula (5.2).

2. In (5.7), suppose $\{\mathbf{Q}, \mathbf{R}, \mathbf{P}\}$ is taken as the QR decomposition of $\mathbf{D}^{\frac{1}{2}}\mathbf{L}$. Since $\mathbf{D}^{\frac{1}{2}}\mathbf{L}$ is by itself an upper triangular matrix, we actually have $\mathbf{P} = \mathbf{I}$ and $\mathbf{Q} = \mathbf{I}$. *In this case, the GTD-TC reduces to the original PLT-TC.*

In the following, we will introduce three new transform coder schemes based on GTD theory.

Geometric Mean Decomposition – GMD

Suppose the GMD is used for the transform coder: in (5.7), $\mathbf{R}$ has all diagonal terms equal to $\bar{\sigma} = (\prod_{i=1}^{M} \sigma_i)^{\frac{1}{M}}$. The bit loading formula becomes

$$b_i = b + \frac{1}{2}\log_2 \frac{\bar{\sigma}^2}{\det(\mathbf{R}_x)^{\frac{1}{M}}} = b, \tag{5.11}$$

because $\det(\mathbf{R}_x) = \bar{\sigma}^{2M}$. The preceding equation says that all the quantizers are assigned the same number of bits. This is a consequence of the fact that D_{ii} in Eq. (5.5) are identical for all i. That is, the variances of the quantizer inputs are all identical, which means that the dynamic ranges of the signals being quantized are identical. This is a desirable property in practice.

Bi-Diagonal Transformation – Hessenberg Form

A matrix $\mathbf{B}$ is said to be bidiagonal if it has the form demonstrated below for the 4×4 case.

$$\mathbf{B} = \begin{bmatrix} b_{00} & b_{01} & 0 & 0 \\ 0 & b_{11} & b_{12} & 0 \\ 0 & 0 & b_{22} & b_{23} \\ 0 & 0 & 0 & b_{33} \end{bmatrix}.$$

If the GTD form of $\mathbf{D}^{\frac{1}{2}}\mathbf{L}^T$ is $\mathbf{QBP}^T$, where $\mathbf{B}$ is a bi-diagonal matrix, then we call it the bi-diagonal transform coder (BID-TC). It can be seen that

$$\mathbf{R}_x = \mathbf{LDL}^T = \mathbf{PB}^T\mathbf{BP}^T,$$

where $\mathbf{B}^T\mathbf{B}$ is a tri-diagonal matrix demonstrated below for size 4×4:

$$\mathbf{B}^T\mathbf{B} = \begin{bmatrix} c_{00} & c_{01} & 0 & 0 \\ c_{10} & c_{11} & c_{12} & 0 \\ 0 & c_{21} & c_{22} & c_{23} \\ 0 & 0 & c_{32} & c_{33} \end{bmatrix}$$

with $c_{mk} = c_{km}$. This tri-diagonal form $\mathbf{B}^T\mathbf{B}$ is also known as the Hessenberg form [25] of $\mathbf{R}_x$. The advantages of the BID-TC coder lie in its reduced computational complexity. To reduce a symmetric matrix to a tri-diagonal form by orthonormal transformation is computationally much less complex compared to eigenvalue decomposition [25]. The detail of reducing a symmetric matrix to the tri-diagonal form is discussed in [25], and requires only several Householder transformations. The LDU decomposition for a symmetric tri-diagonal matrix is also easy, which requires only $O(M)$ operations now, instead of $O(M^2)$ for general symmetric matrices. Therefore, the design cost for the BID-TC is less than KLT whereas the KLT requires iterative EVD computations. Also, due to the bi-diagonal structure of $\mathbf{B}$, the implementation cost for the inner PLT part is also reduced, which is only in the order of $O(M)$. This can be seen in Fig. 5.5, which shows the MINLAB(I) structure for the BID-TC encoder. Signal feedforward paths are only required for the adjacent data streams. The number of signal feedforward paths is much less than for the original PLT.

The detail comparison between the design and implementation costs for various GTD based coders are summarized in Table 5.1.

Combination of GMD and Progressive Transmission

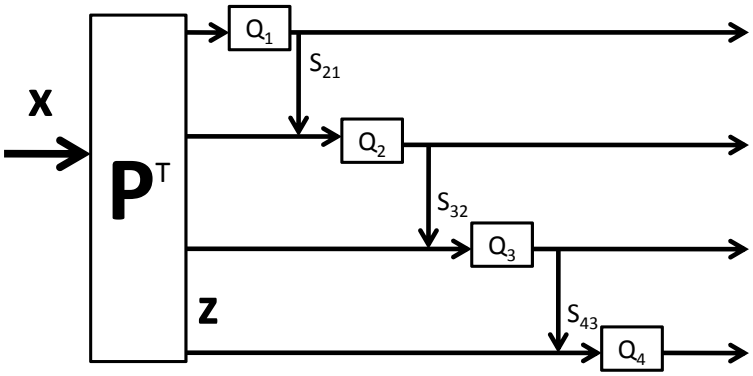


Figure 5.5: The BID Transform coder implemented using MINLAB(I) structure.

Table 5.1: Design and Implementation Costs of Transform Coders

	Design cost	Impl. cost (precoder, PLT)
KLT	EVD, O(M^3)	O(M^2), 0
PLT	LDU, O(M^2)	0, O(M^2)
GMD-TC	EVD and GMD [38], O(M^3)	O(M^2), O(M^2)
BID-TC	Hessenberg form O(M^3) and easy LDU O(M)	O(M^2), O(M)
General GTD-TC	EVD and GTD [38], O(M^3)	O(M^2), O(M^2)

There are some applications where rapid transmission is required and a coarse signal approximation is first produced [61]. When more bits are available, the system progressively enhances the performance by sending more information. Fig. 5.6 shows the example in which we divide the signal data streams after the linear transformation into three groups. The first group is the significant group where the K_1 data streams contain a coarse approximation of the signal. The second group is the less significant group where the K_2 data streams contain detailed information about the signal. The third group of K_3 streams is the least significant group where the remaining $M - K_1 - K_2$ data streams contain components which are close to zero after the linear transformation $\mathbf{P}^T$.

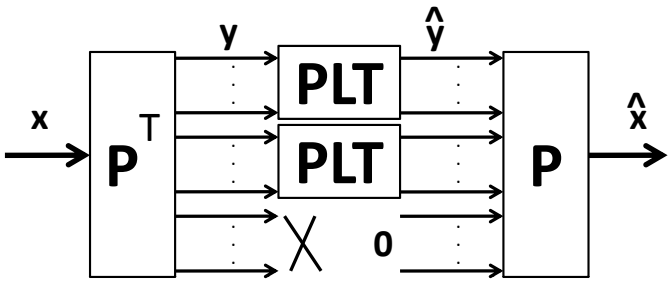


Figure 5.6: Use of GTD-TC in the progressive transmission context.

Suppose we adopt the GTD form in (5.7). We are looking for a transformation such that the diagonal terms of $\mathbf{R}$ have the pattern

$$\text{diag}(\mathbf{R}) = [\bar{\sigma_1}, \cdots, \bar{\sigma_1}, \bar{\sigma_2}, \cdots, \bar{\sigma_2}, \sigma_{K_1+K_2+1}, \cdots, \sigma_M],$$

where

$$\bar{\sigma_1}^2 = (\prod_{i=1}^{K_1} \sigma_i^2)^{\frac{1}{K_1}}, \quad \bar{\sigma_2}^2 = (\prod_{i=K_1+1}^{K_1+K_2} \sigma_i^2)^{\frac{1}{K_2}}.$$

Here $[\sigma_1^2, \cdots, \sigma_M^2]$ are the eigenvalues of $\mathbf{R}_x$ with non-increasing order. $\mathbf{P}$ is the orthonormal matrix obtained from the GTD theory. Note that this decomposition exists for any K_1, K_2 combination, since the multiplicative majorization property holds. Because the eigenvalues are in non-increasing order, the first K_1 substreams actually represent the first K_1 principal components of the vector $\mathbf{x}$, and the next K_2 substreams represent the next K_2 principal components. Suppose for the significant group the total bit budget is $b_1 K_1$, for the less significant group the total bit budget is $b_2 K_2$, and for the least significant group the average number of bits are zero. As shown in Fig. 5.6, for the first and the second group we use the local PLT for each of them. It can be seen that the bit loading formula under the high bit rate assumption will be

$$b_i = b_1 + \frac{1}{2}\log_2 \frac{\mathbf{R}_{ii}^2}{(\prod_{i=1}^{K_1} \sigma_i)^{\frac{1}{K_1}}} = b_1$$

for the first group, and

$$b_i = b_2 + \frac{1}{2}\log_2 \frac{\mathbf{R}_{ii}^2}{(\prod_{i=K_1+1}^{K_1+K_2} \sigma_i)^{\frac{1}{K_2}}} = b_2$$

for the second group. That is, uniform bit loading is used across the quantizers within each group. The data streams in the third group are dropped (i.e., assigned zero bits).

It can be seen that the resulting AM-MSE of this transform coder is

$$\frac{1}{M}(K_1 c 2^{-2b_1} \bar{\sigma_1} \ + K_2 c 2^{-2b_2} \bar{\sigma_2} \ + \Sigma_{i=K_1+K_2+1}^{M} \sigma_i).$$

When we only have a very low bit budget, we can allocate the bits to the first group to get the coarse approximation of the signal. When we have more bits available, the information in the second group is exploited to get the detail information of the signal. Hence the progressive transmission scheme can be implemented when we are able to use uniform quantizers within each group. This

shows one example of the flexibility that our proposed GTD-TC scheme can have. One can have more groups of data streams where each group has a different bit budget.

5.2.3 Simulations

In this subsection we provide the numerical simulations for GTD based coders. The signal $\mathbf{x}$ is generated by a zero-mean Gaussian vector process with prescribed covariance matrix $\mathbf{R}_x$. The number of data streams $M = 8$ in the experiments. Uniform roundoff quantizers are assumed. Each quantizer adapts its step size according to the variance of the Gaussian input (pp. 818 of [107]). We run the simulation for input covariance with high and low condition numbers, respectively. In Fig. 5.7 and Fig. 5.10, the condition number is 10^7. In Fig. 5.8 and in Fig. 5.11 the condition number is 10^3. For each case, we run the Monte Carlo simulations for calculating AM-MSE and root-mean-square error (RMSE, which is the square root of AM-MSE). In each trial, we first generate the input covariance matrix by multiplying a fixed diagonal matrix $\mathbf{\Lambda}_x$ with a randomly generated orthonormal matrix on the left and its transpose on the right. Two choices of $\mathbf{\Lambda}_x$ are used. For the so-called high condition number example,

$$\mathbf{\Lambda}_x = \mathrm{diag}\begin{bmatrix} 10^7 & 10^6 & 10^5 & 10^4 & 10^3 & 10^2 & 10^1 & 1 \end{bmatrix}$$

and for the low condition number example,

$$\mathbf{\Lambda}_x = \mathrm{diag}\begin{bmatrix} 10^3 & 10^3 & 10^2 & 10^2 & 10^1 & 10^1 & 1 & 1 \end{bmatrix}.$$

The input vector $\mathbf{x}$ is then generated according to this covariance matrix. In the following we provide simulation comparisons of different transform coders with and without optimal bit allocation.

Optimal bit allocation: Fig. 5.7 and Fig. 5.8 compare the RMSE performance of different transform coders with optimal bit allocation, for input covariance matrix with high and low condition numbers, respectively. "Transform-wBL" means we adopt the specified transform with optimal bit loading. For example "KLTwBL" uses the KLT with the bit loading formula (5.2). "PLTwBL" is the method mentioned in [79], with the optimal bit loading formula (5.5). "UNCwBL" is the case when we have no transform; we directly quantize the input $\mathbf{x}$ with optimal bit allocation as the bit

loading formula

$$b_i = b + \frac{1}{2}\log_2 \frac{\sigma_{x,i}^2}{(\prod_{k=1}^{M} \sigma_{x,i}^2)^{\frac{1}{M}}}.$$

We perform a rounding operation on the bit loading formula to obtain integer values, and adjust it a little bit to fit the bit budget: First we check if the bit budget is satisfied with equality. If the number of bits is more/less than the bit budget, we decrease/increase 1 bit from the substream with most/least number of bits. We repeat this until the bit budget is satisfied with equality. While suboptimal, we believe this algorithm is not far from optimal in the high bit rate case. Since the input to the quantizers x_i are correlated to each other in general, direct scalar quantization without transformation results in performance loss compared to the GTD-TCs even when the optimal bit loading scheme is applied. "BIDwBL" is the bi-diagonal transform coder. The bit loading formula is as in (5.9). "GMDTFC" is the GMD transform coder. Since the signal variance in each data stream is the same, no bit loading is needed. This allows us to build the same scalar quantizers for all data streams. It can be seen from the figure that with optimal bit loading, all GTD-TCs perform about the same. This is consistent with the analysis made in Sec. 5.2.2. Direct quantization without transforms (UNCwBL) results in about 5 bits per data stream performance loss for Fig. 5.7 and 1.7 bits loss for Fig. 5.8.

Fig. 5.9 plots the coding gain defined as

$$G_{TC} = \frac{MSE_{PCM}}{MSE_{TC}}, \tag{5.12}$$

which is the ratio of the MSE of direct quantization MSE_{PCM} (often referred to as pulse coded modulation) to the MSE of the transform coder MSE_{TC}. It can be seen that the coding gain performance of each method is approximately the same in the high bit rate regime.

Uniform bit allocation: Fig. 5.10 and Fig. 5.11 compare the RMSE performance of different transform coders with uniform bit allocation, for input covariance matrix with high and low condition numbers, respectively. Here "transform-nBL" means we adopt some specific transform with no optimal bit loading, i.e., we allocate the same number of bits to each data stream. However, the step size of each scalar quantizer is adapted according to variance of the Gaussian input (P.818 of [107]). "KLTnBL" uses KLT for the transform. "PLTnBL" is the method mentioned in [79] but with no bit loading. "UNCnBL" is the case when we have no transform but directly quantize the input $\mathbf{x}$.

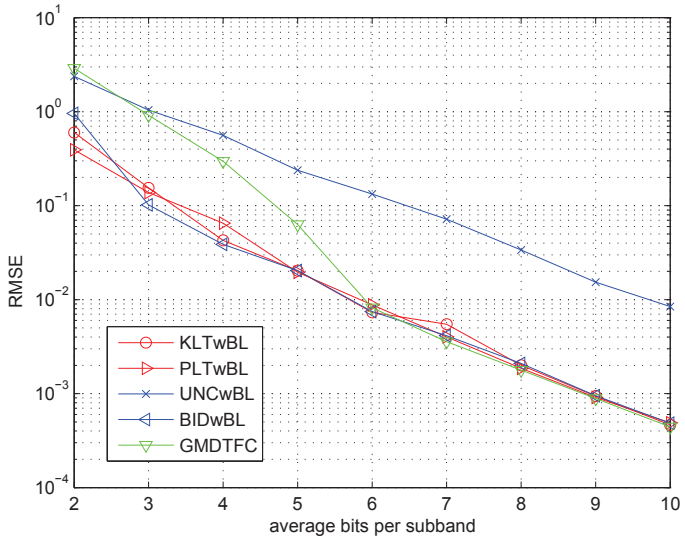


Figure 5.7: Performance of different transform coders with optimal bit allocation. Input covariance matrix has a high condition number (10^7).

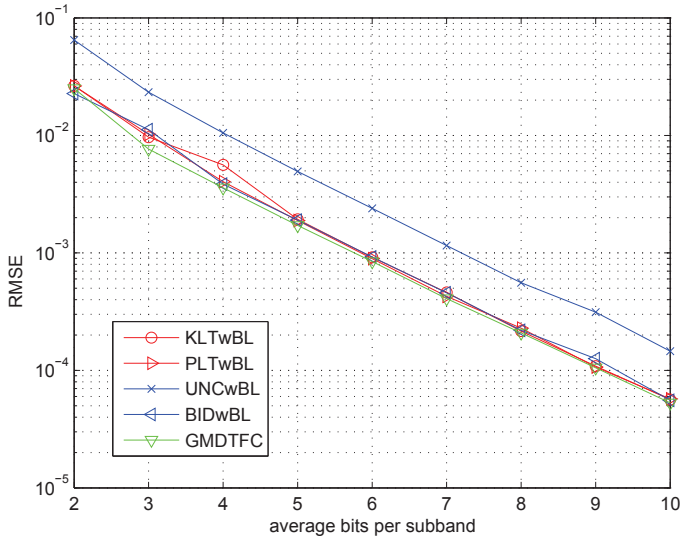


Figure 5.8: Performance of different transform coders with optimal bit allocation. Input covariance matrix has a low condition number (10^3).

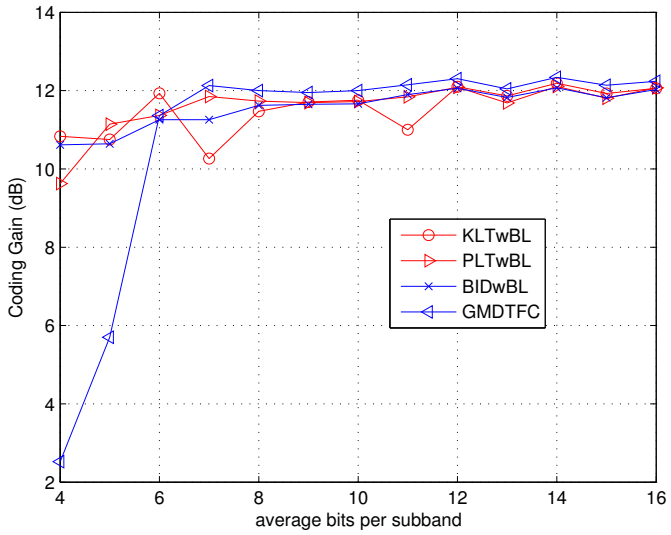


Figure 5.9: Comparison of coding gain of different transform coders with optimal bit allocation. Input covariance matrix has a high condition number (10^7).

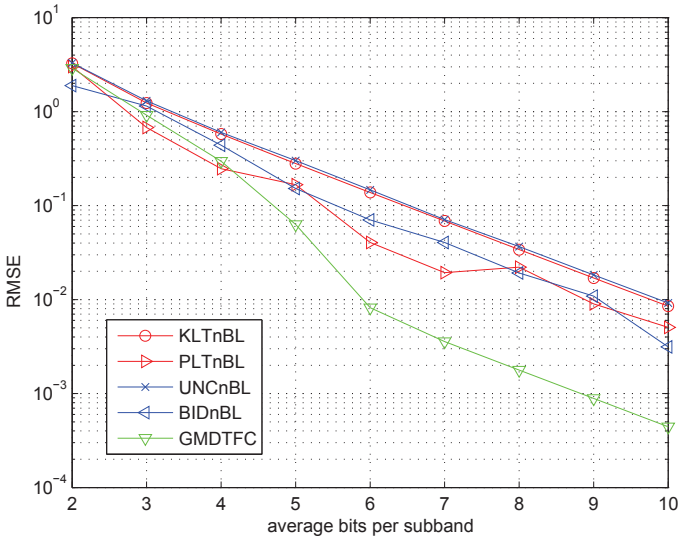


Figure 5.10: Performance of different transform coders with uniform bit allocation. Input covariance matrix has a high condition number (10^7).

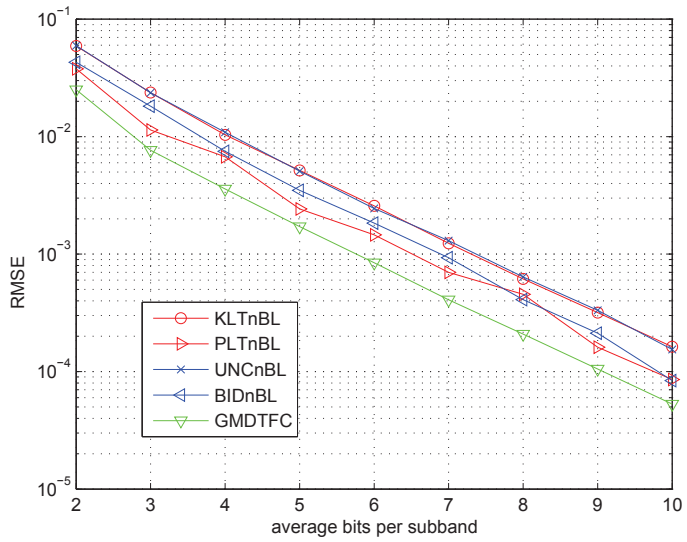


Figure 5.11: Performance of different transform coders with uniform bit allocation. Input covariance matrix has a low condition number (10^3).

No bit loading is applied either. "BIDwBL" is the bi-diagonal transform coder with no bit loading. "GMDTFC" is the GMD transform coder. It can be seen from the figure that with no bit loading applied, GMD performs much better than the other methods, since the GMD without bit allocation is theoretically as good as the other methods with optimal bit allocation.

In the simulation results, the reader will notice that for values of b (average number of bits) exceeding three (the low condition number case), and exceeding six (for the high condition number case), the theoretical predictions are indeed verified to be true. Namely, with no bit allocation, GMD performs much better than KLT, PLT, and the BID. These later methods with no bit allocation have performance comparable to direct quantization. Furthermore, with optimal bit allocation, all these methods (GMD, KLT, and BID) have identical performances. For small values of b [61], these theoretical predictions (which are based on the high bit rate assumption) are seen to be (understandably) less and less true. The low bit rate effect appears to be more severe for the case where the input covariance matrix has high condition number. Also, from the simulations we see that the coding gain improvement of the proposed GTD-TC is more significant for the high condition number case.

5.3 Dithered GMD Transform Coder for Low Rate Applications

In the previous section we showed that all GTD transform coders achieve the maximum coding gain with optimal bit allocation. Several new transform coders in the GTD transform coder family were proposed and shown to have some good properties. In particular, the GMD transform coder yields the maximum coding gain with uniform bit allocation, i.e., no bit allocation is needed.

However, many applications require very low rate transform coders [61]. In those scenarios, the GMD transform coder performs poorly (as can be seen from Fig. 5.10). This is mainly due to two reasons. First, in the high rate case, the uniform quantizer acts like an independent additive noise source. This approximation is no longer valid in the low rate case, where the quantization noise strongly depends on the quantizer input. Second, in the middle-PLT part of the GMD quantizer, the quantized data is used for prediction. However, the prediction coefficients themselves are obtained from the unquantized data. The effect of this mismatch is no longer negligible in the low rate regime.

In this section, we propose two transform coding structures: the GMD subtractive dithered (GMD-SD) transform coder and the GMD non-subtractive dithered (GMD-NSD) transform coder. These two transform coders solve the two difficulties mentioned above in the low rate case. The first difficulty is solved by using dithered quantization [116]. If the decoder has perfect knowledge of the dither signal, the GMD-SD transform coder can be used. In absence of knowledge of the dither signal at the decoder, we propose using the GMD-NSD transform coder. The dither signal is chosen differently in each case in order for the quantization error to be uncorrelated with the quantizer input. The second difficulty is solved by redesigning the prediction coefficients. The predictors are derived from the second order statistics of the quantized data to accommodate the effect of quantization noise. Based on these approaches, we are able to improve the coding performance significantly in the low rate case compared to the original GMD coder.

5.3.1 Dithered GMD Quantizer

The dithered quantizer is shown to render the quantization noise statistics. Suppose that Q is a scalar quantizer with step size Δ. In a dithered quantizer, the sum of the input signal x and the dither w is passed through the scalar quantizer Q, where the dither is independent of the input. The output signals of subtractive and nonsubtractive dithered quantizers are $\hat{x} = Q(x + w) - w$ and $\hat{x} = Q(x + w)$, respectively. Assume that the maximum quantization error is less than $\Delta/2$.

It is shown in [27, 116] that (i) if the dither w is uniformly distributed on $(-\Delta/2, \Delta/2]$ then the quantization error of the subtractive dithered quantizer $Q(x+w)-(x+w)$ is independent of the input signal and is uniformly distributed on $(-\Delta/2, \Delta/2]$, and (ii) if the dither w is the sum of $n \geq 2$ independent random variables uniformly distributed on $(-\Delta/2, \Delta/2]$, then the kth moment of the quantization error of the nonsubtractive dithered quantizer $Q(x+w)-x$ is independent of the input distribution for $k = 1, 2 \ldots, n$, and its variance is equal to $(n+1)\Delta^2/12$. In this subsection, we propose to use dithered quantization along with the GMD coder. Also, a design method for the predictors in the PLT structure used in the GMD coder is proposed to incorporate the low rate mismatch.

We consider transform coder structures with dithered quantizers combined with the GMD coder. Fig. 5.12 shows an example of the structure of a GMD subtractive dithered (GMD-SD) coder with $M = 3$. The dither w_i is added to the input of uniform quantizer Q_i for $i = 1, 2, \ldots, M$. After the uniform quantizer, the dither is subtracted from the quantized signal. The resulting signal is then multiplied with the predictor coefficients $s_{1i}, s_{2i}, \ldots, s_{Mi}$ for the use of following substream quantizers. The quantized signal $\widehat{y}_i$ is then stored or transmitted to the decoder side. The subtractive dither quantizer assumes the decoder has knowledge about the dither signal. In the decoder, the dither is first subtracted. The resulting signal then undergoes the inverse operation of the prediction process and the $M \times M$ matrix $\mathbf{P}$, which yields the reconstructed signal $\widehat{\mathbf{x}}$.

In the non-subtractive dithered quantizer, the dither knowledge is lacking in the decoder. Fig. 5.13 shows the structure of GMD non-subtractive dithered (GMD-NSD) coder. In the encoder of GMD-NSD, the quantized signal is directly multiplied with the predictor coefficients for the use of following substream quantizers, without being first subtracted from the dither. For both cases, the transform coder can be modeled as in Fig. 5.14, where the ith dithered quantizer is modeled as an additive noise source n_i. Note that the statistics of noise source n_i are different in these two cases.

In the following, we will use the model in Fig. 5.14 to design the predictor coefficients. The transformed signal $\mathbf{z}$ is passed through a prediction-based lower triangular transform coder [79] implemented using the MINLAB structure [82]. The resulting encoded signal is $\widehat{\mathbf{y}} = \mathbf{S}(\mathbf{P}^T\mathbf{x} + \mathbf{n})$, where $\widehat{\mathbf{y}} = [\widehat{y}_1, \widehat{y}_2, \cdots, \widehat{y}_M]^T$, $\mathbf{n} = [n_1, n_2, \cdots, n_M]^T$, and the prediction matrix $\mathbf{S}$ is a unit-diagonal lower triangular matrix that consists of prediction coefficients in the MINLAB structure. The covariance matrix of $\widehat{\mathbf{y}}$ can be written as

$$\widehat{\mathbf{R}}_y = \mathbf{S}E[(\mathbf{z}+\mathbf{n})(\mathbf{z}+\mathbf{n})^T]\mathbf{S}^T = \mathbf{S}(\mathbf{R}_z + \mathbf{R}_n)\mathbf{S}^T,$$

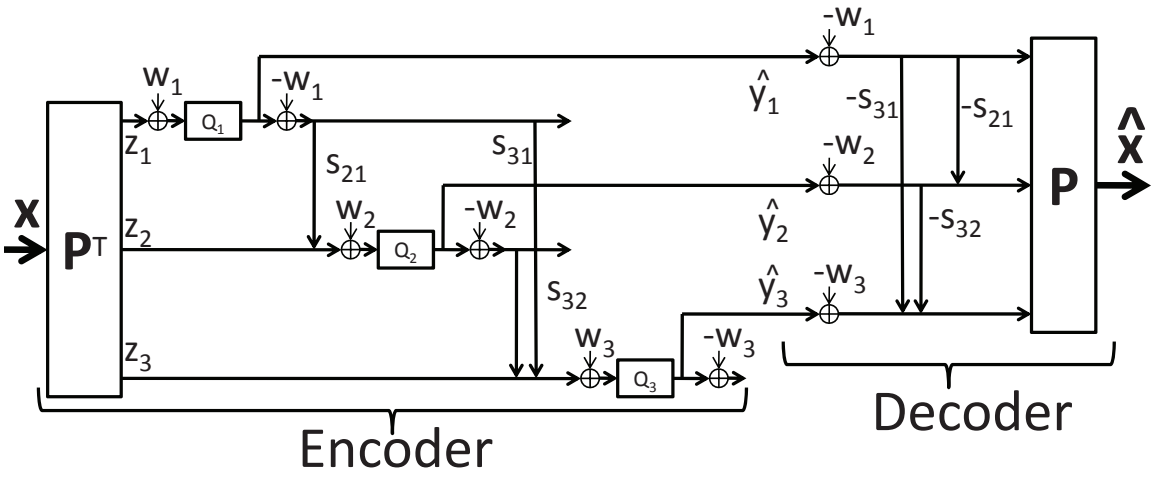


Figure 5.12: Subtractive dithered GMD transform coder.

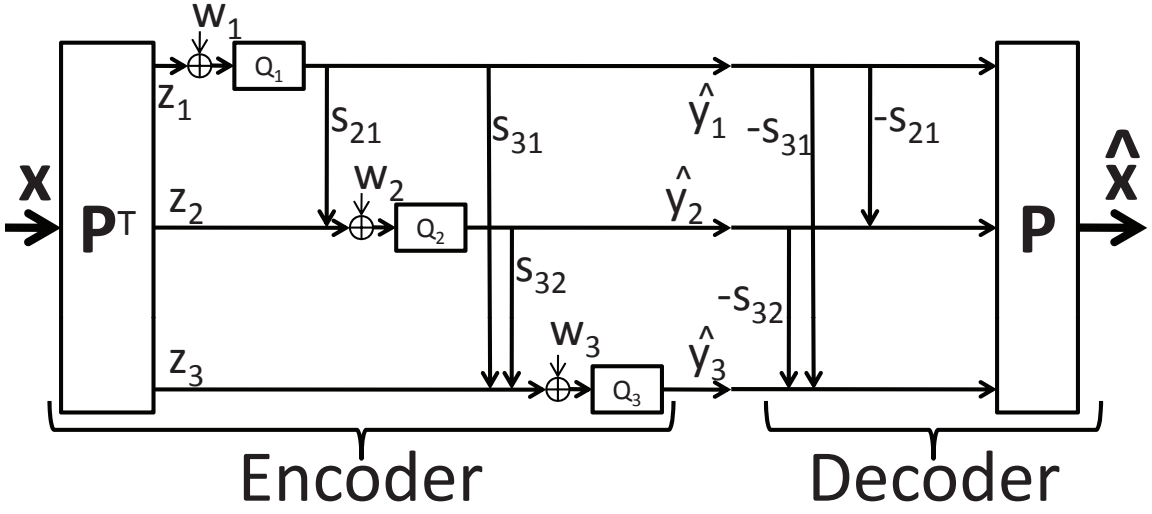


Figure 5.13: Nonsubtractive dithered GMD transform coder.

where we assume the noise samples are uncorrelated with the signal $\mathbf{z}$. Therefore, the MMSE prediction matrix can be obtained by viewing the signal covariance matrix of the middle-part (PLT part) as $\mathbf{R}_z + \mathbf{R}_n$ instead of $\mathbf{R}_z$. By similar derivation as in Sec. III of [79], the MMSE lower triangular prediction matrix can be written as

$$\mathbf{S} = \mathbf{L}_1^{-1}, \tag{5.13}$$

where $\mathbf{L}_1$ is the unit-diagonal lower triangular matrix in the LDU decomposition of $\mathbf{R}_z + \mathbf{R}_n$:

$$\mathbf{R}_z + \mathbf{R}_n = \mathbf{L}_1 \mathbf{D} \mathbf{L}_1^T. \tag{5.14}$$

Suppose that the noise variance depends only on the quantizer step size, and the noise samples are uncorrelated with each other.This is achievable by using a uniformly distributed dither for subtractive dithered quantizer and a triangular-pdf dither for nonsubtractive dithered quantizer (see [27, 116]). We will first assume that the step size of each quantizer can be made the same (same step-size rule). This implies the signal substream to each quantizer has the same variance, or equivalently $\widehat{\mathbf{R}}_y = d\mathbf{I}$, where d is some constant. Later we will prove that this is possible without bit allocation, but with the aid of properly designed precoder $\mathbf{P}^T$. Under this same step size assumption, the noise covariance matrix $\mathbf{R}_n = \sigma^2\mathbf{I}$, where σ^2 is the noise variance that depends on the step size. With the use of prediction matrix $\mathbf{S}$ in (5.13), the covariance matrix of the encoded signal $\widehat{\mathbf{y}}$ is $\widehat{\mathbf{R}}_y = \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix. The question now is whether there exists an orthogonal matrix $\mathbf{P}$ so that $\widehat{\mathbf{R}}_y = d\mathbf{I}$. The following theorem asserts the existence of such $\mathbf{P}$.

Theorem 5.3.1 *Suppose* $\mathbf{A}$ *is some Cholesky factor of* $\mathbf{R}_x$*, i.e.,* $\mathbf{R}_x = \mathbf{A}\mathbf{A}^T$*. Consider the geometric mean decomposition*

$$\begin{bmatrix} \mathbf{A}^T \\ \sigma\mathbf{I} \end{bmatrix} = \mathbf{Q}\mathbf{R}\mathbf{P}^T, \tag{5.15}$$

where $\mathbf{R}$ *has equal diagonal entries* $\mathbf{r} = [r\ r \cdots\ r]$*. If the precoder* $\mathbf{P}^T$ *in Fig. 5.14 is taken as the one in (5.15), then* $\widehat{\mathbf{R}}_y = \mathbf{D} = d\mathbf{I}$ *for some constant* d*.* ♢

Proof: From (5.14), by using the MMSE prediction matrix $\mathbf{S} = \mathbf{L}_1^{-1}$, the covariance matrix $\widehat{\mathbf{R}}_y = \mathbf{D}$. Thus we only need to prove that the LDU decomposition of $\mathbf{R}_z + \sigma^2\mathbf{I} = \mathbf{P}^T\mathbf{R}_x\mathbf{P} + \sigma^2\mathbf{I}$

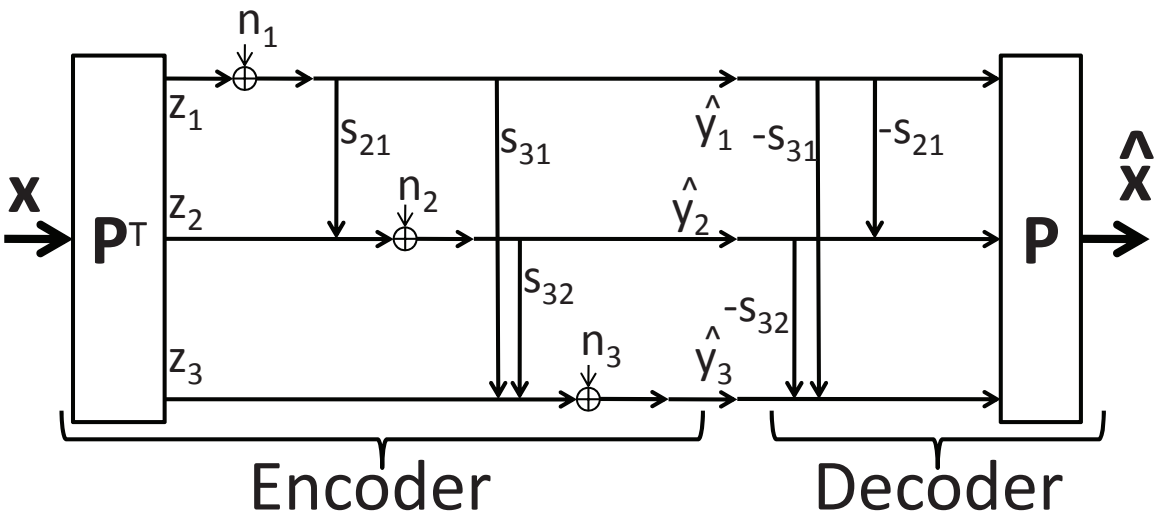


Figure 5.14: The equivalent model of dithered GMD transform coder.

has $\mathbf{D} = d\mathbf{I}$. To prove this, observe that

$$\begin{aligned}
\mathbf{P}^T\mathbf{R}_x\mathbf{P} + \sigma^2\mathbf{I} &= \mathbf{P}^T \begin{bmatrix} \mathbf{A} & \sigma\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}^T \\ \sigma\mathbf{I} \end{bmatrix} \mathbf{P} \\
&= \mathbf{P}^T\mathbf{P}\mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R}\mathbf{P}^T\mathbf{P} \\
&= \mathbf{R}^T\mathbf{R} = \mathbf{L}_1\mathbf{D}\mathbf{L}_1^T,
\end{aligned}$$

where $\mathbf{L}_1$ is taken as $r^{-1}\mathbf{R}^T$, and $\mathbf{D} = r^2\mathbf{I}$. This completes the proof. □

This theorem suggests a method for finding the precoder $\mathbf{P}^T$. With such $\mathbf{P}^T$ as the precoder and $\mathbf{S}$ in (5.13) as the prediction matrix, we are able to have $\widehat{\mathbf{R}}_y = r^2\mathbf{I}$. Therefore, the scalar quantizers of the dithered GMD transform coder can use uniform step size without bit allocation.

The design procedure for the GMD-SD (or GMD-NSD) transform coder, given the input covariance matrix, is summarized in the following:

1. Determine the uniform quantizer step size Δ according to the bit rate.
2. Determine σ from the quantizer step size, e.g., $\sigma = \sqrt{\Delta^2/12}$ for GMD-SD using the uniform pdf dither, and $\sigma = \sqrt{\Delta^2/4}$ for GMD-NSD using the triangular pdf dither.
3. Compute the Cholesky factor $\mathbf{A}$ of $\mathbf{R}_x$: $\mathbf{A}\mathbf{A}^T = \mathbf{R}_x$.
4. Compute the geometric mean decomposition as in (5.15).
5. Compute the LDU decomposition: $\mathbf{P}^T\mathbf{R}_x\mathbf{P} + \sigma^2\mathbf{I} = \mathbf{L}_1\mathbf{D}\mathbf{L}_1^T$, and take $\mathbf{S} = \mathbf{L}_1^{-1}$.

6. Construct the transform coder structure using $\mathbf{P}^T$ and $\mathbf{S}$ as in Fig. 5.12 and Fig. 5.13 for GMD-SD and GMD-NSD transform coder, respectively.

The complexity of the successive decompositions in the above algorithm is in the same order as that in the GMD transform coder described in [122]. A complete comparison of the design and implementation costs of different transform coders can be seen in the previous section.

5.3.2 Numerical Example

In this subsection we provide numerical comparisons of several transform coders. The signal $\mathbf{x}$ is generated by a zero-mean Gaussian vector process with prescribed covariance matrix $\mathbf{R}_x$. The number of data streams $M = 8$ in the experiments. Uniform roundoff quantizers are assumed. Each quantizer adapts its step size according to the variance of the Gaussian input (pp. 818 of [107]). For each case, we run the Monte Carlo simulations for calculating the arithmetic mean of mean square error (AM-MSE). In each trial, we first generate the input covariance matrix by multiplying a fixed diagonal matrix $\mathbf{\Lambda}_x$ with a randomly generated orthogonal matrix on the left and its transpose on the right. The input vector $\mathbf{x}$ is then generated according to this covariance matrix. Fig. 5.15 shows the AM-MSE performance of different transform coders for $\mathbf{\Lambda}_x = 10^{-5} \times \text{diag}([10^7\ 10^6\ \cdots\ 10^0])$. The following five methods use the same number of bits in each of their quantizers: "KLT," which uses Karhunen-Loéve transform; "PLT," which uses the original structure proposed in [79]; "Uncoded," which directly quantizes the signal; "BID," which uses the bi-diagonal decomposition; "GMD," which uses the geometric mean decomposition. The KLT and the PLT with optimal bit allocation but no dithering (represented as "KLTwBL" and "PLTwBL," respectively) are also simulated for comparison. It should be noted that under the high bit rate assumption, "GMD," "KLTwBL," and "PLTwBL" coders achieve the same minimum AM-MSE performance.

The performances of the two new stuctures, "GMD-SD" and "GMD-NSD," are also shown in Fig. 5.15. In "GMD-SD," the dither signal is generated from the uniform pdf that satisfies Schuchman's condition [27]. In particular, it exhibits a variance of $\Delta^2/12$, where Δ is the step size. In "GMD-NSD," the dither signal is generated from triangular-pdf (Sec. III.C. in [116]). This dither pdf renders both the first and second moments of the total error independent of the quantizer input. In particular, it is the unique choice of zero-mean dither pdf which renders the first two moments of the total error independent of the input while minimizing the second [116]. The variance of the error is $\Delta^2/4$. These two dithered GMD quantizers are designed by the method described in Sec.

5.3.1.

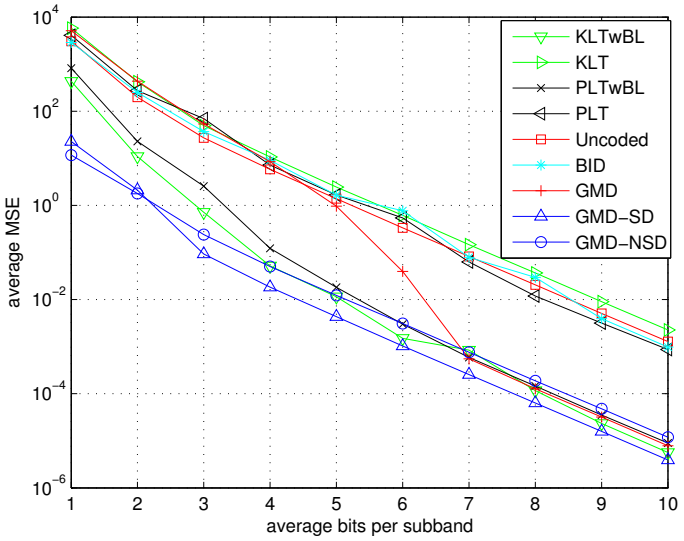


Figure 5.15: Performance of different transform coders.

From Fig. 5.15, we see that in the low rate regime, the two proposed dithered transform coders have better performance than all the other transform coders. The two optimal-bit-allocated coders, "KLTwBL" and "PLTwBL," also perform worse than the two dithered transform coders due to the collapse of the high rate assumption. It can also be seen that in the extremely low rate case (one and two bits), there is a performance degradation in "GMD-SD" and "GMD-NSD." This is because at such low rate, the step size of the dither signal is too large, which makes the chance of overflow much higher for Gaussian sources and violates the non-overflow assumption in the dither quantizer theory [27]. In the high rate regime, the two proposed dithered GMD transform coders perform comparably with the three coders ("KLTwBL", "PLTwBL", and "GMD") which are designed under the high rate assumption. The AM-MSE of "GMD-NSD" is approximately three times of the AM-MSE of "GMD-SD". For fixed AM-MSE, "GMD-NSD" needs to have about one more bit than "GMD-SD" needs. The AM-MSE of "GMD" is between those two. The results of this example suggest the following: If the dither signal is available both at the decoder, then "GMD-SD" is a better candidate since it has the lowest AM-MSE in the low rate and high rate regimes; if the dither is not available in the decoder, then we should use the "GMD" coder in the high rate regime, and use "GMD-NSD" in the low rate regime.

5.4 Concluding Remarks

In this chapter we have shown the use of GTD in transform coding problems. In the first part of this chapter, a general family of optimal transform coders (TC) was introduced based on the GTD. The coding gain of the entire family, with optimal bit allocation, is maximized. This family includes KLT and PLT coders as special cases. Moreover, it can predict many novel structures which achieve the same maximized coding gain. This shows the power of GTD in transform coding. One thing to note for practical use is that, in situations involving the KLT, the discrete cosine transform (DCT) is often used instead of the KLT. This is because the DCT is signal independent, computationally efficient, and a good approximation of the KLT for a large class of signals with low-pass spectra [62]. An analogous low-complexity approximation for the precoder $\mathbf{P}$ which arises in the GTD implementation is not known at this time.

In the second part of this chapter we addressed the case of low bit rate coding using the dithered GMD coder. We proposed two dithered GMD transform coders: the GMD subtractive dithered transform coder (GMD-SD) where the decoder has access to the dither information and the GMD non-subtractive dithered transform coder (GMD-NSD) where the decoder has no knowledge about the dither. Both of these coders use uniform bit loading schemes. We have shown that the proposed dithered GMD transform coders perform significantly better than the original GMD coder in the low rate case.

Chapter 6

The Role of GTD in Filter Bank Optimization

The M-channel filter bank (FB) and subband coder (SBC) are commonly used in many signal processing applications [62, 107, 114]. During the past several years, there has been a great deal of interest in the theory and design of optimal signal adapted orthonormal and biorthogonal filter banks in terms of theoretical coding gain criterion [1, 20, 58, 69, 70, 97, 106, 109]. The theoretical optimal SBC problem can be stated as follows: If the average bit rate b is given, what is the best choice of analysis/synthesis filters $\{H_i(e^{j\omega}), F_i(e^{j\omega})\}$ (equivalently, the polyphase matrices $\{\mathbf{E}(e^{j\omega}), \mathbf{R}(e^{j\omega})\}$) that minimizes the average reconstruction error? Two well-known special cases of this general scenario are the biorthogonal and orthonormal FBs. The FB is said to be biorthogonal if $\mathbf{R}(e^{j\omega})\mathbf{E}(e^{j\omega}) = \mathbf{I}$ for all ω. The FB is said to be orthonormal if $\mathbf{E}(e^{j\omega})$ is unitary for all ω and $\mathbf{R}(e^{j\omega}) = \mathbf{E}^{\dagger}(e^{j\omega})$.

If $\mathbf{E}(e^{j\omega})$ is memoryless (i.e., $\mathbf{E}(e^{j\omega})$ is a constant matrix for all ω), it is called the *transform coding* problem, and was addressed by Huang and Schultheiss [33]. The optimized system is called the *KLT coder*. Later in [79], the authors proposed the prediction-based lower triangular transform coder (*PLT coder*) which used the MINLAB structure [82], and showed that it can also achieve the same optimality as the KLT coder. When the filter order is unconstrained (i.e., $\mathbf{E}(e^{j\omega})$ has infinite memory), theoretical results on the optimal orthonormal filter bank are provided in [106]. It was shown that there are two necessary and sufficient conditions for the optimal orthonormal SBC, namely, *total decorrelation* and *spectrum majorization*. The theory of optimal orthonormal FB is closely related to the principal component filter banks (PCFB) [3, 100]. For the case of biorthogonal filter banks, it was conjectured that the optimal structure is the cascade of the optimal orthonormal filter bank, and a set of half-whitening filters applied to the signal in each individual subband [109]. This

conjecture was later proven to be true in [70]. [70] also showed the two fundamental properties, total decorrelation and spectrum majorization, are also two necessary conditions for optimality of the biorthogonal filter banks. The finite impulse response (FIR) solutions to the orthonormal and biorthogonal filter banks are also discussed extensively in the literature [44, 69, 70, 102, 58].[1]

In Chapter 5 of this thesis, we proposed the GTD transform coder, which combines the idea of linear precoding and the estimation stage of MINLAB structure. The GTD transform coder is shown to be a general family of optimal transform coders that achieves the same theoretical coding gain optimality as the KLT and the PLT coders. In this chapter we propose a novel structure for the perfect reconstruction subband coders as shown in Fig. 6.1. This new structure can be seen as a generalization of the GTD transform coder, where the precoder $\mathbf{E}(e^{j\omega})$ and the estimators $P_{ij}(e^{j\omega})$ are all linear filters instead of one-tap multipliers. For this reason, we call such filter banks the *GTD filter banks.* We optimize the theoretical coding gain performance of the proposed structures for the unconstrained filter order case. We will discuss two cases in detail: The first case is the orthonormal precoder case where $\mathbf{E}(e^{j\omega})$ is paraunitary, and is called the *orthonormal GTD filter bank*; the second case is the general biorthogonal case where the only constraint is $\mathbf{R}(e^{j\omega})\mathbf{E}(e^{j\omega}) = \mathbf{I}$, and is called the *biorthogonal GTD filter bank*. The more general case where biorthogonality is not imposed is not considered here, as it is more involved.

The optimization of the perfect reconstruction GTD filter banks is challenging due to the non-linear nature of the objective functions and the constraints. Interestingly, the theory for GTD FB developed in this chapter is parallel to the theory in traditional FB. We will first show that there are also two fundamental properties both in the optimal orthonormal GTD FB and in the optimal biorthogonal GTD FB, namely, *total decorrelation* and *spectrum equalization*. From these properties, we will derive the optimal solutions for each case. It will be shown that the frequency dependent GTD of the Cholesky factor of the input spectrum density matrix is crucial in designing the optimal orthonormal GTD FB. For the case of biorthogonal GTD FB, we will show that the optimal solution is achieved by cascading an optimal orthonormal GTD FB with a set of scalar filters that are half-whitening filters for the subband signals. The performance of the two GTD FB classes are all superior to traditional FB. The optimal orthonormal GTD FB and the optimal biorthogonal GTD FB achieve exactly the *determinant bounds* derived in [109] for the traditional orthonormal FB and biorthogonal FB, respectively, which are generally not achievable by traditional FB. We will

[1] The filter banks discussed in this paragraph will be referred to as *traditional filter banks* throughout this chapter.

also show that if the FB is designed via the frequency dependent geometric mean decomposition (GMD), then the FB is optimal. Another advantage of the GMD FB is that the bit loading scheme is uniform. Therefore, it does not suffer from the bit loading granularity problem as in the traditional FB [70].

The theory of optimal filter banks is not only useful in data compression but also in digital communication [3, 18]. The notion of duality in the optimal DMT systems and biorthogonal subband coders has been reported in [53]. We also find GTD filter banks useful in wireless communication systems over slowly time-varying frequency selective channels with linear precoding and zero-forcing decision feedback equalizers. Our focus is on the quality of service (QoS) problem, namely, in minimizing the transmitted power subject to specified bit error rate and bit rate constraint. We will show that the optimal systems are related to the frequency dependent GTD of the channel response matrix.

The content of this chapter is mainly drawn from [133], and portions of it have been presented in [131, 132].

6.1 Outline

This chapter is organized as follows. Sec. 6.2 formulates the perfect reconstruction GTD FB optimization problems for subband coders. Sec. 6.3 and Sec. 6.4 provide mathematical derivations and solutions to the optimal orthonormal GTD FBs and the biorthogonal GTD FBs, respectively. Some discussions and performance comparisons of these FBs are presented in Sec. 6.5. Sec. 6.6 introduces the use of GTD filter banks in the context of wireless communications. Concluding remarks are made in Sec.6.7.

6.2 Subband Coder Signal Model

The GTD subband coder structure is shown in Fig. 6.1 for $M = 4$. Here the vector process $\mathbf{x}(n)$ is the M-blocked version of the scalar process $x(n)$. We assume that the vector process $\mathbf{x}(n)$ is wide sense stationary (WSS). The power spectral density (psd) matrix of the WSS vector $\mathbf{x}(n)$ is denoted by $\mathbf{S}_{xx}(e^{j\omega})$. We will find in sequel that the eigenvalues of $\mathbf{S}_{xx}(e^{j\omega})$ appear in various denominators. To make this meaningful, we will assume throughout that $\mathbf{S}_{xx}(e^{j\omega})$ is nonsingular. We denote by $\{\eta_i(e^{j\omega})\}$ the set of eigenvalues of $\mathbf{S}_{xx}(e^{j\omega})$ ordered such that $\eta_i(e^{j\omega}) \geq \eta_{i+1}(e^{j\omega})$ for

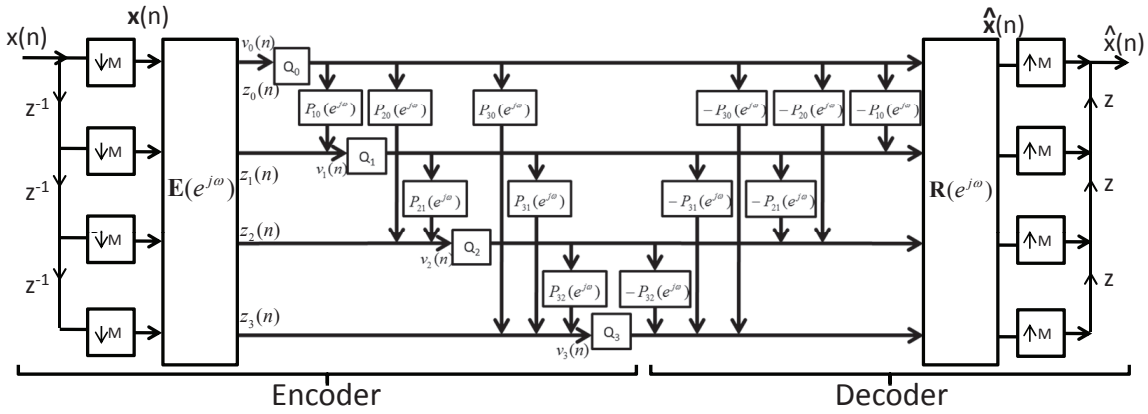


Figure 6.1: The biorthogonal GTD subband coders for $M = 4$.

all ω.[2]

The signal $\mathbf{x}(n)$ first passes through a filter $\mathbf{E}(e^{j\omega})$. Let $\mathbf{z}(n) = [z_0(n)\ z_1(n)\ \cdots\ z_{M-1}(n)]^T$ denote the output of $\mathbf{E}(e^{j\omega})$. Before the quantizers, the signal $\mathbf{z}(n)$ passes through a frequency-dependent PLT stage, where the MINLAB structure [82] is used to ensure unity noise gain. The kth quantizer input $v_k(n)$ is the sum of the signal $z_k(n)$ and the filtered version of the quantized signal $v_0(n)$, $v_1(n)$, up to $v_{k-1}(n)$. The filter $P_{ik}(e^{j\omega})$ is the estimation filter from the kth stream to the ith stream. The decoder performs the inverse operations on the quantized data. The validity of the MINLAB structure assumptions must rely on the high-bit-rate assumption where we assume that the prediction based on the quantized data is not too much different from that on the unquantized data. Under this assumption, the signal $\mathbf{v}(n)$ before the quantizer is the filtered version of $\mathbf{x}(n)$ passing through the filter $\mathbf{L}(e^{j\omega})\mathbf{E}(e^{j\omega})$, where $\mathbf{L}(e^{j\omega})$ is the filter used to represent the frequency dependent PLT stage. In particular, $\mathbf{L}(e^{j\omega})$ is a lower triangular matrix with unity on its diagonals for all frequencies.

Since $\mathbf{x}(n)$ is zero-mean and WSS, the quantizer inputs $v_i(n)$'s are therefore zero-mean and jointly WSS with psd matrix

$$\mathbf{S}_{vv}(e^{j\omega}) = \mathbf{L}(e^{j\omega})\mathbf{E}(e^{j\omega})\mathbf{S}_{xx}(e^{j\omega})\mathbf{E}^{\dagger}(e^{j\omega})\mathbf{L}^{\dagger}(e^{j\omega}). \tag{6.1}$$

The ith quantizer input signal variance is therefore

$$\sigma_{v_i}^2 = \int_0^{2\pi} [\mathbf{S}_{vv}(e^{j\omega})]_{ii} \frac{d\omega}{2\pi}. \tag{6.2}$$

[2]Note that if the vector process is obtained from blocking a scalar process, its power spectrum density matrix has pseudo-circulant structure [110, 107]. However, we will see that the theory developed in this section does not depend nor utilize this structure. Hence, the results of this section are not restricted to the blocked version of a scalar input process, but are also true for any WSS vector process with well-defined power spectrum density matrix.

To derive the coding gain expression, we model the quantizers with additive noise sources $q_i(n)$. We assume these noise sources are jointly WSS, white, with zero mean and with variances of the form

$$\sigma_{q_i}^2 = c2^{-2b_i}\sigma_{v_i}^2 \tag{6.3}$$

where b_i is the number of bits assigned to the ith quantizer. So the quantizer noise psd $\mathbf{S}_{qq}(e^{j\omega})$ is a constant diagonal matrix with diagonal elements $\sigma_{q_i}^2$. This is the *high-bit-rate* assumption, and is also used in the previous chapter. The average bit rate $b = \frac{1}{M}\sum_{i=0}^{M-1} b_i$ is assumed to be fixed. The reconstruction error vector is $\mathbf{e}(n) = \widehat{\mathbf{x}}(n) - \mathbf{x}(n)$. Based on unity noise gain property in the MINLAB structure [82], the error vector $\mathbf{e}(n)$ can be regarded as the output of the synthesis matrix $\mathbf{R}(e^{j\omega})$ in response to the quantization error $\mathbf{q}(n)$. Thus, the psd matrix of the error $\mathbf{e}(n)$ is $\mathbf{S}_{ee}(e^{j\omega}) = \mathbf{R}(e^{j\omega})\mathbf{S}_{qq}(e^{j\omega})\mathbf{R}^\dagger(e^{j\omega})$ The average mean square error of the coder ε_{coder} is[3]

$$\begin{aligned}
\varepsilon_{coder} &= \frac{1}{M}E[\mathbf{e}^\dagger(n)\mathbf{e}(n)] = \frac{1}{M}\mathrm{Tr}(E[\mathbf{e}(n)\mathbf{e}(n)^\dagger]) \\
&= \frac{1}{M}\int_0^{2\pi}\mathrm{Tr}(\mathbf{S}_{ee})\frac{d\omega}{2\pi} = \frac{1}{M}\int_0^{2\pi}\mathrm{Tr}(\mathbf{R}\mathbf{S}_{qq}\mathbf{R}^\dagger)\frac{d\omega}{2\pi} \\
&= \frac{1}{M}\int_0^{2\pi}\mathrm{Tr}(\mathbf{R}^\dagger\mathbf{R}\mathbf{S}_{qq})\frac{d\omega}{2\pi} \\
&= \frac{1}{M}\sum_{i=0}^{M-1}\sigma_{q_i}^2\int_0^{2\pi}[\mathbf{R}^\dagger\mathbf{R}]_{ii}\frac{d\omega}{2\pi}
\end{aligned}$$

Using (6.3) and $\sigma_{v_i}^2 = \int_0^{2\pi}[\mathbf{L}\mathbf{E}\mathbf{S}_{xx}\mathbf{E}^\dagger\mathbf{L}^\dagger]_{ii}\frac{d\omega}{2\pi}$, we get

$$\sigma_{q_i}^2 = c2^{-2b_i}\int_0^{2\pi}[\mathbf{L}\mathbf{E}\mathbf{S}_{xx}\mathbf{E}^\dagger\mathbf{L}^\dagger]_{ii}\frac{d\omega}{2\pi}$$

Substituting into the preceding equation, this yields

$$\begin{aligned}
\varepsilon_{coder} &= \frac{1}{M}\sum_{i=0}^{M-1}c2^{-2b_i}\int_0^{2\pi}[\mathbf{L}\mathbf{E}\mathbf{S}_{xx}\mathbf{E}^\dagger\mathbf{L}^\dagger]_{ii}\frac{d\omega}{2\pi}\times\int_0^{2\pi}[\mathbf{R}^\dagger\mathbf{R}]_{ii}\frac{d\omega}{2\pi} \\
&\geq c2^{-2b}\prod_{i=0}^{M-1}\left(\int_0^{2\pi}[\mathbf{L}\mathbf{E}\mathbf{S}_{xx}\mathbf{E}^\dagger\mathbf{L}^\dagger]_{ii}\frac{d\omega}{2\pi}\times\int_0^{2\pi}[\mathbf{R}^\dagger\mathbf{R}]_{ii}\frac{d\omega}{2\pi}\right)^{\frac{1}{M}},
\end{aligned}$$

[3]For simplicity, from now on we drop the argument "$e^{j\omega}$" when there is no confusion. For example, by $\mathbf{S}_{ee}$ we mean the psd matrix $\mathbf{S}_{ee}(e^{j\omega})$.

where we have used the AM-GM inequality [29]. This becomes an equality if the terms in the summation are identical for all i, and can be accomplished by choosing the b_i according to the optimum bit loading formula[4] similar to that in [109], i.e.,

$$b_i = b + 0.5 \log_2 \sigma_{v_i}^2 K_i^2 - 0.5 \sum_i \log_2 \sigma_{v_i}^2 K_i^2 / M, \tag{6.4}$$

where $K_i^2 = \int_0^{2\pi} [\mathbf{R}^\dagger \mathbf{R}]_{ii} \frac{d\omega}{2\pi}$. Let us define

$$\phi = \prod_{i=0}^{M-1} \int_0^{2\pi} [\mathbf{L}\mathbf{E}\mathbf{S}_{xx}\mathbf{E}^\dagger\mathbf{L}^\dagger]_{ii} \frac{d\omega}{2\pi} \int_0^{2\pi} [\mathbf{R}^\dagger\mathbf{R}]_{ii} \frac{d\omega}{2\pi}. \tag{6.5}$$

Therefore, the average MSE of the coder under optimal bit allocation becomes

$$\epsilon_{coder} = c 2^{-2b} \phi^{1/M}.$$

The coding gain of a coder is defined by comparing the average mean square value ε_{coder} of the reconstruction error $\mathbf{x}(n) - \hat{\mathbf{x}}(n)$ with the mean square value ε_{direct} of the direct quantization error (roundoff quantizer) with the same bit rate b. An expression for the coding gain G_C can be written as

$$G_C = \frac{\varepsilon_{direct}}{\varepsilon_{coder}}. \tag{6.6}$$

Thus, maximizing the coding gain is equivalent to minimizing ϕ by choosing $\{\mathbf{E}(e^{j\omega}), \mathbf{R}(e^{j\omega}), \mathbf{L}(e^{j\omega})\}$ subject to some constraints. In this section, we consider problems similar to what was considered in [106] and [109] – a theoretical performance bound of the infinite order perfect reconstruction filter banks. We consider two classes of subband coders. The first class is when the precoder $\mathbf{E}(e^{j\omega})$ in Fig. 6.1 is restricted to be paraunitary, i.e., $\mathbf{E}(e^{j\omega})$ is unitary for all ω. We call such filter banks the *orthonormal GTD filter banks*, since the columns of $\mathbf{E}(e^{j\omega})$ are orthonormal for every frequency. In

[4]The optimum bit loading formula for conventional perfect reconstruction filter banks usually has the granularity problem, i.e., the number of bits in the formula needs to be rounded off to the nearest integer when used in practice. This results in some performance loss. However, it will be seen later that this problem does not exist in the optimal orthonormal and biorthogonal GTD FBs where uniform bit loading can be applied.

this case, the optimization problem can be written as

$$\begin{aligned} \min_{\mathbf{E},\mathbf{R},\mathbf{L}} \quad & \phi \\ s.t. \quad & \text{(a) } \mathbf{R}(e^{j\omega})\mathbf{E}(e^{j\omega}) = \mathbf{I} \\ & \text{(b) } \mathbf{E}(e^{j\omega}) \text{ is paraunitary.} \end{aligned} \tag{6.7}$$

The second class is when the paraunitary precoder constraint is absent; we call such filter banks *the biorthogonal GTD filter banks*, since the matrix $\mathbf{E}(e^{j\omega})$ and $\mathbf{R}(e^{j\omega})$ are biorthogonal pairs. In this case, the optimization problem is

$$\begin{aligned} \min_{\mathbf{E},\mathbf{R},\mathbf{L}} \quad & \phi \\ s.t. \quad & \mathbf{R}(e^{j\omega})\mathbf{E}(e^{j\omega}) = \mathbf{I}. \end{aligned} \tag{6.8}$$

6.3 Optimal Orthonormal GTD Filter Banks

In this section, we discuss the coding gain optmization for the orthonormal GTD filter banks. Because of the paraunitary precoder constraint and the perfect reconstruction constraint, it can be seen that $\mathbf{R}(e^{j\omega})$ is also paraunitary. Substitute this into (6.5), we have

$$\phi = \prod_{i=0}^{M-1} \sigma_{v_i}^2 = \prod_{i=0}^{M-1} \int_0^{2\pi} [\mathbf{LES}_{xx}\mathbf{E}^\dagger\mathbf{L}^\dagger]_{ii} \frac{d\omega}{2\pi},$$

which is purely the product of subband signal variances. Thus, the problem of maximizing the coding gain is the same as minimizing the product of subband variances. In the following we will derive a set of necessary and sufficient conditions for the optimality of the orthonormal GTD filter banks.

For orthonormal subband coders, Vaidyanathan proved that total decorrelation is necessary for the optimal coders [106]. The same condition is also necessary for the optimal orthonormal GTD filter banks.

Theorem 6.3.1 *(Total Decorrelation Is Necessary) For fixed input psd $\mathbf{S}_{xx}(e^{j\omega})$, suppose a coder is optimal (in the coding gain sense) within the class of all orthonormal GTD filter banks with optimal bit allocation.*

Then, the random processes before each quantizer are uncorrelated with each other, that is

$$E[v_i(n)v_k^*(m)] = 0 \text{ for } i \neq k, \text{ and for all } n,\ m. \tag{6.9}$$

This condition will also be referred to as total decorrelation *of the subbands.* ♢

Proof: See Appendix A. □

Thus, for optimality, the random processes $v_i(\cdot)$ and $v_k(\cdot)$ must be decorrelated, not just the random variables $v_i(n)$ and $v_k(n)$ for each time n. Equivalently, the psd matrix of the vector process $\mathbf{v}(n) = [v_0(n)\ v_1(n)\ \cdots\ v_{M-1}(n)]^T$ must be diagonal, i.e.,

$$\mathbf{S}_{vv}(e^{j\omega}) = \text{diag}([S_{v_0}(e^{j\omega}), \cdots, S_{v_{M-1}}(e^{j\omega})]^T). \tag{6.10}$$

For the optimal orthonormal GTD filter banks if the precoder $\mathbf{E}(e^{j\omega})$ is given, the optimal estimator matrices will be the one that satisfies (6.10). We can thus write the optimal estimator matrix $\mathbf{L}(e^{j\omega})$ as a function of the precoder $\mathbf{E}(e^{j\omega})$. Based on this observation, we developed the following Corollary which is useful later in deriving the optimal precoder.

Corollary 6.3.2 *Given any $\mathbf{E}(e^{j\omega})$, the optimal $\mathbf{L}(e^{j\omega})$ in the orthonormal GTD FBs is given by $\mathbf{L}(e^{j\omega}) = \mathbf{L}_e^{-1}(e^{j\omega})$. Here $\mathbf{L}_e(e^{j\omega})$ is a lower triangular matrix with unity on the diagonals, and such that*

$$\mathbf{E}(e^{j\omega})\mathbf{S}_{xx}(e^{j\omega})\mathbf{E}^\dagger(e^{j\omega}) = \mathbf{L}_e(e^{j\omega})\mathbf{D}(e^{j\omega})\mathbf{L}_e^\dagger(e^{j\omega}), \tag{6.11}$$

where $\mathbf{D}(e^{j\omega})$ is a diagonal matrix for all frequencies. Note that (6.11) can be regarded as an LDU decomposition of $\mathbf{E}(e^{j\omega})\mathbf{S}_{xx}(e^{j\omega})\mathbf{E}^\dagger(e^{j\omega})$. ♢

Proof: This is a direct consequence of Theorem 6.3.1. □

We say that the set of the subband signals, or the set of subband signal power spectra $\{S_{v_k}(e^{j\omega})\}$, has *spectrum equalizing* (SE) property if

$$\frac{S_{v_0}(e^{j\omega})}{\int_0^{2\pi} S_{v_0}(e^{j\omega})\frac{d\omega}{2\pi}} = \cdots = \frac{S_{v_{M-1}}(e^{j\omega})}{\int_0^{2\pi} S_{v_{M-1}}(e^{j\omega})\frac{d\omega}{2\pi}} \ \textit{a.e.}, \tag{6.12}$$

where *a.e.* is the abbreviation for *almost everywhere*, which means the set of ω such that (6.12) fails

has measure zero. Note that $\sigma_{v_i}^2 = \frac{1}{2\pi}\int_0^{2\pi} S_{v_i}(e^{j\omega})d\omega$. Thus, (6.12) can also be written as

$$\frac{S_{v_0}(e^{j\omega})}{\sigma_{v_0}^2} = \cdots = \frac{S_{v_{M-1}}(e^{j\omega})}{\sigma_{v_{M-1}}^2} \quad a.e. \tag{6.13}$$

We are now ready to introduce the second necessary condition.

Theorem 6.3.3 *(Spectrum Equalization Is Necessary for Optimal Orthonormal GTD Subband Coders) For fixed input psd $\mathbf{S}_{xx}(e^{j\omega})$ and under the optimal bit allocation constraint, suppose an orthonormal GTD subband coder is optimal (in maximizing the coding gain) among the class of all orthonormal GTD subband coders. Then, the power spectras of the signals before quantizers have the spectrum equalizing property (6.13).*

Proof: See Appendix B. □

It can be shown that neither of the two conditions is individually sufficient. However, in the following we will prove that *if we put them together, that turns out to be sufficient!*

Theorem 6.3.4 *(Optimal Orthonormal GTD Subband Coders) When optimal bit loading is applied, the coding gain of an optimal orthonormal GTD subband coder is maximum for a given input psd $\mathbf{S}_{xx}(e^{j\omega})$ if and only if the signals before the quantizers $v_i(n)$ satisfy the following two properties:*

1. *Total decorrelation as in (6.9).*

2. *Spectrum equalization as in (6.13).*

Proof: In view of earlier theorems, it only remains to prove that the total decorrelation and the spectrum equalizing property together imply optimality. If the filter pair $\{\mathbf{E}(e^{j\omega}), \mathbf{T}(e^{j\omega})\}$ performs total decorrelation, $\mathbf{S}_{vv}(e^{j\omega})$ must be diagonal. If this filter pair also results in the spectrum equalizing property, then the product of the subband variances will be

$$\begin{aligned}
\prod_{i=0}^{M-1} \sigma_{v_i}^2 &= \prod_{i=0}^{M-1} \int_0^{2\pi} S_{v_i}(e^{j\omega}) \frac{d\omega}{2\pi} \\
&= \left(\int_0^{2\pi} \left(\prod_{i=0}^{M-1} S_{v_i}(e^{j\omega}) \right)^{\frac{1}{M}} \frac{d\omega}{2\pi} \right)^M \\
&= \left(\int_0^{2\pi} \det \mathbf{S}_{vv}(e^{j\omega})^{\frac{1}{M}} \frac{d\omega}{2\pi} \right)^M \\
&= \left(\int_0^{2\pi} \det \mathbf{S}_{xx}(e^{j\omega})^{\frac{1}{M}} \frac{d\omega}{2\pi} \right)^M ,
\end{aligned}$$

where in the second equality we have used the spectrum equalizing property, in the third equality we have used the fact that $\mathbf{S}_{vv}(e^{j\omega})$ is diagonal, and in the fourth equality we have used the fact that $\det \mathbf{S}_{vv} = \det \mathbf{L}\mathbf{E}\mathbf{S}_{xx}\mathbf{E}^{\dagger}\mathbf{L}^{\dagger} = \det \mathbf{S}_{xx}$.

Therefore, if both total decorrelation and equalizing property are satisfied, the product of the subband variances is a fixed and unique quantity, which depends only on $\mathbf{S}_{xx}(e^{j\omega})$. Since total decorrelation and spectrum equalizing property are necessary for the optimality and since there is only one value of the product of the subband variances satisfying these two conditions simultaneously, it follows that these two conditions leads to optimality. This completes the proof. □

We now discuss the set of the solutions that satisfy these two necessary and sufficient conditions. By Eq. (6.10), in the optimal orthonormal GTD FBs the psd matrix of the vector process $\mathbf{v}(n)$ will be the diagonal matrix $\mathbf{D}(e^{j\omega})$ for every frequency ω. Using (6.11), we can rewrite the Cholesky factor of $\mathbf{S}_{xx}$ as

$$\mathbf{S}_{xx}^{\dagger/2} = \mathbf{Q}\mathbf{D}^{\frac{1}{2}}\mathbf{\Phi}\mathbf{L}_e^{\dagger}\mathbf{E}, \tag{6.14}$$

where $\mathbf{Q}$ is some paraunitary matrix, $\mathbf{D}^{1/2}$ is a diagonal matrix such that $[\mathbf{D}^{\frac{1}{2}}]_{ii} = [\mathbf{D}]_{ii}^{\frac{1}{2}}$, and $\mathbf{\Phi}$ is some diagonal matrix with unit-magnitude diagonal elements representing the phase ambiguity in the Cholesky decomposition.

We can see that Eq. (6.14) is actually a GTD form [38], where the middle upper triangular matrix has the diagonal elements as in the diagonal matrix $\mathbf{D}(e^{j\omega})$. In order for the SE property (6.13) to be true, the precoder $\mathbf{E}(e^{j\omega})$ should be chosen appropriately. More specifically, the precoder $\mathbf{E}(e^{j\omega})$ should be chosen such that the diagonal matrix $\mathbf{D}(e^{j\omega})$ in frequency ω has diagonal elements

$$[\mathbf{D}(e^{j\omega})]_{ii} = \sqrt[M]{\det \mathbf{S}_{xx}(e^{j\omega})}\, a_i, \tag{6.15}$$

where the values in the set of numbers $\{a_i\}$ satisfy $a_i > 0$ for all i, and $\prod_{i=0}^{M-1} a_i = 1$. This also means that $[\mathbf{D}(e^{j\omega})]_{ii}/[\mathbf{D}(e^{j\omega})]_{jj} = a_i/a_j$ for all ω and all $i, j = 0, 1, \cdots, M-1$. In order for this to hold, we require that the scaled singular values of $\mathbf{S}_{xx}(e^{j\omega})$ multiplicatively majorizes (see Theorem 1 in [122]) the numbers $\{a_0, \cdots, a_{M-1}\}$, i.e.,

$$\frac{\{\eta_0(e^{j\omega}), \cdots, \eta_{M-1}(e^{j\omega})\}}{\sqrt[M]{\det \mathbf{S}_{xx}(e^{j\omega})}} \succ_{\times} [a_0, \cdots, a_{M-1}] \tag{6.16}$$

for all ω. If for some vector $[a_0, \cdots, a_{M-1}]$, Eq. (6.16) is satisfied, then there exists a set of GTD filter

banks such that the coding gain is maximized. In this case, we can plug in the subband variance $\sigma_{v_i}^2 = \int_0^{2\pi} [\mathbf{D}(e^{j\omega})]_{ii} \frac{d\omega}{2\pi}$ to the bit loading formula (6.4) and get the optimal bit loading scheme for the optimal orthonormal GTD filter banks:

$$b_i = b + 0.5 \log_2 a_i. \tag{6.17}$$

Therefore, a_i can be interpreted as the indication of how to allocate the number of bits to each subband. Note that there might be many sets of $\{a_i\}$ satisfying (6.16) for $\mathbf{S}_{xx}(e^{j\omega})$, and therefore we have some freedom in choosing the bit loading vector according to the input psd.

For the case when $[a_0, \cdots, a_{M-1}] = [1, \cdots, 1]$, eq. (6.16) is always satisfied. In this case, the GTD (6.14) actually corresponds to the geometric mean decomposition (GMD), which always exists for any full-rank matrix [38]. We call such filter banks the *orthonormal GMD filter banks*. It is always optimal because it satisfies the two necessary and sufficient conditions. The desirable property of the orthonormal GMD filter banks is that the subband power spectrum will be completely equalized, i.e., $S_{v_i}(e^{j\omega}) = S_{v_j}(e^{j\omega})$ for all i, j and for all ω. Also, the subband variances will also be equal for all subbands. The optimal bit loading formula (6.17) becomes *uniform bit loading*, i.e., $b_i = b$ for all i. If the average bit budget b is an integer, then the granularity problem in the optimal bit loading formula is no longer present in the orthonormal GMD filter banks!

In the following we summarize the above discussions and provide a design procedure for the optimal orthonormal GTD filter banks:

Design of optimal orthonormal GTD filter banks:

1. Find a set of numbers $\{a_0, a_1, \cdots, a_{M-1}\}$ such that $a_i > 0$ and $\prod_{i=0}^{M-1} a_i = 1$. If (6.16) is satisfied, go to step 2; otherwise, find another set of numbers $\{a_i\}$.

2. Perform the frequency dependent GTD (6.14) on the Cholesky factor of $\mathbf{S}_{xx}(e^{j\omega})$, where $\mathbf{D}(e^{j\omega})$ is related to $\mathbf{a}$ as (6.15). The precoder $\mathbf{E}(e^{j\omega})$ is obtained as in (6.14), and the estimation filters are obtained as $\mathbf{L}(e^{j\omega}) = \mathbf{L}_e^{-1}(e^{j\omega})$.

3. Design the optimal bit loading scheme as (6.17).

Now let us discuss the performance of the optimal GTD filter banks. From the above proof in Thm. 6.3.4, we know that the optimal GTD coder will produce the product of the subband variances

$\prod_{i=0}^{M-1}\sigma_{v_i}^2 = \left(\int_0^{2\pi}\det\mathbf{S}_{xx}(e^{j\omega})^{\frac{1}{M}}\frac{d\omega}{2\pi}\right)^M$. The MSE of the direct quantization is

$$\begin{aligned}\varepsilon_{direct} &= c2^{-b}\frac{1}{M}\sum_{k=0}^{M-1}\int_0^{2\pi}[\mathbf{S}_{xx}(e^{j\omega})]_{kk}\frac{d\omega}{2\pi}\\ &= c2^{-b}\int_0^{2\pi}\frac{\mathrm{Tr}(\mathbf{S}_{xx}(e^{j\omega}))}{M}\frac{d\omega}{2\pi}.\end{aligned}$$

Substituting these in (6.6), the maximized coding gain can thus be calculated as

$$G_C = \frac{\int_0^{2\pi}\frac{1}{M}\mathrm{Tr}(\mathbf{S}_{xx}(e^{j\omega}))d\omega}{\int_0^{2\pi}\sqrt[M]{\det\mathbf{S}_{xx}(e^{j\omega})}d\omega}. \tag{6.18}$$

Eq. (6.18) gives a nice closed-form expression for the optimal performance that orthonormal GTD FBs can have. For the traditional filter bank optimization problem discussed in the literature [106], it is well known that the coding performance can be further improved by relaxing the orthonormal constraint. It is thus natural to ask how much performance improvement we can get if we relax the orthonormal precoder constraint in the GTD FBs. This will be addressed in the next section.

6.4 Biorthogonal GTD Filter Banks

To solve the optimization problem (6.8), we first give a necessary condition.

Theorem 6.4.1 *(Total Decorrelation Is Necessary for Optimality) The triplet* $\{\mathbf{E}(e^{j\omega}),\mathbf{R}(e^{j\omega}),\mathbf{L}(e^{j\omega})\}$ *is optimal for (6.8) only if the subband processes* $v_i(n)$ *are totally uncorrelated to each other, i.e.,* $E[v_i(n)v_k^*(m)] = 0$ *for* $i\neq k$ *and for all* $n,\ m$. ♢

Proof: We use the proof technique similar to Theorem 6.3.1 and thus the detail is left to the reader. □

From this theorem, it can also be shown that total decorrelation is achievable by choosing appropriate $\mathbf{L}(e^{j\omega})$ for any given $\mathbf{E}(e^{j\omega})$. In the following we provide a corollary which is similar to Corollary 6.3.2:

Corollary 6.4.2 *Given any* $\mathbf{E}(e^{j\omega})$ *and* $\mathbf{R}(e^{j\omega})$*, the optimal* $\mathbf{L}(e^{j\omega})$ *in the biorthogonal GTD FBs is given by* $\mathbf{L}(e^{j\omega}) = \mathbf{L}_e^{-1}(e^{j\omega})$*. Here* $\mathbf{L}_e(e^{j\omega})$ *is a lower triangular matrix with* 1 *on the diagonals such that the*

LDU decomposition of $\mathbf{E}(e^{j\omega})\mathbf{S}_{xx}(e^{j\omega})\mathbf{E}^{\dagger}(e^{j\omega})$ *is* $\mathbf{L}_e(e^{j\omega})\mathbf{D}(e^{j\omega})\mathbf{L}_e^{\dagger}(e^{j\omega})$, *where* $\mathbf{D}(e^{j\omega})$ *is the diagonal matrix for all frequencies.* $\diamondsuit$

Substituting the result of Corollary 6.4.2 in (6.5), we can establish a lower bound on ϕ:

$$\begin{aligned}
\phi &\geq \prod_{i=0}^{M-1} \int_0^{2\pi} [\mathbf{L}_e^{-1}\mathbf{E}\mathbf{S}_{xx}\mathbf{E}^{\dagger}\mathbf{L}_e^{-\dagger}]_{ii} \int_0^{2\pi} [\mathbf{R}^{\dagger}\mathbf{R}]_{ii} \frac{d\omega}{2\pi} && (6.19)\\
&\geq \left(\int_0^{2\pi} \prod_{i=0}^{M-1} \left([\mathbf{L}_e^{-1}\mathbf{E}\mathbf{S}_{xx}\mathbf{E}^{\dagger}\mathbf{L}_e^{-\dagger}]_{ii} [\mathbf{R}^{\dagger}\mathbf{R}]_{ii} \right)^{\frac{1}{2M}} \frac{d\omega}{2\pi} \right)^{2M} && (6.20)\\
&\geq \left(\int_0^{2\pi} \left(\det \mathbf{L}_e^{-1}\mathbf{E}\mathbf{S}_{xx}\mathbf{E}^{\dagger}\mathbf{L}_e^{-\dagger} \det \mathbf{R}^{\dagger}\mathbf{R} \right)^{\frac{1}{2M}} \frac{d\omega}{2\pi} \right)^{2M} && (6.21)\\
&= \left(\int_0^{2\pi} (\det \mathbf{S}_{xx})^{\frac{1}{2M}} \frac{d\omega}{2\pi} \right)^{2M}, && (6.22)
\end{aligned}$$

where in (6.20) we have used Hölder's inequality for integrals (6.9 in [29]), in (6.21) we have used Hadamard inequality for positive definite matrices (7.8.1 in [31]), and in (6.22) we have used the fact that $\det \mathbf{L}_e = 1$ and $\det \mathbf{R}\mathbf{E} = \det \mathbf{I} = 1$.

Now the question is, whether the bound (6.22) is achievable. The theorem below answers this question in the affirmative, and in particular, this bound can be achieved by a restricted case (where $\mathbf{E}$ can be decomposed as a paraunitary matrix $\mathbf{U}$ and a set of scalar filters) of biorthogonal GTD coder which is shown in Fig. 6.2.

Theorem 6.4.3 *(The Cascade of Optimal Orthonormal GTD Filter Bank and a Set of Half-Whitening Filters is Optimal) The structure in Fig. 6.2, which is a restricted class of the biorthogonal GTD filter banks, achieves (6.22). In particular, (6.22) can be achieved by using* $\mathbf{U}(e^{j\omega})$ *and* $\{\lambda_i(e^{j\omega})\}$, *where* $\mathbf{U}(e^{j\omega})$ *is the precoder solution to the optimal orthonormal GTD filter banks in Section 6.3, and* $\{\lambda_i(e^{j\omega})\}$ *is a set of half-whitening filters determined by the input psd. In particular, we can use the same filter for all subbands, i.e.,* $\lambda_i(e^{j\omega}) = \lambda(e^{j\omega})$. $\diamondsuit$

Proof: The proof is by construction. Consider the solution of the optimal orthonormal GTD filter banks in Sec. 6.3 which performs the frequency dependent GTD for the Cholesky factor of $\mathbf{S}_{xx}(e^{j\omega})$ (we reproduce Eq.(6.14) here without some modifications on the notations):

$$\mathbf{S}_{xx}^{\dagger/2} = \mathbf{Q}\mathbf{D}^{\frac{1}{2}}\mathbf{\Phi}\mathbf{L}_x^{\dagger}\mathbf{P}, \tag{6.23}$$

where both $\mathbf{Q}$ and $\mathbf{P}$ are unitary matrices for all frequencies, $\mathbf{L}_x$ is a lower triangular matrix with 1 on the diagonals. The middle diagonal matrix $\mathbf{D}(e^{j\omega})$ is such that $[\mathbf{D}(e^{j\omega})]_{ii} = \sqrt[M]{\det \mathbf{S}_{xx}(e^{j\omega})}a_i$, where the values in the set $\{a_i\}$ satisfy $a_i > 0$ and $\prod_{i=0}^{M-1} a_i = 1$.

Therefore, $\mathbf{S}_{xx} = \mathbf{P}^\dagger \mathbf{L}_x \mathbf{D} \mathbf{L}_x^\dagger \mathbf{P}$. Note that $\mathbf{P}$ is exactly the precoder of the optimal orthonormal GTD coders described in Sec. 6.3. Now, consider the system structure in Fig. 6.2. Let $\mathbf{\Lambda}(e^{j\omega}) = \text{diag}(\lambda_0(e^{j\omega}), \cdots, \lambda_{M-1}(e^{j\omega}))$ denote the frequency response of the scalar filters. If we take $\mathbf{U} = \mathbf{P}$, the psd matrix of $\mathbf{y}(n)$ can be expressed as $\mathbf{S}_{yy} = \mathbf{\Lambda}\mathbf{L}_x\mathbf{D}\mathbf{L}_x^\dagger\mathbf{\Lambda}^\dagger$. It can be proved that $\mathbf{\Lambda}\mathbf{L}_x = \mathbf{L}_y\mathbf{\Lambda}$ where $\mathbf{L}_y$ is a lower triangular matrix with 1 on the diagonals and $[\mathbf{L}_y]_{ij} = \frac{\lambda_j}{\lambda_i}[\mathbf{L}_x]_{ij}$. Therefore, $\mathbf{S}_{yy}$ can be expressed as $\mathbf{S}_{yy} = \mathbf{L}_y\mathbf{\Lambda}\mathbf{D}\mathbf{\Lambda}^\dagger\mathbf{L}_y^\dagger$. We can take the PLT part to be $\mathbf{L} = \mathbf{L}_y^{-1}$, and the resulting psd matrix will be $\mathbf{S}_{vv} = \mathbf{\Lambda}\mathbf{D}\mathbf{\Lambda}^\dagger$. Substituting these quantities in (6.5), we have

$$\begin{aligned}
\phi &= \prod_{i=0}^{M-1} \int_0^{2\pi} [\mathbf{D}]_{ii} |\lambda_i|^2 \frac{d\omega}{2\pi} \int_0^{2\pi} |\lambda_i|^{-2} \frac{d\omega}{2\pi} && (6.24)\\
&\geq \prod_{i=0}^{M-1} \left(\int_0^{2\pi} ([\mathbf{D}]_{ii})^{1/2} \frac{d\omega}{2\pi} \right)^2 && (6.25)\\
&= \left(\int_0^{2\pi} (\det \mathbf{S}_{xx})^{\frac{1}{2M}} \frac{d\omega}{2\pi} \right)^{2M}, && (6.26)
\end{aligned}$$

where (6.25) is from the Cauchy-Schwartz inequality, and (6.26) is from the fact that $[\mathbf{D}(e^{j\omega})]_{ii} = \sqrt[M]{\det \mathbf{S}_{xx}(e^{j\omega})}a_i$ and $\prod_{i=0}^{M-1} a_i = 1$. The equality in (6.25) can be satisfied by choosing

$$\begin{aligned}
\lambda_i(e^{j\omega}) &= \alpha_i \left([\mathbf{D}]_{ii}(e^{j\omega}) \right)^{-1/4} \\
&= \frac{\alpha_i}{a_i^{1/4}} \left(\det \mathbf{S}_{vv}(e^{j\omega}) \right)^{-\frac{1}{4M}}, && (6.27)
\end{aligned}$$

where α_i is any nonzero scalar multiplier. Thus, λ_i is the *half-whitening filter* in the ith subband. If we choose $\alpha_i = a_i^{1/4}$, for all the subband we can use the same half-whitening filter that has the frequency response

$$\lambda_i(e^{j\omega}) = \lambda(e^{j\omega}) = \left(\det \mathbf{S}_{xx}(e^{j\omega}) \right)^{-\frac{1}{4M}}.$$

Therefore we just need to design one scalar filter $\lambda(e^{j\omega})$ for all the subbands.

To summarize, we have constructed a biorthogonal GTD filter bank in the structure of Fig. 6.2 that achieves exactly (6.22), and is thus optimal for the problem (6.8). □

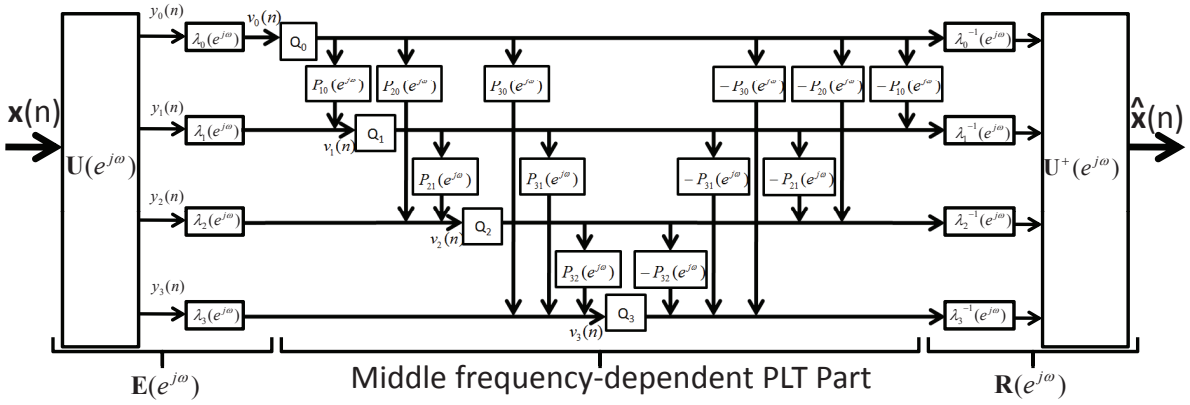


Figure 6.2: A restricted case of the biorthogonal GTD subband coders for $M = 4$.

Theorem 6.4.3 proves that the bound (6.22) is achievable. The proof also suggests a design method for the optimal biorthogonal GTD filter banks. By working out the optimal bit loading formula (6.4) of this design scheme, we can find that the bit loading is exactly the same as (6.17). That is, *the optimal biorthogonal GTD FBs have the same bit loading scheme as the corresponding optimal orthonormal GTD FBs.* This design method is summarized below.

Design of optimal biorthogonal GTD filter banks:

1. Perform the design procedure for optimal orthonormal GTD filter banks as described in Sec. 6.3 to obtained the precoder $\mathbf{U}(e^{j\omega})$ and the estimators $\mathbf{L}_x(e^{j\omega})$.

2. Design the half-whitening filter according to (6.27). The precoder is chosen as $\mathbf{E}(e^{j\omega}) = \mathbf{\Lambda}(e^{j\omega})\mathbf{U}(e^{j\omega})$. The estimators need to be recomputed according to Corollary 6.4.2.

3. Design the optimal bit loading scheme as (6.17).

For the case when the optimal orthonormal GTD filter banks are designed as GMD filter banks, the diagonal matrix $\mathbf{D}(e^{j\omega})$ can be written as $\mathbf{D}(e^{j\omega}) = \sqrt[M]{\det \mathbf{S}_{xx}(e^{j\omega})}\mathbf{I}$. Therefore, for all i we have $[\mathbf{D}(e^{j\omega})]_{ii} = \sqrt[M]{\det \mathbf{S}_{xx}(e^{j\omega})}$. This makes the optimal bit loading formula (6.4) correspond to uniform bit loading, which does not have the granularity problem if the average bit budget is an integer!

Theorem 6.4.3 not only shows the lower bound (6.22) can be achieved but also provides more insight to this problem. The optimal solution to (6.8) must have ϕ equal to (6.22), and thus it must satisfy all the equalities from (6.19) to (6.22). These conditions are necessary conditions of the

optimal systems, and will be examined in the following. First, in (6.20) for the Hölder's inequality

$$\prod_{i=0}^{M-1} \int_0^{2\pi} [\mathbf{L}_e^{-1}\mathbf{E}\mathbf{S}_{xx}\mathbf{E}^\dagger\mathbf{L}_e^{-1}]_{ii} \frac{d\omega}{2\pi} \geq \left(\int_0^{2\pi} \prod_{i=0}^{M-1} [\mathbf{L}_e^{-1}\mathbf{E}\mathbf{S}_{xx}\mathbf{E}^\dagger\mathbf{L}_e^{-1}]_{ii} \frac{d\omega}{2\pi} \right)^M$$

to have equality, we require the spectrum equalizing property (6.13) to be satisfied. This gives us another necessary condition for the optimal biorthogonal GTD filter banks:

Corollary 6.4.4 *Spectrum equalization of Subband Signals is Necessary for Optimal Biorthogonal GTD Filter Banks: The subband signals of the optimal biorthogonal GTD filter banks have the spectrum equalizing property (6.13).* ♢

Second, in (6.21) for the Hadamard inequality

$$\prod_{i=0}^{M-1} [\mathbf{R}^\dagger\mathbf{R}]_{ii} \geq \det(\mathbf{R}^\dagger\mathbf{R})$$

to have equality, $\mathbf{R}^\dagger\mathbf{R}$ needs to be diagonal. This shows that the matrix $\mathbf{R}(e^{j\omega})$ must be decomposable as a paraunitary matrix multiplying with a diagonal matrix. Therefore, the optimal solution to (6.7) will definitely have the structure as in Fig. 6.2. This is stated as the following corollary:

Corollary 6.4.5 *(Optimal Biorthogonal GTD FBs Necessarily Has a Decomposable Structure) The optimal biorthogonal GTD filter banks have the structure as shown in Fig. 6.2.* ♢

To summarize from Theorem 6.4.1, Corollary 6.4.4, and Corollary 6.4.5, the optimal biorthogonal GTD filter bank also has *total decorrelation* and *spectrum equalization* as the two necessary conditions. Furthermore, the optimal systems can be decomposed as the structure in Fig. 6.2. Surprisingly, all of these results have a parallel fashion to the theory of traditional biorthogonal filter banks developed in [109] and [70]: Theorem 2.3 in [70] suggesting that total decorrelation is necessary, Lemma 2.6 in [70] suggesting that spectral majorization is necessary, and Theorem 2.7 in [70] suggesting that the optimal biorthogonal filter banks have the decomposable structure as shown in Fig. 3 of [70].

6.5 Performance Comparison of Optimal Filter Banks Designs

In Table 6.1 we compare the performance of the optimal subband coders in different coder classes. Here $\{\eta_i(e^{j\omega})\}$ is the set of *ordered* eigenvalues of $\mathbf{S}_{xx}(e^{j\omega})$. It is interesting to see that the perfor-

Table 6.1: Features of Optimal Filter Banks Used in Subband Coders

	ϕ (Coding Gain = $\sigma_x^2/\phi^{1/M}$)	Nec. and Suff. Conditions
Orth. SBC [106]	$\prod_{i=0}^{M-1}\int_0^{2\pi}\eta_i\frac{d\omega}{2\pi}$	Total Decor., and Spec. Maj.
Biorth. SBC [70]	$\left(\prod_{i=0}^{M-1}\int_0^{2\pi}\sqrt{\eta_i}\frac{d\omega}{2\pi}\right)^2$	Total Decor., Spec. Maj., and opt. orth. FBs + scalar filters
Orth. GTD SBC	$\left(\int_0^{2\pi}(\det\mathbf{S}_{xx})^{\frac{1}{M}}\frac{d\omega}{2\pi}\right)^M$	Total Decor. and Spec. Eq.
Biorth. GTD SBC	$\left(\int_0^{2\pi}(\det\mathbf{S}_{xx})^{\frac{1}{2M}}\frac{d\omega}{2\pi}\right)^{2M}$	Total Decor., Spec. Eq., and opt. orth. GTD FBs + scalar filters

mance of the optimal orthonormal GTD SBC and the optimal biorthogonal GTD SBC are exactly the *determinant bounds* (not achievable for most input statistics) derived in [109] (see Eq.(39) and Eq.(38) in [109]) for the orthonormal SBCs and the biorthogonal SBCs, respectively. We also list the necessary and sufficient conditions for the optimal solutions in each case.

In the following we compare the coding gain performance of the optimal subband coders in these four cases. We use an AR(1) test input with parameter ρ and an AR(2) process with poles at $z_\pm = \rho e^{\pm j\theta}$. The AR(1) process is often used to model simple images in the literature, and AR(2) process models certain types of image texture as mentioned in [69]. The recursion of the autocorrelation function in AR(2) is $r_n = 2\rho\cos\theta r_{n-1} - \rho^2 r_{n-2}$ with $r_0 = 1$ and $r_1 = (2\rho\cos\theta)/(1+\rho^2)$. We also compare the performance with the theoretical bound on the coding gain, namely the prediction gain given by [107]: $G_{th} = \sigma_x^2/\exp\left(\int_{-\pi}^{\pi}\ln S_{xx}(e^{j\omega})\frac{d\omega}{2\pi}\right)$, where $S_{xx}(e^{j\omega})$ denotes the psd of signal $x(n)$.

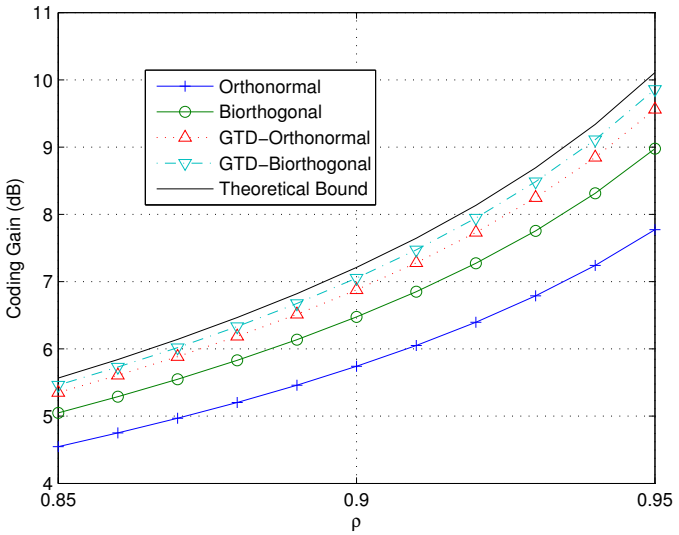


Figure 6.3: Coding gain of subband coders with $M = 3$ for the AR(1) process with ρ from 0.85 to 0.95.

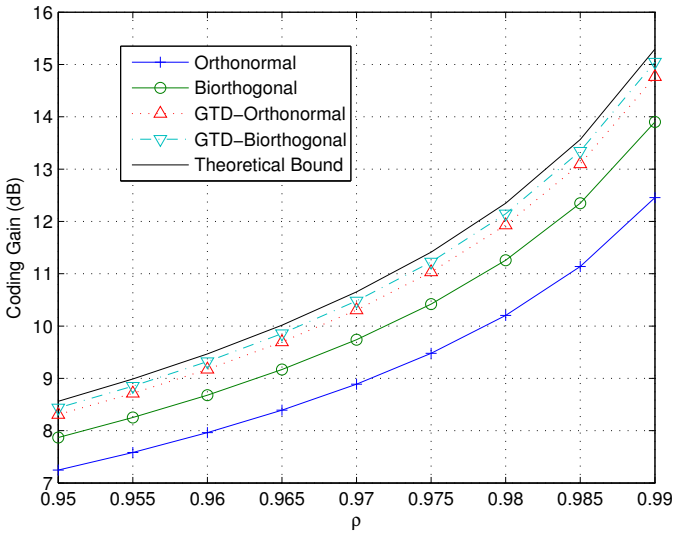


Figure 6.4: Coding gain of subband coders with $M = 4$ for the AR(2) process with ρ from 0.95 to 0.99 and $\theta = \pi/3$.

Fig. 6.3 shows the coding gain for $M = 3$ and ρ from 0.85 to 0.95 for the AR(1) input. Fig. 6.4 shows the coding gain for $M = 4$, ρ from 0.95 to 0.99 and $\theta = \pi/3$. It can be seen that in both cases the optimal biorthogonal GTD coder is only about 0.1dB away from the theoretical bound and is about 1dB better than the optimal biorthogonal subband coders [70]. This suggests the advantage of using the GTD filter banks.

Fig. 6.5 and Fig. 6.6 show the coding gain of M from 2 to 10 for the AR(1) process with $\rho = 0.95$ and the AR(2) process with $\rho = 0.975$ and $\theta = \pi/3$, respectively. It is known [107] that for $M \to \infty$ the optimal orthonormal SBC approaches the theoretical bound (and therefore all four coders approach the bound asymptotically as $M \to \infty$). However, it can be seen that the performance of the biorthogonal GTD coder is close to the bound even for small M. It is also interesting to see that this coder class has monotone coding gain behavior with respect to the block size M for the AR(1) input. However, such monotone behavior is not present for the AR(2) process. This fact was actually also reported for the traditional orthonormal SBC in the literature [107]. However, It was proved that the coding gain for block size M is definitely less than or equal to the coding gain for block size kM where k is any positive integer. Fig. 6.5 and Fig. 6.6 suggest that this phenomenon may also be true in the GTD subband coders.

We have shown the usefulness of the frequency dependent GTD in optimizing the perfect reconstruction filter banks in subband coders. In the next section we will show that the concept is

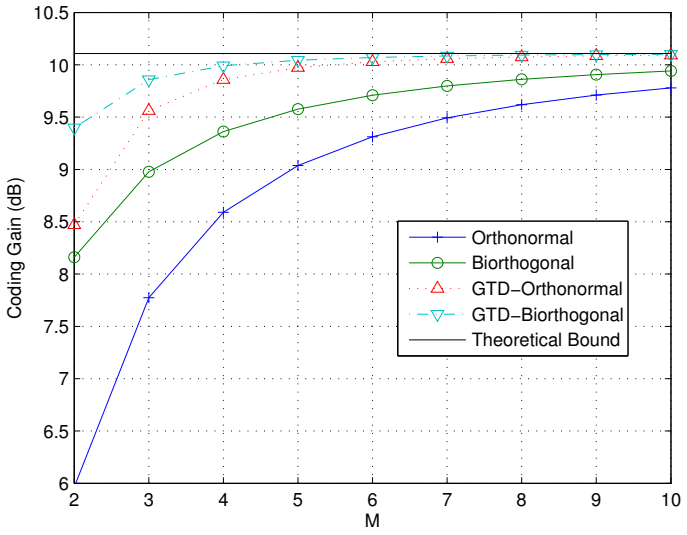


Figure 6.5: Monotone behavior of the coding gain as a function of the number of channels for the AR(1) process with $\rho = 0.95$.

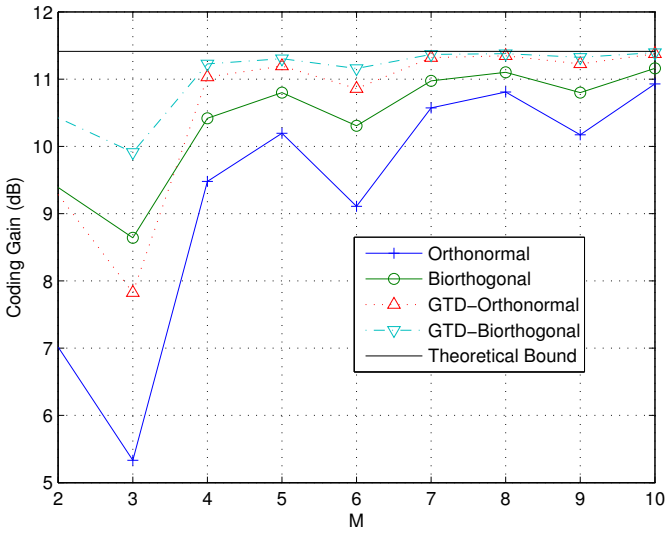


Figure 6.6: Nonmonotone behavior of the coding gain as a function of the number of channels for the AR(2) process with $\rho = 0.975$ and $\theta = \pi/3$.

applicable in digital communication as well.

6.6 The Role of Frequency Dependent GTD in Transceivers for the QoS Problem

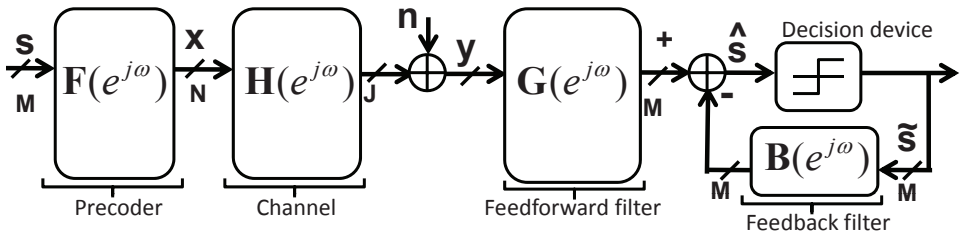


Figure 6.7: Schematic of a frequency selective transceiver with linear precoder and zero-forcing DFE.

In this section we will discuss the role of frequency dependent GTD in optimizing the transceivers for communication systems with frequency selective channels. We consider the wireless communication system with N transmitting antennas and J receiving antennas. The transceiver structure is shown in Fig. 6.7. The channel frequency response is modeled as $J \times N$ matrix $\mathbf{H}(e^{j\omega})$. It is assumed throughout that $\mathbf{H}(e^{j\omega})$ is fixed and the perfect channel state information (CSI) is known at both ends of the communication links. The additive channel Gaussian noise is with no loss of generality assumed to be white, i.e., $\mathbf{S}_{nn}(e^{j\omega}) = \sigma_n^2\mathbf{I}$. The transmitted symbol vector $\mathbf{s}(n)$ is first precoded using the precoder matrix $\mathbf{F}(e^{j\omega})$. At the receiver, zero-forcing (ZF) decision feedback equalization (DFE) is adopted. The received signal $\mathbf{y}(n)$ is first passed through a feedforward filter $\mathbf{G}(e^{j\omega})$. The simple successive decision feedback algorithm can be performed afterwards. The matrix $\mathbf{B}(e^{j\omega})$ is used to represent the causal decision feedback filters across layers, and thus is strictly upper triangular for every frequency.

We consider the optimization of linear precoder $\mathbf{F}(e^{j\omega})$, feedforward filters $\mathbf{G}(e^{j\omega})$, and feedback filters $\mathbf{B}(e^{j\omega})$, subject to the zero-forcing constraint for a channel matrix $\mathbf{H}(e^{j\omega})$. Here zero-forcing condition is imposed here for tractability. The optimal decision feedback receiver for many reasonable objective functions is the MMSE-DFE. The extension to the MMSE-DFE case is currently under investigation. More specifically, we consider the quality of service (QoS) problem of minimizing the transmitted power subject to the specified BER and total bit rate constraint. Under the

assumption of correct decisions of the previous layers, the zero-forcing condition implies

$$\mathbf{G}(e^{j\omega})\mathbf{H}(e^{j\omega})\mathbf{F}(e^{j\omega}) = \mathbf{I} + \mathbf{B}(e^{j\omega}) \text{ for all } \omega.$$

The signal before the decision device will be $\hat{\mathbf{s}} = \mathbf{s} + \mathbf{n}_0$, where $\mathbf{n}_0$ is the noise filtered by $\mathbf{G}(e^{j\omega})$. The kth subchannel thus has the noise power

$$\sigma_{e_k}^2 = \sigma_n^2 \int_0^{2\pi} [\mathbf{G}(e^{j\omega})\mathbf{G}^\dagger(e^{j\omega})]_{kk} \frac{d\omega}{2\pi}.$$

Assume the components $s_k(n)$ of the transmitted symbol vector $\mathbf{s}(n)$ are zero-mean, uncorrelated (both in time and space) processes representing independent data streams with power P_k. So, the input covariance is

$$\mathrm{E}[\mathbf{s}(n)\mathbf{s}(m)^\dagger] = \mathrm{diag}(P_0, P_1, \cdots, P_{M-1})\delta_{nm} \tag{6.28}$$

and the psd matrix of the transmitted symbol is

$$\mathbf{\Lambda}_s(e^{j\omega}) = \mathrm{diag}(P_0, \cdots, P_{M-1}).$$

We consider the case when the kth data stream is a b_k-bit QAM constellation with transmitted power P_k. As discussed in previous chapters, the probability of error for the kth symbol stream can be written as

$$P_{e_k} \approx 4(1 - 2^{\frac{-b_k}{2}})Q\left(\sqrt{\frac{3P_k}{(2^{b_k} - 1)\sigma_{e_k}^2}}\right), \tag{6.29}$$

where $Q(\tau) = \int_\tau^\infty e^{-t^2/2}dt/\sqrt{2\pi}$. Under the high bit rate assumption ($b_k >> 1$) we have $2^{b_k} - 1 \approx 2^{b_k}$ and $1 - 2^{-b_k/2} \approx 1$. By rearranging Eq. (6.29) we get

$$\frac{P_k}{\sigma_{e_k}^2} \approx \frac{2^{b_k}}{3}\left(Q^{-1}\left(\frac{P_{e_k}}{4}\right)\right)^2, \tag{6.30}$$

where $Q^{-1}(\cdot)$ denotes the inverse function of $Q(\cdot)$. We can rewrite the above equation to be

$$P_k = c_k 2^{b_k}\sigma_{e_k}^2, \tag{6.31}$$

where c_k represents the constant related to the probability of error in the kth stream. The transmitted power is a function of the precoder $\mathbf{F}(e^{j\omega})$:

$$\begin{aligned} P_{trans} &= \sum_{k=0}^{M-1} \int_0^{2\pi} [\mathbf{F}\mathbf{\Lambda}_s\mathbf{F}^\dagger]_{kk} \frac{d\omega}{2\pi} = \int_0^{2\pi} \mathrm{Tr}(\mathbf{F}\mathbf{\Lambda}_s\mathbf{F}^\dagger) \frac{d\omega}{2\pi} \\ &= \int_0^{2\pi} \mathrm{Tr}(\mathbf{F}^\dagger\mathbf{F}\mathbf{\Lambda}_s) \frac{d\omega}{2\pi} = \sum_{k=0}^{M-1} P_k \int_0^{2\pi} [\mathbf{F}^\dagger\mathbf{F}]_{kk} \frac{d\omega}{2\pi} \end{aligned}$$

This quantity is what we wish to minimize by designing the transceivers. By substituting (6.31) into the above equation, the transmitted power can be written as

$$\begin{aligned} P_{trans} &= \sum_{k=0}^{M-1} c_k 2^{b_k} \sigma_n^2 \left(\int_0^{2\pi} [\mathbf{F}^\dagger\mathbf{F}]_{kk} \frac{d\omega}{2\pi} \int_0^{2\pi} [\mathbf{G}\mathbf{G}^\dagger]_{kk} \frac{d\omega}{2\pi} \right) \\ &\geq c 2^b \left(\prod_{k=0}^{M-1} \int_0^{2\pi} [\mathbf{F}^\dagger\mathbf{F}]_{kk} \frac{d\omega}{2\pi} \int_0^{2\pi} [\mathbf{G}\mathbf{G}^\dagger]_{kk} \frac{d\omega}{2\pi} \right)^{\frac{1}{M}} \end{aligned}$$

where we have used the AM-GM inequality, and c is a constant related to the specified probability of error such that $c = M\sigma_n^2 \sqrt[M]{\prod_{k=0}^{M-1} c_k}$. The equality can be achieved by appropriately choosing the bit loading b_k similar to [123][5]. Let us define

$$\psi = \prod_{k=0}^{M-1} \int_0^{2\pi} [\mathbf{F}^\dagger\mathbf{F}]_{kk} \frac{d\omega}{2\pi} \int_0^{2\pi} [\mathbf{G}\mathbf{G}^\dagger]_{kk} \frac{d\omega}{2\pi}. \tag{6.32}$$

The transceiver optimization problem can thus be written in the following manner. If the precoder is restricted to have orthonormal columns, the optimization problem is

$$\begin{aligned} \min_{\mathbf{G},\mathbf{F},\mathbf{B}} \quad & \psi \\ s.t. \quad & \text{(a) } \mathbf{G}(e^{j\omega})\mathbf{H}(e^{j\omega})\mathbf{F}(e^{j\omega}) = \mathbf{I} + \mathbf{B}(e^{j\omega}) \\ & \text{(b) } \mathbf{F}(e^{j\omega}) \text{ has orthonormal columns.} \end{aligned} \tag{6.33}$$

[5]Here we relax the non-negative integer bit loading constraint for tractability. Later on we will see that the GMD transceiver, which is an instance of optimal solutions, applies uniform bit loading and thus has no bit loading granularity issue.

If the orthonormal column constraint is not present, the problem becomes

$$\begin{aligned} &\min_{\mathbf{G},\mathbf{F},\mathbf{B}} \quad \psi \\ &\quad s.t. \qquad \mathbf{G}(e^{j\omega})\mathbf{H}(e^{j\omega})\mathbf{F}(e^{j\omega}) = \mathbf{I} + \mathbf{B}(e^{j\omega}). \end{aligned} \tag{6.34}$$

In the following we will discuss in detail how to solve these two problems. The design procedure is to first find some absolute lower bounds on the transmitted power for each case. Then we will show that these bounds are achievable, and the examples of the optimal systems will be given. We will see that the frequency dependent GTD of the channel frequency response plays an important role in the optimal transceiver design.

For both problems (6.33) and (6.34), the following lemma characterizes the optimal feedforward filters if the precoder and the decision feedback filters are determined.

Lemma 6.6.1 *For both problems (6.33) and (6.34), when the precoder $\mathbf{F}(e^{j\omega})$ and the feedback filters $\mathbf{B}(e^{j\omega})$ are given, the optimal feedforward filters $\mathbf{G}(e^{j\omega})$ for minimizing ψ subject to the zero-forcing constraint is*

$$\mathbf{G}_{opt}(e^{j\omega}) = \left(\mathbf{I} + \mathbf{B}(e^{j\omega})\right)\left(\mathbf{H}(e^{j\omega})\mathbf{F}(e^{j\omega})\right)^{\sharp}, \tag{6.35}$$

where $\mathbf{A}^{\sharp}$ denotes the minimum norm pseudo-inverse of the matrix $\mathbf{A}$, i.e., $\mathbf{A}^{\sharp} = (\mathbf{A}^{\dagger}\mathbf{A})^{-1}\mathbf{A}^{\dagger}$ ♢

Proof: See Appendix. □

Substituting (6.35) into ψ in (6.32), we get

$$\psi \geq \left(\prod_{k=0}^{M-1}\int_0^{2\pi}[\mathbf{F}^{\dagger}\mathbf{F}]_{kk}\frac{d\omega}{2\pi}\right) \times \left(\prod_{k=0}^{M-1}\int_0^{2\pi}[(\mathbf{I}+\mathbf{B})(\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{F})^{-1}(\mathbf{I}+\mathbf{B})^{\dagger}]_{kk}\frac{d\omega}{2\pi}\right).$$

Let us examine the factors on the right hand side of the above inequality. For the first term, we have the following inequalities:

$$\prod_{k=0}^{M-1}\int_0^{2\pi}[\mathbf{F}^{\dagger}\mathbf{F}]_{kk}\frac{d\omega}{2\pi} \geq \left(\int_0^{2\pi}\sqrt[M]{\prod_{k=0}^{M-1}[\mathbf{F}^{\dagger}\mathbf{F}]_{kk}}\frac{d\omega}{2\pi}\right)^{M} \tag{6.36}$$

$$\geq \left(\int_0^{2\pi}\sqrt[M]{\det(\mathbf{F}^{\dagger}\mathbf{F})}\frac{d\omega}{2\pi}\right)^{M}, \tag{6.37}$$

where (6.36) is from the Hölder's inequality for integral, and (6.37) is from Hadamard inequality

for positive definite matrices. For the second term, we have

$$\begin{aligned}
&\left(\prod_{k=0}^{M-1}\int_0^{2\pi}[(\mathbf{I}+\mathbf{B})(\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F})^{-1}(\mathbf{I}+\mathbf{B})^\dagger]_{kk}\frac{d\omega}{2\pi}\right)\\
\geq\ &\left(\int_0^{2\pi}\sqrt[M]{\prod_{k=0}^{M-1}[(\mathbf{I}+\mathbf{B})(\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F})^{-1}(\mathbf{I}+\mathbf{B})^\dagger]_{kk}}\frac{d\omega}{2\pi}\right)^M \qquad (6.38)\\
\geq\ &\left(\int_0^{2\pi}\sqrt[M]{\det((\mathbf{I}+\mathbf{B})(\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F})^{-1}(\mathbf{I}+\mathbf{B})^\dagger)}\frac{d\omega}{2\pi}\right)^M \qquad (6.39)\\
=\ &\left(\int_0^{2\pi}\sqrt[M]{\frac{1}{\det(\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F})}}\frac{d\omega}{2\pi}\right)^M, \qquad (6.40)
\end{aligned}$$

where (6.38) is from the Hölder's inequality for integral, (6.39) is from Hadamard inequality for any positive definite matrices, and (6.40) is from the fact that $\det(\mathbf{I}+\mathbf{B}(e^{j\omega}))=1$ for all ω.

To summarize, we have established a lower bound on ψ:

$$\psi^{\frac{1}{M}} \geq \int_0^{2\pi}\sqrt[M]{\det(\mathbf{F}^\dagger\mathbf{F})}\frac{d\omega}{2\pi}\int_0^{2\pi}\sqrt[M]{\frac{1}{\det(\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F})}}\frac{d\omega}{2\pi}, \qquad (6.41)$$

where the right hand side is purely a function of the precoder $\mathbf{F}$. We can also optimize $\mathbf{F}$ to find a better bound. However, one wonders whether these inequalities are achievable with equality, and if so, what the conditions are. In the following we will answer these questions and derive the optimal $\mathbf{F}(e^{j\omega})$ for both problems (6.33) and (6.34).

6.6.1 Transceivers with Orthonormal Precoder Constraint

For the case of orthonormal precoder, we have $\det(\mathbf{F}^\dagger\mathbf{F})=1$ for all ω, and thus the first term on the right hand side of (6.41) is unity. The optimization problem can be rewritten as

$$\begin{aligned}
\min_{\mathbf{F}}\quad &\int_0^{2\pi}\sqrt[M]{\frac{1}{\det(\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F})}}\frac{d\omega}{2\pi}\\
\text{s.t.}\quad &\mathbf{F}(e^{j\omega}) \text{ has orthonormal columns.} \qquad (6.42)
\end{aligned}$$

The following lemma solves the optimization problem (6.42).

Lemma 6.6.2 *The minimum achievable objective value to the optimization problem (6.42) is*

$$\psi_1 = \int_0^{2\pi}\sqrt[M]{\frac{1}{\prod_{k=0}^{M-1}\sigma_{h,k}^2(e^{j\omega})}}\frac{d\omega}{2\pi},$$

where $\sigma_{h,k}(e^{j\omega})$ is the kth largest singular value of the channel matrix $\mathbf{H}(e^{j\omega})$ at frequency ω. $\diamondsuit$

Proof: See Appendix. $\square$

Because of this lemma, we can have the following lower bound for ψ subject to the orthonormal precoder constraint:

$$\psi \geq \left(\int_0^{2\pi} \sqrt[M]{\frac{1}{\prod_{k=0}^{M-1} \sigma_{h,k}^2(e^{j\omega})}} \frac{d\omega}{2\pi} \right)^M.$$

Now the question is whether this lower bound is achievable. This question is answered by examining the equality conditions of (6.36) to (6.40). Because $\mathbf{F}$ has orthonormal columns, equalities in (6.36) and (6.37) are automatically satisfied. On the other hand, for (6.38) and (6.39) to have equalities, we require that

$$[(\mathbf{I}+\mathbf{B})(\mathbf{F}^\dagger \mathbf{H}^\dagger \mathbf{H}\mathbf{F})^{-1}(\mathbf{I}+\mathbf{B})^\dagger]_{kk} = A(e^{j\omega})a_k,$$

where $A(e^{j\omega})$ is some scalar multiplier, a_k are some elements such that $\prod_{k=0}^{M-1} a_k = 1$, and also $(\mathbf{I}+\mathbf{B})(\mathbf{F}^\dagger \mathbf{H}^\dagger \mathbf{H}\mathbf{F})^{-1}(\mathbf{I}+\mathbf{B})^\dagger$ is a diagonal matrix. We assume at frequency ω, $\mathbf{H}(e^{j\omega})$ has rank K_ω, and $K_\omega \geq M$ for all ω because of the zero-forcing assumption, i.e., there is no channel null.

Now consider some vector $[a_0, \cdots, a_{M-1}]$ such that

$$\frac{[\sigma_{h,0}(e^{j\omega}), \cdots, \sigma_{h,M-1}(e^{j\omega})]}{\sqrt[M]{\prod_{k=0}^{M-1} \sigma_{h,k}(e^{j\omega})}} \succ_\times [a_0, \cdots, a_{M-1}] \tag{6.43}$$

for all ω. If (6.43) is satisfied, it means that there exists the GTD form of the channel matrix for every frequency such that

$$\mathbf{H}(e^{j\omega}) = \mathbf{Q}(e^{j\omega})\mathbf{R}(e^{j\omega})\mathbf{P}^\dagger(e^{j\omega}), \tag{6.44}$$

where $\mathbf{R}(e^{j\omega})$ is a $K_\omega \times K_\omega$ upper triangular matrix with diagonal elements $r_k(e^{j\omega})$ so that the first M elements satisfy $|r_k(e^{j\omega})| = \sqrt[M]{\prod_{k=0}^{M-1} \sigma_{h,k}(e^{j\omega})}a_k$, and $\mathbf{Q}(e^{j\omega})$ and $\mathbf{P}(e^{j\omega})$ are both matrices with appropriate dimensions and orthonormal columns.

We are now ready to design our transceiver based on this specific GTD form of the channel matrix. We begin by choosing the precoder as

$$\mathbf{F}(e^{j\omega}) = [\mathbf{P}(e^{j\omega})]_{N\times M}. \tag{6.45}$$

We then choose the feedforward matrix as

$$\mathbf{G}(e^{j\omega}) = \left(\text{diag}([\mathbf{R}(e^{j\omega})]_{M\times M})\right)^{-1}\mathbf{G}_0(e^{j\omega}), \tag{6.46}$$

where

$$\mathbf{G}_0(e^{j\omega}) = [\mathbf{Q}^{\dagger}(e^{j\omega})]_{M\times J}. \tag{6.47}$$

Since $\mathbf{P}(e^{j\omega})$ and $\mathbf{Q}(e^{j\omega})$ have orthonormal columns, the columns of $\mathbf{F}(e^{j\omega})$ are orthonormal, and so are the rows of $\mathbf{G}_0(e^{j\omega})$. Finally the feedback filters $\mathbf{B}(e^{j\omega})$ are determined by the zero-forcing condition, and can be shown to have the form

$$\mathbf{B}(e^{j\omega}) = \left(\text{diag}([\mathbf{R}(e^{j\omega})]_{M\times M})\right)^{-1}[\mathbf{R}(e^{j\omega})]_{M\times M} - \mathbf{I}. \tag{6.48}$$

It can be verified that with such design, all the equalities in (6.36) - (6.39) are satisfied, and thus it is optimal for the optimization problem (6.33). The following theorem summarizes the discussions.

Theorem 6.6.3 *(Optimal solution for transceivers with orthonormal precoder) Suppose* $\mathbf{a} = [a_0, \cdots, a_{M-1}]^T$ *is a vector such that (6.43) is satisfied. If the transceiver is designed as in (6.44) - (6.48), the transceiver is optimal for the problem (6.33), and the minimum transmitted power is*

$$P_{min,orth} = c2^b \int_0^{2\pi} \sqrt[M]{\frac{1}{\prod_{k=0}^{M-1}\sigma_{h,k}^2(e^{j\omega})}}\frac{d\omega}{2\pi}.$$

♢

Proof: This can be directly verified by substituting the equations (6.44) - (6.48). □

The optimal system described in Theorem 6.6.3 is based on the frequency dependent GTD (6.44). This demonstrates again the role of GTD in optimizing the transceivers. In the next subsection we will relax the orthonormal precoder constraint for optimizing the transceivers.

6.6.2 Transceivers with Arbitrary Precoder

In this section we discuss the transceiver design when the precoder is not restricted to have orthonormal columns, and we will call such transceiver an unconstrained ZF transceiver. First we will give a lower bound for the transmitted power of the unconstrained ZF transceiver. We will

then prove that this lower bound is achievable and the solution will again be related to the frequency dependent GTD of the channel matrix.

The right hand side of (6.41) can be used as a lower bound on the required transmitted power. We can further optimize the bound by choosing the precoder $\mathbf{F}$. The optimization problem can be written as

$$\min_{\mathbf{F}} \int_0^{2\pi} \sqrt[M]{\det(\mathbf{F}^\dagger \mathbf{F})} \frac{d\omega}{2\pi} \int_0^{2\pi} \sqrt[M]{\det(\mathbf{F}^\dagger \mathbf{H}^\dagger \mathbf{H}\mathbf{F})^{-1}} \frac{d\omega}{2\pi}. \tag{6.49}$$

To solve the problem (6.49), we first use the Hölder's inequality for integrals [29]:

$$\int_0^{2\pi} \sqrt[M]{\det(\mathbf{F}^\dagger \mathbf{F})} \frac{d\omega}{2\pi} \int_0^{2\pi} \sqrt[M]{\det(\mathbf{F}^\dagger \mathbf{H}^\dagger \mathbf{H}\mathbf{F})^{-1}} \frac{d\omega}{2\pi} \geq \left(\int_0^{2\pi} \sqrt[2M]{\frac{\det(\mathbf{F}^\dagger \mathbf{F})}{\det(\mathbf{F}^\dagger \mathbf{H}^\dagger \mathbf{H}\mathbf{F})}} \frac{d\omega}{2\pi} \right)^2, \tag{6.50}$$

where the equality holds when $\det(\mathbf{F}^\dagger \mathbf{F}) \times \det(\mathbf{F}^\dagger \mathbf{H}^\dagger \mathbf{H}\mathbf{F})$ is constant for all ω. The following lemma gives the minimum value of the right hand side of the above inequality.

Lemma 6.6.4 *Assume $\mathbf{H}(e^{j\omega})$ has rank at least M for every frequency, and $\sigma_{h,k}(e^{j\omega})$ is the kth largest singular value of the channel matrix $\mathbf{H}(e^{j\omega})$ at frequency ω, then*

$$\int_0^{2\pi} \sqrt[2M]{\frac{\det(\mathbf{F}^\dagger \mathbf{F})}{\det(\mathbf{F}^\dagger \mathbf{H}^\dagger \mathbf{H}\mathbf{F})}} \frac{d\omega}{2\pi} \geq \int_0^{2\pi} \sqrt[2M]{\frac{1}{\prod_{k=0}^{M-1} \sigma_{h,k}^2}} \frac{d\omega}{2\pi}. \tag{6.51}$$

♢

Proof: For every frequency ω, we can apply the techniques used in Appendix of [123] and prove that

$$\frac{\det(\mathbf{F}^\dagger(e^{j\omega})\mathbf{F}(e^{j\omega}))}{\det(\mathbf{F}^\dagger(e^{j\omega})\mathbf{H}^\dagger(e^{j\omega})\mathbf{H}(e^{j\omega})\mathbf{F}(e^{j\omega}))} \geq \frac{1}{\prod_{k=0}^{M-1} \sigma_{h,k}^2(e^{j\omega})}.$$

By taking the integral of the $2M$th root on both sides with respect to frequency $\omega \in [0, 2\pi)$, the lemma is proved. □

Combining (6.41), (6.49) and Lemma 6.6.4, we actually showed a lower bound on the transmitted power in the unconstrained ZF transceivers to be

$$P_{trans} \geq c 2^b \left(\int_0^{2\pi} \sqrt[2M]{\frac{1}{\prod_{k=0}^{M-1} \sigma_{h,k}^2}} \frac{d\omega}{2\pi} \right)^2.$$

In the following we will provide a design example showing that this bound is indeed achiev-

able. Consider a set of numbers $\{a_0, \cdots, a_{M-1}\}$ satisfies (6.43) and so that the specific GTD (6.44) exists. We are now ready to design our transceiver based on this specific GTD form of the channel matrix. We begin by choosing the precoder as

$$\mathbf{F}(e^{j\omega}) = \lambda(e^{j\omega})[\mathbf{P}(e^{j\omega})]_{N\times M}, \tag{6.52}$$

where $\lambda(e^{j\omega})$ is chosen to satisfy the equality in (6.50), or

$$\frac{|\lambda(e^{j\omega})|^2}{|\lambda(e^{j\omega})|^{-2}\prod_{k=0}^{M-1}\sigma_{h,k}^2(e^{j\omega})} = d$$

for some constant d.

Therefore,

$$\lambda(e^{j\omega}) = d^{\frac{1}{4}}\left(\prod_{k=0}^{M-1}\sigma_{h,k}^2(e^{j\omega})\right)^{\frac{1}{4}}\phi(e^{j\omega}), \tag{6.53}$$

where $\phi(e^{j\omega})$ is some arbitrary phase response with magnitude $|\phi(e^{j\omega})| = 1$. We then choose the feedforward matrix as

$$\mathbf{G}(e^{j\omega}) = \lambda^{-1}(e^{j\omega})\left(\text{diag}([\mathbf{R}(e^{j\omega})]_{M\times M})\right)^{-1}\mathbf{G}_0(e^{j\omega}), \tag{6.54}$$

where

$$\mathbf{G}_0(e^{j\omega}) = [\mathbf{Q}^\dagger(e^{j\omega})]_{M\times J}. \tag{6.55}$$

Finally the feedback filters $\mathbf{B}(e^{j\omega})$ are determined by the zero-forcing condition, and can be shown to have the form

$$\mathbf{B}(e^{j\omega}) = \left(\text{diag}([\mathbf{R}(e^{j\omega})]_{M\times M})\right)^{-1}[\mathbf{R}(e^{j\omega})]_{M\times M} - \mathbf{I}. \tag{6.56}$$

It can be verified that with such design, all the equalities in (6.36) - (6.39), (6.50), and (6.51), are satisfied, and thus it is optimal for the optimization problem (6.34). The following theorem summarizes the discussions.

Theorem 6.6.5 *(Optimal solution for transceivers with general precoder) Suppose* $\mathbf{a} = [a_0, \cdots, a_{M-1}]^T$ *is a vector such that (6.43) is satisfied. If the transceiver is designed as in (6.44), and (6.52) - (6.56), the*

transceiver is optimal for the problem (6.34), and the minimum transmitted power is

$$P_{unc,min} \geq c2^b \left(\int_0^{2\pi} \sqrt[2M]{\frac{1}{\prod_{k=0}^{M-1} \sigma_{h,k}^2}} \frac{d\omega}{2\pi} \right)^2 .$$

♢

We have derived the optimal transceiver designs for the case of orthonormal precoder and unconstrained precoder. The results of Theorem 6.6.3 and 6.6.5 give elegant design methods for both cases – the optimal orthonormal precoder transceiver design can be obtained from a frequency dependent GTD form of the channel matrix; the optimal unconstrained ZF transceivers can be obtained by designing the optimal orthonormal transceiver first and then cascading with the filter $\lambda(e^{j\omega})$. This is very similar to the case of the subband coder case where the optimal biorthogonal GTD subband coders can be obtained by using the optimal orthonormal GTD subband coders and a set of scalar filters.

6.7 Concluding Remarks

We proposed the perfect reconstruction GTD filter bank structure for subband coding and developed the optimal orthonormal and biorthogonal GTD filter banks when the filter order is unconstrained. The optimal solution is related to the frequency-dependent GTD of the Cholesky factor of the input psd matrix. The performance comparison between the GTD filter banks and the optimal traditional PR filter banks was given. In particular, the optimal GTD filter banks achieve the determinant bounds of the traditional PR filter banks [109]. The theory of optimal GTD filter banks is parallel to that of the optimal traditional filter banks, and these results were summarized in Table I. Furthermore, we extended the use of GTD filter banks to wireless communication systems, and showed that the optimal transceiver in the QoS problem is related to the frequency-dependent GTD of the channel response matrix.

While this chapter contributes to the understanding of the role of GTD in optimizing the perfect reconstruction filter banks, it is also clear that there are some key unsolved issues. For example, finding the FIR approximation for the optimal GTD filter banks will be an important problem. Several more open problems, and the detail discussion, will be made in Sec. 7.2.

6.8 Appendix

6.8.1 Proof of Theorem 6.3.1

Suppose a pair of the subband processes, say $v_0(\cdot)$ and $v_1(\cdot)$, are not uncorrelated. Then, $E[v_0(n)v_1^*(n-k)] = r \neq 0$ for some k. We now show how to decrease the product of the variances by redesigning the estimators. Suppose we use a delay z^{-k} and an additional predictor $-r$ from the 0th stream to the 1st stream to produce the uncorrelated pair $w_0(n)$ and $w_1(n)$ (see Fig. 6.8, where $\mathbf{L}_0(e^{j\omega})$ denotes the remaining frequency dependent PLT part). Note that this fixed estimator $-r$ works for all n by the WSS property. The delay element can be absorbed into the paraunitary filter $\mathbf{E}(e^{j\omega})$, as in the proof of Theorem 1 of [106]. The additional predictor can be absorbed into $\mathbf{L}(e^{j\omega})$ without destroying its structure (i.e., lower triangular with 1's on the diagonal). Also, since $w_1(n)$ is different from $v_1(n)$, all the other estimators need to be changed correspondingly. However, it can be seen that it is possible to make $w_i(n) = v_i(n)$ for $i \geq 2$ by changing these remaining estimators. Thus the structure in Fig. 6.8 is the same as using a modified pair of filters $\{\mathbf{E}_{new}(e^{j\omega}), \mathbf{L}_{new}(e^{j\omega})\}$ where $\mathbf{E}_{new}(e^{j\omega})$ is still paraunitary and $\mathbf{L}_{new}(e^{j\omega})$ is still lower triangular with diagonal entries all equal to unity. We now check if the product of the first two subband variances has been reduced, i.e., if $\sigma_{w_0}^2\sigma_{w_1}^2 < \sigma_{v_0}^2\sigma_{v_1}^2$.

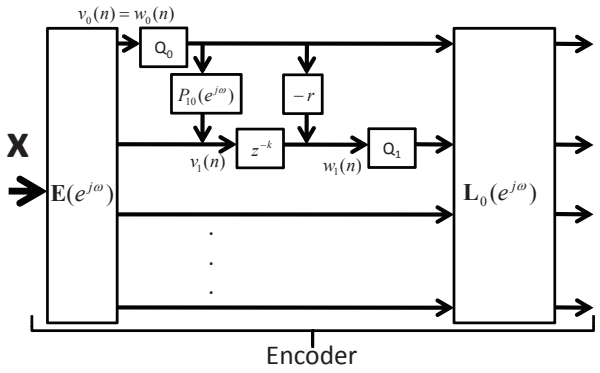


Figure 6.8: Increasing the coding gain by exploiting residual correlation.

Let $\mathbf{R}_w$ and $\mathbf{R}_v$ be the correlation matrices of the vectors $[w_0(n)\ w_1(n)]^T$ and $[v_0(n)\ v_1(n-k)]^T$. Note that by using $\{\mathbf{E}_{new}(e^{j\omega}), \mathbf{L}_{new}(e^{j\omega})\}$, the determinant is preserved, and thus $\det \mathbf{R}_w = \det \mathbf{R}_v$. Note that the diagonal elements of $\mathbf{R}_w$ and $\mathbf{R}_v$ are the quantities $\sigma_{w_i}^2$ and $\sigma_{v_i}^2$. Since $w_0(n)$

and $w_1(n)$ are uncorrelated, we have

$$\sigma_{w_0}^2\sigma_{w_1}^2 = \det \mathbf{R}_w = \det \mathbf{R}_v = \sigma_{v_0}^2\sigma_{v_1}^2 - |r|^2 < \sigma_{v_0}^2\sigma_{v_1}^2,$$

where r denotes the nonzero cross-correlation between $v_0(n)$ and $v_1(n-k)$. Thus, we have shown that for optimality $v_0(\cdot)$ and $v_1(\cdot)$ need to be totally decorrelated. For the case where $(i,k) \neq (0,1)$, if $v_i(\cdot)$ and $v_k(\cdot)$ are not totally decorrelated, similar arguments can be used. However, one has to be careful about the predictor filters, since by adding additional predictor filter $-r$ from ith stream to kth stream, all the remaining estimators need to be changed correspondingly. It can be verified that after the changes, the overall structure is still representable by Fig. 6.1, thus it is still inside the orthonormal GTD coder class we are discussing. Therefore, we have proved optimality implies total decorrelation between $v_i(\cdot)$ and $v_k(\cdot)$ for $i \neq k$. This completes the proof.

6.8.2 Proof of Theorem 6.3.3

Here we first consider the case of $M = 2$. Assume $S_{v_0}(e^{j\omega})/\sigma_{v_0}^2 \neq S_{v_1}(e^{j\omega})/\sigma_{v_1}^2$ for all ω in a set that has nonzero measure. Suppose we are able to produce $w_0(n)$ and $w_1(n)$ from $v_0(n)$ and $v_1(n)$ such that $S_{w_0}(e^{j\omega}) = S_{w_1}(e^{j\omega}) = \sqrt{S_{v_0}(e^{j\omega})S_{v_1}(e^{j\omega})}$. Then, we have

$$\begin{aligned} \sigma_{v_0}^2\sigma_{v_1}^2 &= \frac{1}{4\pi^2}\int_0^{2\pi} S_{v_0}(e^{j\omega})d\omega \int_0^{2\pi} S_{v_1}(e^{j\omega})d\omega \\ &> \frac{1}{4\pi^2}\left(\int_0^{2\pi} \sqrt{S_{v_0}(e^{j\omega})S_{v_1}(e^{j\omega})}d\omega\right)^2 \qquad (6.57) \\ &= \sigma_{w_0}^2\sigma_{w_1}^2, \end{aligned}$$

where the inequality (6.57) is from the Hölder's inequality on square-integrable real-value functions. The inequality is strict since we have $S_{v_0}(e^{j\omega})/\sigma_{v_0}^2 \neq S_{v_1}(e^{j\omega})/\sigma_{v_1}^2$ for some ω on a set that has nonzero measure.

It only remains to prove that such $[w_0(n)\ w_1(n)]^T$ can be obtained from $[v_0(n)\ v_1(n)]^T$ with permissible transformations in the proposed coder structure. By Theorem 6.3.1 we know that the psd matrix of $[v_0(n)\ v_1(n)]^T$ is diagonal for all frequencies since they are totally decorrelated. Taking the determinant on both sides of (6.1), we have

$$S_{v_0}(e^{j\omega})S_{v_1}(e^{j\omega}) = \det \mathbf{S}_{vv}(e^{j\omega}) = \det \mathbf{S}_{xx}(e^{j\omega}).$$

Consider the following decomposition:

$$\mathbf{S}_{xx}^{\dagger/2}(e^{j\omega}) = \mathbf{Q}^{\dagger}(e^{j\omega})\mathbf{R}(e^{j\omega})\mathbf{P}(e^{j\omega}) \tag{6.58}$$

where

$$\mathbf{R}(e^{j\omega}) = \sqrt[4]{S_{v_0}(e^{j\omega})S_{v_1}(e^{j\omega})}\begin{bmatrix} 1 & r(e^{j\omega}) \\ 0 & 1 \end{bmatrix} \tag{6.59}$$

is an upper triangular matrix with diagonals equal to the geometric mean of $\sqrt{S_{v_0}(e^{j\omega})}$ and $\sqrt{S_{v_1}(e^{j\omega})}$. Here $\mathbf{Q}(e^{j\omega})$ and $\mathbf{P}(e^{j\omega})$ are both 2×2 unitary matrices for all frequencies ω. The existence of this decomposition is ensured by the GMD theory [38] for every frequency ω. Let $[w_0(n)\ w_1(n)]^T$ be the signal constructed by passing $[x_0(n)\ x_1(n)]^T$ through filter $\mathbf{P}(e^{j\omega})$ and the predictor matrix $\mathbf{R}_1(e^{j\omega})$, where

$$\mathbf{R}_1(e^{j\omega}) = \begin{bmatrix} 1 & 0 \\ -r^*(e^{j\omega}) & 1 \end{bmatrix}.$$

Thus, we can calculate the psd of $[w_0(n)\ w_1(n)]^T$ as follows:

$$\begin{aligned} & \mathbf{S}_{ww}(e^{j\omega}) \\ = \ & \mathbf{R}_1(e^{j\omega})\mathbf{P}(e^{j\omega})\mathbf{S}_{xx}(e^{j\omega})\mathbf{P}^{\dagger}(e^{j\omega})\mathbf{R}_1^{\dagger}(e^{j\omega}) \\ = \ & \begin{bmatrix} 1 & 0 \\ -r^*(e^{j\omega}) & 1 \end{bmatrix}\mathbf{R}^{\dagger}(e^{j\omega})\mathbf{R}(e^{j\omega})\begin{bmatrix} 1 & -r(e^{j\omega}) \\ 0 & 1 \end{bmatrix} \\ = \ & \begin{bmatrix} \sqrt{S_{v_0}(e^{j\omega})S_{v_1}(e^{j\omega})} & 0 \\ 0 & \sqrt{S_{v_0}(e^{j\omega})S_{v_1}(e^{j\omega})} \end{bmatrix}, \end{aligned}$$

where in the derivation we have substituted in (6.58) and (6.59). Therefore, if we use $\{\mathbf{E}(e^{j\omega}), \mathbf{L}(e^{j\omega})\} = \{\mathbf{P}(e^{j\omega}), \mathbf{R}_1(e^{j\omega})\}$, we are able to decrease the product of the stream signal variances. This completes the proof for the case $M = 2$. For greater M, a similar proof technique can be used, and this is left to the reader.

6.8.3 Proof of Lemma 6.6.1

The proof of this lemma follows from similar arguments as in the proof of Lemma 1 in [123]. Suppose there is another $\mathbf{G}'(e^{j\omega})$ satisfying the zero forcing constraint with the given $\mathbf{F}(e^{j\omega})$ and

$\mathbf{B}(e^{j\omega})$. By Lemma 1 in [123] we can prove that for every frequency ω, $[\mathbf{G}(e^{j\omega})'\mathbf{G}^{\dagger}(e^{j\omega})']_{kk} \geq [\mathbf{G}_{opt}(e^{j\omega})\mathbf{G}_{opt}^{\dagger}(e^{j\omega})]_{kk}$. Since ψ is related to the integral of $[\mathbf{G}(e^{j\omega})\mathbf{G}^{\dagger}(e^{j\omega})]_{kk}$ as in (6.32), by integrating the above inequality for ω from 0 to 2π we can prove this lemma.

6.8.4 Proof of Lemma 6.6.2

Suppose $\{\sigma_{f,k}(e^{j\omega})\}$ denote the singular values of $\mathbf{F}^{\dagger}(e^{j\omega})\mathbf{H}^{\dagger}(e^{j\omega})\mathbf{H}(e^{j\omega})\mathbf{F}(e^{j\omega})$ in descending order. By the interlacing property [31] for the Hermitian matrices $\mathbf{H}^{\dagger}(e^{j\omega})\mathbf{H}(e^{j\omega})$, we have

$$\sigma_{h,0}^2(e^{j\omega}) \geq \sigma_{f,0}(e^{j\omega}) \geq \sigma_{h,1}^2(e^{j\omega}) \geq \cdots \geq \sigma_{f,M-1}(e^{j\omega}).$$

So,

$$1/\prod_{k=0}^{M-1}\sigma_{f,k}(e^{j\omega}) \geq 1/\prod_{k=0}^{M-1}\sigma_{h,k}^2(e^{j\omega}).$$

By integrating this equation, this lemma can be proved.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis, we have studied many important problems of modern signal processing and communication by using the theory of majorization and generalized triangular decomposition.

In Chapter 1, an overview of transceiver optimization and signal-adapted filter bank optimization problems was given. In Chapter 2, we reviewed the mathematical preliminaries needed to understand this thesis. In particular, the theory of additive and multiplicative majorization were introduced. The connection between the notion of majorization and the matrix theory were then reviewed. Finally, the generalized triangular decomposition, as well as the block-diagonal geometric mean decomposition, were introduced.

In Chapter 3 and 4, the roles of majorization and GTD in modern communication were studied. Chapter 3 considered the transceiver optimization for frequency flat MIMO channels and Chapter 4 considered the transceiver design for frequency selective MIMO channels. In Chapter 3, we first studied the problem of jointly optimizing the DFE transceiver with linear precoding and bit allocation, under the total power constraint. We have proposed a general family of GTD transceivers, which optimally solves the DFE transceiver optimization problem. The GTD family also yields optimal solutions for the QoS problem and the bit rate maximization problem. Many existing systems are identified to be special cases of the GTD-based system, and some new GTD-based transceivers were also indicated. The QR-based GTD has the advantage of offering a simple way to perform limited-feedback by sending the bit allocation information from the receiver to transmitter. In the second part of Chapter 3, we focused on the linear transceiver and DFE transceiver design under any linear constraints on the transmit covariance matrix. These constraints include total power

constraint, individual power constraints on the antennas, spectral masks in cable systems to control crosstalk among users, limiting power along some directions, and many more. A two-step approach was proposed to tackle this problem. We first showed that the minimum MSE problem (AM-MSE of GM-MSE, depending on types of transceiver considered) can be solved based on a general semi-definite programming (SDP) framework. The theory of majorization was later used to obtained minimum BER solutions.

In Chapter 4 we studied the transceiver design problem for MIMO frequency selective channels. We focused on the DFE transceiver with linear precoder for the zero-padded frequency selective channels. Using the block-diagonal GMD, we proposed the ZP-BD-GMD transceivers, for both zero-forcing DFEs and MMSE DFEs. Because the block diagonal structure of the ZP-BD-GMD transceivers, the implementation is greatly simplified. Performance of the ZP-BD-GMD were then analyzed. Many desirable properties of the system were also discovered, and the proofs of these properties were presented systematically.

In Chapter 5 and 6, we studied the roles of majorization and GTD in data compression systems. Chapter 5 revisited the transform coding problem and Chapter 6 considered the filter bank optimization problem. In Chapter 5, a general family of optimal transform coders (TC) was introduced based on the GTD. The use of GTD allows the signal variance to be distributed across the subbands. The coding gain of the entire GTD transform coder family, with optimal bit allocation, is maximized. This family includes KLT and PLT coders as special cases. Moreover, many novel transform coders were proposed. In particular, the GMD transform coder can achieve the maximized coding gain with uniform bit loading, thus solving the bit granularity problem. While the previous results are only applicable in the high bit rate case, in the second part of this chapter we addressed the low bit rate coding using the dithered GMD coder. We have proposed two dithered GMD transform coders: the GMD subtractive dithered transform coder (GMD-SD) and the GMD non-subtractive dithered transform coder (GMD-NSD). Both of these two coders use uniform bit loading schemes. We have shown that the proposed dithered GMD transform coders perform significantly better than the original GMD coder in the low rate case.

In Chapter 6 we focused on the signal adapted filter bank optimization. We studied the use of GTD to design the perfect reconstruction filter bank as a subband coder for optimizing the theoretical coding gain. The theory of orthonormal GTD filter banks and biorthogonal GTD filter banks were derived. We have shown that there are two fundamental properties in the optimal solutions,

namely, total decorrelation and spectrum equalization. The optimal GTD filter banks, for both orthonormal and biorthogonal cases, can be obtained by performing the frequency dependent GTD on the Cholesky factor of the input power spectrum density matrices. The connection between the theory of GTD filter banks and the traditional linear filter banks were discussed. We then extended the use of GTD filter banks to wireless communication systems, where linear precoding and zero-forcing decision feedback equalization were used in frequency selective channels. We considered the quality of service (QoS) problem of minimizing the transmitted power subject to the bit error rate and total bit rate constraints. Optimal systems with orthonormal precoder and unconstrained precoder were both derived and shown to be related to the frequency dependent GTD of the channel frequency response.

7.2 Future Work

There are various topics worthy of future research. While the thesis contributes to the understanding of the roles of majorization and GTD in many signal processing problems, it is also clear that there are some key unsolved issues. We summarize several open problems as follows:

1. *Robust Transceiver Designs Against Channel Estimation Errors:* In Chapter 3 and 4 we studied the transceiver designs using majorization theory and GTD. The results were obtained based on the assumption of perfect CSIR and CSIT. However, imperfect channel state information arises in practical communication systems due to channel estimation errors. It is thus essential to have a robust design against these channel uncertainties. The theoretically optimal transceivers using GTD derived in this thesis is a good start point for continuing this line of research.

2. *Signal Independent Transformation:* In Chapter 5, we studied the use of GTD for transform coding. One thing to note for practical use of transform coder is that, in situations involving the KLT, the discrete cosine transform (DCT) is often used instead of the KLT. This is because the DCT is signal independent, computationally efficient, and a good approximation of the KLT for a large class of signals with low-pass spectra [62]. An analogous low-complexity approximation for the precoder $\mathbf{P}$, which arises in the GTD implementation, is not known and worthy of pursuing.

3. *FIR Approximation*: In Chapter 6, we discussed the performance of the GTD filter banks when

the filter order is unconstrained. The theory was useful in giving insight on how to design the implementable finite order filter banks. While the FIR implementations of the traditional PR filter banks are discussed extensively in the literature, the FIR approximation of the optimal GTD filter banks is under investigation. Ideas from the design of traditional filter banks might also be useful here. For example, [97] proposed the greedy algorithm for approximating FIR paraunitary matrix that may be used to design the precoder $\mathbf{P}(e^{j\omega})$; the phase ambiguity used in [98] to improve the design of the FIR filter banks may also be useful here since the phase ambiguity is also present in the GTD case.

4. *Parallel Theory to PCFB*: Principal component filter banks (PCFB) are closely related to optimal orthonormal filter banks. They are known to be optimal for objective functions that are Schur-convex [65] in the subband variances [3, 34]. By modifying the proofs in Chapter 6, it is possible to show that the GMD filter bank is optimal for a wider class of objective functions that are Schur-convex in the logarithm of the subband variances. However, the algebraic theory that is parallel to the linear filter bank version in [34], as well as possible applications are still under investigation.

5. *Relaxing the PR Constraint and the High Bit Rate Assumption*: Traditional filter bank optimization without the perfect reconstruction constraint was solved in [68]. The optimal GTD filter bank without the PR constraint is challenging due to the nonlinear nature of the estimator stage. On the other hand, the validity of the MINLAB structure relies on the high bit rate assumption [82]. The low rate case for the GMD transform coder was discussed in Chapter 5. How to extend the GTD filter bank theory to the low bit rate case is currently also an open problem.

Bibliography

[1] S. O. Aase and T. A. Ramstad, "On the optimality of nonunitary filter banks in subband coders," *IEEE Trans. on Image Processing*, vol. 4, pp. 1585 - 1591, Dec. 1995.

[2] A. N. Akansu and Y. Liu, "On signal decomposition techniques," *Optical Engr.,* vol. 30, pp.912 - 920, Jul. 1991.

[3] S. Akkarakaran and P. P. Vaidyanathan, "Filter bank optimization with convex objectives, and the optimality of principal component forms," *IEEE Trans. Sig. Proc.,* vol. 49, pp. 100 - 114, Jan. 2001.

[4] A. Antoniou and W. S. Lu, *Practical Optimization: Algorithms and Engineering Applications,* Springer. 2007.

[5] A. Barg and D. Y. Nogin, "Bounds on packings of spheres in the Grassmann manifold," *IEEE Trans. Info. Theory,* vol. 48, no. 9, pp. 2450 - 2454, Sep. 2002.

[6] T. Berger and D. W. Tufts, "Optimum pulse amplitude modulation Part I: Transmitter-receiver design and bounds from information theory," *IEEE Trans. Info. Theory,* vol. 13, no. 2, pp. 196 - 208, Apr. 1967.

[7] S. Bergman, D. P. Palomar, and B. Ottersten, "Joint bit allocation and precoding for MIMO systems with decision feedback detection," *IEEE Trans. Sig. Proc.* pp. 4509 - 4521, Nov. 2009.

[8] S. Boyd and L. Vandenberghe, *Convex Optimization,* Cambridge Univ. Press, 2004.

[9] J. Campello, "Optimal discrete bit loading for multicarrier modulation systems," *IEEE International Symp. Info. Theory*, pp. 193, Aug. 1998.

[10] S. S. Chan, T. N. Davidson, and K. M. Wong, "Asymptotically minimum BER linear block precoders for MMSE equalisation," *IEE Proc. Commun.* pp. 297 - 304, Aug. 2004.

[11] D. Chan, and R. W. Donaldson, "Optimum pre- and postfiltering of sampled signals with application to pulse modulation and data compression systems," *IEEE Trans. Comm. Tech.,* vol. 19, no. 2, pp. 141 - 157, Apr. 1971.

[12] C. Y. Chen and P. P. Vaidyanathan, "Precoded FIR and redundant V-BLAST systems for frequency-selective MIMO channels," *IEEE Trans. Sig. Proc.*, vol. 55, no. 7, pp. 3390 - 3404, Jul. 2007.

[13] P. R. Chevillat and G. Ungerboeck, "Optimum FIR transmitter and receiver filters for data transmission over band-limited channels," *IEEE Trans. Comm.,* vol. 50, pp. 1074 - 1080, July 2002.

[14] S. T. Chung, A. Lozano, H. C. Huang, A. Sutivong, and J. M. Cioffi, "Approaching the MIMO capacity with a low-rate feedback channel in V-BLAST," *EURASIP Journal on Applied Signal Processing*, vol. 5 , pp. 762 - 771, 2004.

[15] J. M. Cioffi and G. D. Forney, "Generalized decision-feedback equalization for packet transmission with ISI and Gaussian noise," *Communications, Computation, Control and Signal Processing: A Tribute to Thomas Kailath, A. Paulraj, V. Roychowdhury, and C. D. Shaper,* Eds. Norwell, MA: Kluwer, 1997.

[16] J. P. Costas, "Coding with linear systems," *Proc. IRE.,* pp. 1101 - 1103, Sept. 1952.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory,* Wiley. 1991.

[18] S. Dasgupta and A. Pandharipande, "Optimum biorthogonal DMT systems for multiservice communication," *IEEE Proc. ICASSP.* vol. IV, pp. 552 - 555, Hong Kong, Apr. 2003.

[19] Y. Ding, T. N. Davidson, Z. Luo, and K. M. Wong, "Minimum BER precoders for zero-forcing equalization," *IEEE Trans. Sig. Proc.*, vol. 51, pp. 2410 - 2423, Sep. 2003.

[20] I. Djokovic and P. P. Vaidyanathan, "On optimal analysis/synthesis filters for coding gain optimization," *IEEE Trans. Sig. Proc.,* vol. 44, pp. 1276 - 1279, May 1996.

[21] M. Effros, H. Feng, and K. Zeger, "Suboptimality of Karhunen-Loéve transformation for transform coding," *IEEE Trans. Info. Theory.,* vol. 50, No. 8, pp. 1605 - 1619, Aug. 2004.

[22] D. Falconer and G. J. Foschini, "Theory of minimum mean square error QAM systems employing decision feedback equalization," *Bell Sys. Tech. J.,* vol. 52, no. 10, pp. 1821 - 1849, Dec. 1973.

[23] H. Gazzah, P. A. Regalia, and J. Delmas, "Asymptotic eigenvalue distribution of block Toeplitz matrices and application to blind SISO channel identification," *IEEE Trans. Info.,* vol. 47, No. 3, pp. 1243-1251, Mar. 2001.

[24] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression,* Norwell, MA: Kluwer Academic Publshers, 1992.

[25] G. H. Golub and C. F. Van Loan, *Matrix Computations,* The Johns Hopkins Univ. Press. 1996.

[26] R. M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Trans. Info. Theory,*, vol. 18, pp. 725-730, Nov. 1972.

[27] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Trans. on Info. Theory,* vol. 39, No. 3, pp. 805 - 812, May. 1993.

[28] T. Guess, "Optimal sequences for CDMA with decision-feedback receivers," *IEEE Trans. Inform. Theory,* vol. 49, pp. 886 - 900, Apr. 2003.

[29] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities,* Cambridge University Press, 2nd Edition, 1988.

[30] A. Horn, "On the eigenvalues of a matrix with prescribed singular values,"*Proceedings of the American Mathematical Society,* Vol. 5, No. 1. pp. 4-7. Feb. 1954.

[31] R. A. Horn, and C. R. Johnson, *Matrix Analysis,* Cambridge Univ. Press, 1985.

[32] R. A. Horn, and C. R. Johnson, *Topics in Matrix Analysis,* Cambridge Univ. Press, 1991.

[33] J. J. Y. Huang and P. M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Communication Systems,* vol. CS-11, pp. 289 - 296, Sept. 1963.

[34] O. S. Jahromi, B. A. Francis, and R. H. Raymond, "Algebraic theory of optimal flterbanks," *IEEE Trans. Sig. Proc.,* vol. 51, pp. 442-457, Feb. 2003.

[35] N. S. Jayant and P. Noll, *Digital Coding of Waveforms,* Englewood Cliffs, NJ: Prentice-Hall, 1984.

[36] Y. Jiang, J. Li, and W. W. Hager, "Joint transceiver design for MIMO communications using geometric mean decomposition, " *IEEE Trans. Sig. Proc.*, vol. 53, pp. 3791 - 3803, Oct. 2005.

[37] Y. Jiang, J. Li, and W. W. Hager, "Uniform channel decomposition for MIMO communications, " *IEEE Trans. Sig. Proc.,* vol. 53, pp. 4283 - 4294, Nov. 2005.

[38] Y. Jiang, W. W. Hager, and J. Li, "Generalized triangular decomposition, " *Mathematics of Computation.,* vol. 77, no. 262, Apr. 2008.

[39] Y. Jiang, W. W. Hager and J. Li, "Tunable channel decomposition for MIMO communications using channel state information," *IEEE Trans. Sig. Proc.,* vol. 54, pp. 4405 - 4418, Nov. 2006.

[40] Y. Jiang, W. W. Hager, and J. Li, "The Geometric Mean Decomposition," *Linear Algebra and its Applications,* pp. 373 - 384, Feb. 2005.

[41] Y. Jiang, D. P. Palomar, and M. K. Varanasi, "Precoder Optimization for Nonlinear MIMO Transceiver Based on Arbitrary Cost Function," *Proc. Conference on Information Sciences and Systems.* pp. 119 - 124, March, 2007.

[42] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation,* Prentice Hall, 2000.

[43] N. Karmarkar, "A new polynomial-time algorithm for linear programming," *Combinatorica*, vol. 4, no. 4, pp. 373-395, 1984.

[44] A. Kirac and P. P. Vaidyanathan, "Theory and design of optimum FIR compaction filters," *IEEE Trans. Sig. Proc.*, vol. 46, no. 4, pp. 903 - 919, Apr. 1998.

[45] K. H. Lee and D. P. Petersen, "Optimal linear coding for vector channels," *IEEE Trans. Comm.*, pp. 1283 - 1290, Dec. 1976.

[46] C. C. Li, S. H. Tsai, Y. P. Lin, and P. P. Vaidyanathan, "Optimal Zero-Forcing Transceiver Design for Maximizing Bit Rate Subject to a Total Transmit Power Constraint" *Proc. 16th European Signal Processing Conference.*, Lausanne, 2008.

[47] S. Lin, W. L. Ho, and Y. C. Liang, "Block diagonal geometric mean decomposition (BD-GMD) for MIMO broadcast channels," *IEEE Trans. Wireless Comm.*, vol. 7, no. 7, pp. 2778 - 2789, Jul. 2008.

[48] Y. P. Lin and S. M. Phoong, "ISI-free FIR filterbank transceivers for frequency-selective channels," *IEEE Trans. Sig. Proc.* pp. 2648 - 2658, Nov. 2001.

[49] Y. P. Lin and S. M. Phoong, "BER minimized OFDM systems with channel independent precoders," *IEEE Trans. Sig. Proc.* vol. 51, pp. 2369 - 2380, Sep. 2003.

[50] Y. P. Lin and S. M. Phoong, "Perfect discrete multitone modulation with optimal transceivers," *IEEE Trans. Sig. Proc.* vol. 48, pp. 1702 - 1711, Jun. 2000.

[51] Y. P. Lin and S. M. Phoong, "Optimal ISI-Free DMT transceivers for distorted channels with colored noise," *IEEE Trans. Sig. Proc.* vol. 49, pp. 2702 - 2712, Nov. 2001.

[52] Y. P. Lin, S. M. Phoong, and P. P. Vaidyanathan, *Filter Bank Transceivers for OFDM and DMT Systems,* Cambridge Univ. Press, 2010.

[53] Y. P. Lin, P. P. Vaidyanathan, S. Akkarakaran, and S. M. Phoong, "On the duality of optimal DMT systems and biorthogonal subband coders," *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Proc.,* Istanbul, Turkey, June 2000.

[54] J. Lofberg, "YALMIP : a toolbox for modeling and optimization in MATLAB," *IEEE Intl. Sym. Computer Aided system Design.,* pp. 284 - 289, Sep. 2004.

[55] D. J. Love, R. W. Heath, V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE Selected Areas in Comm.*, vol. 26, no. 8, pp. 1341 - 1365, Oct. 2008.

[56] D. Love and R. W. Heath Jr., "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Trans, Info. Theory,* vol. 51, no. 8, pp. 2967 - 2976, Aug. 2005.

[57] D. Love, "Tables of complex Grassmannian packings," *[Online],* http://www.ece.purdue.edu/∼djlove/grass.html

[58] W. S. Lu and A. Antoniou, "Design of signal-adapted biorthogonal filter banks," *IEEE Trans. Circuits and Systems - I,* vol. 48, No. 1, pp. 90 - 102, Jan. 2001.

[59] Z. Luo, T. N. Davidson, G. B. Giannakis, and K. M. Wong, "Transceiver optimization for block-based multiple access through ISI channels," *IEEE Trans. Sig. Proc.* vol. 52, pp. 1037 - 1052, Apr. 2004.

[60] B. Maison and L. Vandendorpe, "About the asymptotic performance of multiple-input/multiple-output linear prediction of subband signals," *IEEE Sig. Proc. Letters*, vol. 5, no. 12, pp. 315 - 317, Dec. 1998.

[61] S. Mallat and F. Falzon, "Analysis of low bit rate image transform coding," *IEEE Trans. Sig. Proc.,* vol. 46, pp. 1027 - 1042, Apr. 1998.

[62] H. S. Malvar, "Biorthogonal and nonuniform lapped transforms for transform coding with reduced blocking and ringing artifacts,"*IEEE Trans. on Signal Processing,* vol. 46, pp. 1043 - 1053, Apr. 1998.

[63] H. S. Malvar and D. H. Staelin, "Optimal pre- and postfilters for multichannel signal processing," *IEEE Trans. Acoustics, Speech, and Signal Processing*, pp. 287 - 289, Feb. 1988.

[64] D. L. Mary and D. T. M. Slock, "A theoretical high-rate analysis of causal versus unitary online transform coding," *IEEE Trans. Sig. Proc.,* vol. 54, No. 4, pp. 1472 - 1482, Apr. 2006.

[65] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications,* Academic Press. 1979.

[66] D. G. Messerschmitt, "A unified geometric theory of zero forcing and decision feedback equalization," *Rec. Int. Conf. Comm.,* 1973.

[67] A. Mertins, "MMSE design of redundant FIR precoders for arbitrary channel lengths," *IEEE Trans. Sig. Proc.,* vol. 51, pp. 2402 - 2409, Sept. 2003.

[68] M. K. Mihcak, P. Moulin, M. Anitescu, and K. Ramchandran, "Rate-distortion-optimal subband coding without perfect-reconstruction constraints," *IEEE Trans. Sig. Proc.*, vol. 49, pp.542-57, Mar. 2001.

[69] P. Moulin and M. K. Mihcak, "Theory and design of signal-adapted FIR paraunitary filter banks," *IEEE Trans. Sig. Proc.,* vol. 46, pp. 920 - 929, Apr. 1998.

[70] P. Moulin, M. Anitescu, and K. Ramchandran, "Theory of rate-distortion-optimal, constrained filter banks–application to IIR and FIR biorthogonal designs," *IEEE Trans. Sig. Proc.*, vol. 48, pp. 1120 - 1132, Apr. 2000.

[71] Y. Nesterov and A. Nemirovskij, "Interior-point polynomial algorithms in convex programming," *SIAM Studies in Applied Mathematics,* 1994.

[72] S. Ohno, "Performance of single-carrier block transmissions over multipath fading channels with linear equalization," *IEEE Trans. Sig. Proc.,* vol. 54, pp. 3678 - 3687, Oct. 2006.

[73] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: a unified framework for convex optimization," *IEEE Trans. Sig. Proc.* pp. 2381 - 2401, Sept. 2003.

[74] D. P. Palomar and S. Barbarossa "Designing MIMO communication systems: constellation choice and linear transceiver design," *IEEE Trans. Sig. Proc.,* pp. 3804 - 3818, Oct. 2005.

[75] D. P. Palomar and Y. Jiang, "MIMO transceiver design via majorization theory," *Foundations and Trends in Communications and Information Theory,* vol. 3, issue 4, pp. 331 - 551, Nov. 2006.

[76] D. P. Palomar, "Unified framework for linear MIMO transceivers with shaping constraints," *IEEE Communication Letters* vol. 8, nO. 12, pp. 697 - 699, Dec. 2004.

[77] D. P. Palomar, M. A. Lagunas, and J. M. Cioffi, "Optimum linear joint transmit-receive processing for MIMO channels with QoS constraint," *IEEE Trans. Sig. Proc.*, vol. 51, pp. 2381 - 2401, Sept. 2003.

[78] P. A. Parrilo, *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization,* PhD Thesis, California Institute of Technology, May 2000.

[79] S. M. Phoong and Y. P. Lin, "Prediction-based lower triangular transform," *IEEE Trans. Sig. Proc.,* vol. 48, pp. 1947 - 1955, Jul. 2000.

[80] Y. P. Lin and S. M. Phoong, "Minimum redundancy for ISI free FIR filter bank transceivers," *IEEE Trans. Sig. Proc.,* vol. 50, pp. 842 - 853, Apr. 2002.

[81] S. M. Phoong, Y. Chang, and C. Y. Chen, "DFE-modulated filterbank transceivers for multipath fading channels," *IEEE Trans. Sig. Proc.,* vol. 53, pp.182 - 192, Jan. 2005.

[82] S. M. Phoong and Y. P. Lin, "MINLAB: minimum noise structure for ladder-based biorthogonal filter banks," *IEEE Trans. Sig. Proc.,* vol. 48, pp. 465 - 476, Feb. 2000.

[83] N. Prasad and M. K. Varanasi, "Analysis of decision feedback detection for MIMO Rayleigh-fading channels and the optimization of power and rate allocations" , *IEEE Trans. Inform. Theory,* vol. 50, No. 6, pp. 1009 - 1025, June. 2004.

[84] R. Price, "Nonlinearly feedback-equalized PAM versus capacity, for noisy filter channels," *Rec. Int. Conf. Comm.*, 1972.

[85] J. G. Proakis, *Digital Communicatins,* New York: McGraw-Hill, 2001.

[86] J. A. Saghri, A. G. Tescher, and J. T. Reagan, "Practical transform coding of multispectral imagery," *IEEE Signal Processing Magazine,* pp. 32 - 43, Jan. 1995.

[87] J. Salz, "Optimum mean-square decision feedback equalization," *Bell Sys. Tech. J.*, vol. 52, no. 8, pp. 1341 - 1373, Oct. 1973.

[88] J. Salz, "Digital transmission over cross-coupled linear channels," *AT&T Tech. J.*, vol. 64, no. 6, pp. 1147 - 1159, Jul.-Aug. 1985.

[89] A. Scaglione, G. B. Giannakis, and S. Barbarossa, "Redundant filterbank precoders and equalizers part I: unification and optimal designs," *IEEE Trans. Sig. Proc.* vol. 47, No. 7, pp. 1988 - 2006, Jul. 1999.

[90] M. B. Shenouda and T. N. Davidson, "A framework for designing MIMO systems with decision feedback equalization or Tomlinson-Harashima precoding," *IEEE Select. Areas. Commun.*, pp. 401 - 411, Feb. 2008.

[91] M. B. Shenouda, T. N. Davidson, "A design framework for limited feedback MIMO systems with zero-forcing DFE," *IEEE Trans. Selected Areas in Comm.,* vol. 26, pp. 1578 - 1587, Oct. 2008.

[92] P. W. Shor and N. J. A. Sloane, "A family of optimal packings in the Grassmannian manifolds," *J. Algebraic Combin.*, vol. 7, pp. 157 - 163, 1998.

[93] S. Srinivasan, C. Tu, S. L. Regunathan, R. A. Rossi, Jr., G. J. Sullivan, "HD Photo: a new image coding technology for digital photography," *Applications of Digital Image Processing XXX, Proceedings of SPIE,* vol. 6696, San Diego, CA, Aug. 2007.

[94] G. L. Stuber, J. R. Barry, S. W. McLaughlin, L. Ye, M. A. Ingram, and T. G. Pratt, "Broadband MIMO-OFDM wireless communications," *Proceedings of the IEEE.*, vol. 92, Iss. 2, pp. 271 - 294, Feb. 2004.

[95] J. F. Sturm, "Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones" *Optim. Meth. Software.*, vol. 11 - 12, pp. 625 - 653, 1999.

[96] I. E. Telatar "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun,* vol. 10, no. 6, pp. 585 - 595, Nov. 1999.

[97] A. Tkacenko and P. P. Vaidyanathan, "Iterative greedy algorithm for solving the FIR paraunitary approximation problem", *IEEE Trans. Sig. Proc.*, vol. 54, pp. 146 - 160, Jan. 2006.

[98] A. Tkacenko and P. P. Vaidyanathan, "On the spectral factor ambiguity of FIR energy compaction filter banks", *IEEE Trans. Sig. Proc.,* vol. 54, pp. 380 - 385, Jan. 2006.

[99] M. Trivellato, F. Boccardi, and H. Huang, "Zero-Forcing vs Unitary Beamforming inMultiuserMIMO Systems with Limited Feedback," *Proc. IEEE Intern. Symp. Personal, Indoor and Mobile Radio Commun. (PIMRC),* Cannes, France, Sept. 2008.

[100] M. K. Tsatsanis and G. B. Giannakis, "Principal component filter banks for optimal multiresolution analysis," *IEEE Trans. Sig. Proc.,* vol. 43, pp. 1766 - 1777, Aug. 1995.

[101] D. N. C. Tse, P. Viswanath, and L. Zheng, "Diversity-multiplexing tradeoff in multiple-access channels," *IEEE Trans. Info. Theory,* vol. 50, no. 9, pp. 1859 - 1874, Sep. 2004.

[102] J. Tuqan and P. P. Vaidyanathan, "A state space approach to the design of globally optimal FIR energy compaction filter," *IEEE Trans. Sig. Proc.,* vol. 48, No. 10, pp. 2822 - 2838, Oct. 2000.

[103] M. Unser, "On the optimality of ideal filters for pyramid and wavelet signal approximation," *IEEE Trans. Sig. Proc.,* vol. 41, pp. 3591 - 3596, Dec. 1993.

[104] L. Vandenberghe, S. Boyd, and S. P. Wu "Determinant maximization with linear matrix inequality constraints," *SIAM J. Matrix Anal. Appl.* vol. 19, no. 2, pp. 499 - 533, Apr. 1998.

[105] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Rev.* vol. 38, no. 1. pp. 49 - 95, March 1996.

[106] P. P. Vaidyanathan, "Theory of optimal orthonormal subband coders," *IEEE Trans. Sig. Proc.,* vol. 46, pp. 1528 - 1543, Jun. 1998.

[107] P. P. Vaidyanathan, *Multirate Systems and Filter Banks,* Englewood Cliffs, NJ: Prentice-Hall, 1993.

[108] P. P. Vaidyanathan, *The Theory of Linear Prediction,* Morgan & Claypool publishers, 2008.

[109] P. P. Vaidyanathan and A. Kirac, "Results on optimal biorthogonal filter banks," *IEEE Trans. Circuits and Systems - II,* vol. 45, no. 8, pp. 932 - 947, Aug. 1998.

[110] P. P. Vaidyanathan and S. K. Mitra, "Polyphase networks, block digital filtering, LPTV systems, and alias-free QMF banks: a unified approach based on pseudocirculants," *IEEE Trans. Acoustic, speech, and Sig. Proc.*, vol. 36, no. 3, pp. 381 - 391, Mar. 1988.

[111] P. P. Vaidyanathan, S. M. Phoong, and Y. P. Lin, *Signal Processing and Optimization for Transceiver Systems,* Cambridge Univ. Press 2010.

[112] M. K. Varanasi and T. Guess "Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-accesschannel," *Proc. 34th Asilomar Conf. Signals, Syst., Compu.*, pp. 1405 - 1409, Nov. 1997.

[113] A. Vijaya Krishna, and K. V. S. Hari, "Filter bank precoding for FIR equalization in high-rate MIMO communications," *IEEE Trans. Sig. Proc.,* vol. 54, no. 5, pp. 1645 - 1652, May. 2006.

[114] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding,* Englewood Cliffs, NJ: Prentice-Hall, 1995.

[115] Z. Wang, X. Ma, and G. B. Giannakis, "Optimality of single-carrier zero-padded block transmissions," in *Proc. Wireless Communications Networking Conf.,* vol. 2, pp. 660-664, Orlando, FL, Mar. 17 - 21, 2002.

[116] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, and J. N. Wright, "A theory of nonsubtractive dither," *IEEE Trans. on Sig. Proc.*, vol. 48, No. 2, pp. 499 - 516, Feb. 2000.

[117] H. Widom, "Asymptotic behavior of block Toeplitz matrices and determinants. II," *Advances in Mathematics*, vol. 21, pp. 1 - 29, 1976.

[118] H. S. Witsenhausen, "A determinant maximization problem occurring in the theory of data communication," *SIAM J. Applied Mathematics.* vol. 29, no. 3, pp. 515 - 522, Nov. 1975.

[119] C. C. Weng, C. Y. Chen, and P. P. Vaidyanathan, "Joint optimization of transceivers with decision feedback and bit loading," *Proc. 42nd Asilomar Conference on Signals, Systems, and Computers,* Monterey, CA, Oct. 2008.

[120] C. C. Weng, C. Y. Chen, and P. P. Vaidyanathan, "GTD-based transceivers for decision feedback and Bit Loading," *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Proc.*, Taipei, Apr. 2009.

[121] C. C. Weng, C. Y. Chen, and P. P. Vaidyanathan, "Generalized triangular transform coding," *Proc. 43nd Asilomar Conference on Signals, Systems, and Computers,* Monterey, CA, Nov. 2009.

[122] C. C. Weng, C. Y. Chen, and P. P. Vaidyanathan, "Generalized triangular decomposition in transform coding," *IEEE Trans. on Sig. Proc.*, vol. 58, No. 2, pp. 566 - 574, Feb. 2010.

[123] C. C. Weng, C. Y. Chen, and P. P. Vaidyanathan, "MIMO transceivers with decision feedback and bit loading: theory and optimization," *IEEE Trans. Sig. Proc.*, vol. 58, no. 3, pp. 1334 - 1346, Mar. 2010.

[124] C. C. Weng and P. P. Vaidyanathan, "Joint optimization of transceivers with fractionally spaced equalizers," *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Proc.*, Las Vegas, March-April 2008.

[125] C. C. Weng and P. P. Vaidyanathan, "Per-antenna power constrained MIMO transceivers optimized for BER," *Proc. 42nd Asilomar Conference on Signals, Systems, and Computers,* Monterey, CA, Oct. 2008.

[126] C. C. Weng and P. P. Vaidyanathan, "Transceiver design with vector perturbation technique and iterative power loading," *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Proc.,* Taipei, Apr. 2009.

[127] C. C. Weng and P. P. Vaidyanathan, "Block diagonal GMD for zero-padded MIMO frequency selective channels with zero-forcing DFE," *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Proc.*, Dallas, Mar. 2010.

[128] C. C. Weng and P. P. Vaidyanathan, "MIMO transceiver optimization with linear constraints on transmitted signal covariance components," *IEEE Trans. on Sig. Proc.,* vol. 58, pp. 458 - 462 Jan. 2010.

[129] C. C. Weng, P. P. Vaidyanathan, and H. I. Su. "Dithered GMD transform coding," *IEEE Sig. Proc. Letters,* vol. 17, No. 5, pp. 457 - 460, May 2010.

[130] C. C. Weng and P. P. Vaidyanathan, "Block Diagonal GMD for zero-padded MIMO frequency selective channels," *IEEE Trans. on Sig. Proc.*, vol. 59, no. 2, pp. 713 - 727, Feb. 2011.

[131] C. C. Weng and P. P. Vaidyanathan, "Frequency dependent GTD coders," *Proc. 44nd Asilomar Conference on Signals, Systems, and Computers,* Monterey, CA, Nov. 2010.

[132] C. C. Weng and P. P. Vaidyanathan, "The role of GTD in optimizing the perfect reconstruction filter banks," *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Proc.*, Praque, Czech, May, 2011.

[133] C. C. Weng and P. P. Vaidyanathan, "The role of GTD in optimizing the perfect reconstruction filter banks," submitted to *IEEE Trans. on Sig. Proc.*, 2011.

[134] H. Weyl, "Inequalities between the two kinds of eigenvalues of linear transformations," *Proc. Nat. Acad. Sci.*, vol. 35, pp. 408 - 411, July 1949.

[135] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," *IEEE Symp. on Signals, Systems, and Electronics,* pp. 295 - 300, Pisa, Italy, Oct. 1998.

[136] S. P. Wu, L. Vandenberghe, and S. Boyd "MAXDET: Software for determinant maximization problems. User's Guide, Alpha version," Stanford University, CA, 1996.

[137] F. Xu, T. N. Davidson, J. K. Zhang, and K. M. Wong, "Design of block transcivers with decision feedback detection," *IEEE Trans. Signal Processing,* pp. 965 - 978, March. 2006.

[138] J. Yang and S. Roy, "Joint transmitter-receiver optimization for multi-input multi-output systems with decision feedback," *IEEE Trans. Info. Theory,* pp. 1334 - 1347, Vol. 40, Iss. 5, Sep. 1994.

[139] W. Yu and T. Lan, "Transmitter optimization for the multi-antenna downlink with per-antenna power constraints," *IEEE Trans. Sig. Proc.* pp. 2646 - 2660, Jun. 2007.

[140] X. Zhang, *Matrix inequalities,* Springer, 2002.

[141] J. Zhang, A. Kavcic, X. Ma, and K. M. Wong, "Design of unitary precoders for ISI channels," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. III, pp. 2265 - 2268, Orlando, FL, 2002.

[142] J. Zhang, A. Kavcic, and K. M. Wong, "Equal-diagonal QR decomposition and its application to precoder design for successive-cancellation detection," *IEEE Trans. Info. Theory,* pp. 154 - 172, Jan. 2005.

[143] X. Zhang, D. P. Palomar, and B. Ottersten, "Robust Design of Linear MIMO Transceivers under Channel Uncertainty," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006.