Foundations and Trends® in
Communications and Information Theory
3:4-5 (2006)

MIMO Transceiver Design via Majorization Theory

Daniel P. Palomar and Yi Jiang

now
the essence of knowledge

MIMO Transceiver Design via Majorization Theory

MIMO Transceiver Design via Majorization Theory

Daniel P. Palomar

Dept. Electronic and Computer Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon
Hong Kong
palomar@ust.hk

Yi Jiang

Dept. Electrical and Computer Engineering
University of Colorado
Boulder, Colorado 80309
USA
yjiang@dsp.colorado.edu

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in Communications and Information Theory

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is D. P. Palomar and Y. Jiang, MIMO Transceiver Design via Majorization Theory, Foundation and Trends® in Communications and Information Theory, vol 3, nos 4-5, pp 331–551, 2006

Printed on acid-free paper

ISBN: 978-1-60198-030-4

Foundations and Trends® in Communications and Information Theory

Volume 3 Issues 4-5, 2006

Editorial Scope

Foundations and Trends® in Communications and Information Theory will publish survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and Equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and Image Compression
- Wireless Communications

Information for Librarians

Foundations and Trends® in Communications and Information Theory, 2006, Volume 3, 6 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328. Also available as a combined paper and online subscription.

Foundations and Trends® in
Communications and Information Theory
Vol. 3, Nos. 4-5 (2006) 331–551
© 2007 D. P. Palomar and Y. Jiang
DOI: 10.1561/0100000018

MIMO Transceiver Design via Majorization Theory

Daniel P. Palomar[1] and Yi Jiang[2]

[1] *Dept. Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, palomar@ust.hk*
[2] *Dept. Electrical and Computer Engineering, University of Colorado, Boulder, Colorado 80309, USA, yjiang@dsp.colorado.edu*

Abstract

Multiple-input multiple-output (MIMO) channels provide an abstract and unified representation of different physical communication systems, ranging from multi-antenna wireless channels to wireless digital subscriber line systems. They have the key property that several data streams can be simultaneously established.

In general, the design of communication systems for MIMO channels is quite involved (if one can assume the use of sufficiently long and good codes, then the problem formulation simplifies drastically). The first difficulty lies on how to measure the global performance of such systems given the tradeoff on the performance among the different data streams. Once the problem formulation is defined, the resulting mathematical problem is typically too complicated to be optimally solved as it is a matrix-valued nonconvex optimization problem. This design problem has been studied for the past three decades (the first papers dating back to the 1970s) motivated initially by cable systems and more recently by wireless multi-antenna systems. The approach was to

choose a specific global measure of performance and then to design the system accordingly, either optimally or suboptimally, depending on the difficulty of the problem.

This text presents an up-to-date unified mathematical framework for the design of point-to-point MIMO transceivers with channel state information at both sides of the link according to an *arbitrary* cost function as a measure of the system performance. In addition, the framework embraces the design of systems with given individual performance on the data streams.

Majorization theory is the underlying mathematical theory on which the framework hinges. It allows the transformation of the originally complicated matrix-valued nonconvex problem into a simple scalar problem. In particular, the *additive* majorization relation plays a key role in the design of *linear* MIMO transceivers (i.e., a linear precoder at the transmitter and a linear equalizer at the receiver), whereas the *multiplicative* majorization relation is the basis for *nonlinear decision-feedback* MIMO transceivers (i.e., a linear precoder at the transmitter and a decision-feedback equalizer at the receiver).

Contents

1

Introduction

This chapter starts by introducing in a concise way the concept and relevance of multiple-input multiple-output (MIMO) channels and by highlighting some of the successful schemes for MIMO communication systems that have been proposed such as space–time coding and linear precoding. Then, a first glimpse at linear transceivers is presented, starting from the classical receive beamforming schemes in *smart antennas* and gradually building on top in a natural way. Finally, a historical account on MIMO transceivers is outlined.

1.1 MIMO Channels

MIMO channels arise in many different scenarios such as wireline systems or multi-antenna wireless systems, where there are multiple transmit and receive dimensions. A MIMO channel is mathematically denoted by a channel matrix which provides an elegant, compact, and unified way to represent physical channels of completely different nature.

The use of multiple dimensions at both ends of a communication link offers significant improvements in terms of *spectral efficiency* and *link reliability*. The most important characteristic of MIMO channels is the *multiplexing gain*, obtained by exploiting the multiple dimensions to open up several parallel *subchannels* within the MIMO channel, also termed channel *eigenmodes*, which leads to an increase of rate. The multiplexing property allows the transmission of several symbols simultaneously or, in other words, the establishment of several *substreams* for communication.

1.1.1 Basic Signal Model

The transmission over a general MIMO communication channel with n_T transmit and n_R receive dimensions can be described with the baseband signal model

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

as depicted in Figure 1.1, where $\mathbf{s} \in \mathbb{C}^{n_T \times 1}$ is the transmitted vector, $\mathbf{H} \in \mathbb{C}^{n_R \times n_T}$ is the channel matrix, $\mathbf{y} \in \mathbb{C}^{n_R \times 1}$ is the received vector, and $\mathbf{n} \in \mathbb{C}^{n_R \times 1}$ denotes the noise.

A multicarrier MIMO channel can be similarly described, either explicitly for the N carriers as

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{s}_k + \mathbf{n}_k \quad 1 \leq k \leq N, \tag{1.2}$$

or implicitly as in (1.1) by defining the block-diagonal equivalent matrix $\mathbf{H} = \operatorname{diag}(\{\mathbf{H}_k\})$.

When $n_T = 1$, the MIMO channel reduces to a single-input multiple-output (SIMO) channel (e.g., with multiple antennas only at the receiver). Similarly, when $n_R = 1$, the MIMO channel reduces to a

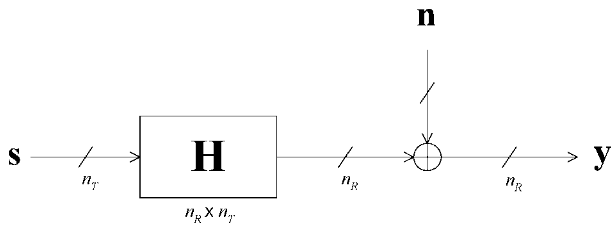


Fig. 1.1 Scheme of a MIMO channel.

multiple-input single-output (MISO) (e.g., with multiple antennas only at the transmitter). When both $n_T = 1$ and $n_R = 1$, the MIMO channel simplifies to a simple scalar or single-input single-output (SISO) channel.

1.1.2 Examples of MIMO Channels

We now briefly illustrate how different physical communication channels can be conveniently modeled as a MIMO channel.

1.1.2.1 Inter-Symbol Interference (ISI) Channel

Consider the discrete-time signal model after symbol-rate sampling

$$y(n) = \sum_{k=0}^{L} h(k)\, s(n-k) + n(n), \tag{1.3}$$

where $h(k)$ are the coefficients of the finite-impulse response (FIR) filter of order L representing the channel.

If the transmitter inserts at least L zeros between blocks of N symbols (termed zero-padding), the MIMO channel model in (1.1) is obtained where the channel matrix $\mathbf{H}$ is a convolutional matrix [122, 131]:

$$\mathbf{H} = \begin{bmatrix} h(0) & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ h(L) & & \ddots & 0 \\ 0 & \ddots & & h(0) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h(L) \end{bmatrix}. \tag{1.4}$$

Alternatively, if the transmitter uses a cyclic prefix of at least L symbols between blocks of N symbols, then the linear convolution becomes a circular convolution and the MIMO channel model in (1.1) is obtained where the channel matrix $\mathbf{H}$ is a circulant matrix[1] [122, 131].

[1] In a circulant matrix, the rows are composed of cyclically shifted versions of a sequence [66].

1.1.2.2 Multicarrier Channel

In a multicarrier communication system, the available bandwidth is partitioned into N subbands and then each subband is independently used for transmission [17, 80]. Such an approach not only simplifies the communication process but it is also a capacity-achieving structure for a sufficiently large N [46, 62, 122].

The signal model follows from a block transmission with a cyclic prefix, obtaining a circulant matrix, combined with an inverse/direct discrete Fourier transform (DFT) at the transmitter/receiver. The MIMO channel model in (1.1) is obtained where the channel matrix $\mathbf{H}$ is a diagonal matrix with diagonal elements given by DFT coefficients [54, 86].

1.1.2.3 Multi-Antenna Wireless Channel

The multi-antenna wireless channel with multiple antennas at both sides of the link (see Figure 1.2) is the paradigmatic example of a MIMO channel. In fact, the publication of [43, 146, 148] in the late 1990s about multi-antenna wireless channels boosted the research on MIMO systems. The popularity of this particular scenario is mainly due to the linear increase of capacity with the number of antennas [43, 148] for the same bandwidth.

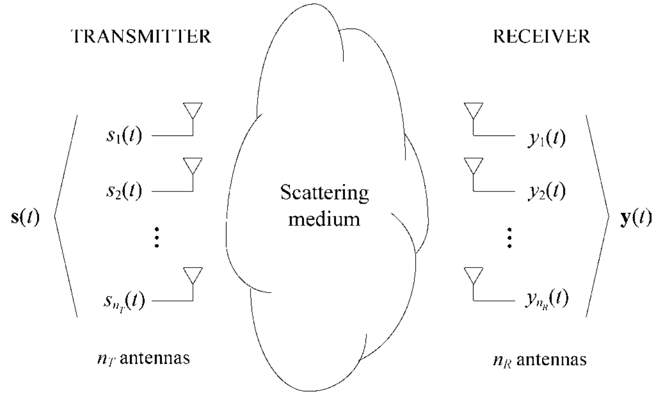


Fig. 1.2 Example of a MIMO channel arising in wireless communications when multiple antennas are used at both the transmitter and the receiver.

If the channel is flat in frequency, then the MIMO channel model in (1.1) follows naturally by defining the (i,j)th element of matrix $\mathbf{H}$ as the channel gain/fading between the jth transmit antenna and the ith receive one. In general, however, the channel will be frequency-selective according to the following matrix convolution:

$$\mathbf{y}(n) = \sum_{k=0}^{L} \mathbf{H}(k)\,\mathbf{s}(n-k) + \mathbf{n}(n) \tag{1.5}$$

where $\mathbf{H}(n)$ are the matrix-coefficients of the FIR matrix filter representing the channel ($[\mathbf{H}(n)]_{ij}$ is the discrete-time channel from the jth transmit antenna to the ith receive one). At this point, the frequency-selective channel in (1.5) can be manipulated as in (1.3) to obtain a block-matrix with each block corresponding to the channel between each transmit–receive pair of antennas; in particular, with zero padding each block will be a convolutional matrix, whereas with cyclic prefix each block will be a circulant matrix [122]. In the case of cyclic prefix, after applying the inverse/direct DFT to each block and a posterior rearrangement of the elements, the multicarrier MIMO signal model in (1.2) is obtained [122], i.e., one basic MIMO channel per carrier.

1.1.2.4 Wireline DSL Channel

Digital Subscriber Line technology has gained popularity as a broadband access technology capable of reliably delivering high data rates over telephone subscriber lines [144]. Modeling a DSL system as a MIMO channel presents many advantages with respect to treating each twisted pair independently [47, 63]. If fact, modeling a wireline channel as a MIMO channel was done three decades ago [90, 129].

The dominant impairment in DSL systems is crosstalk arising from electromagnetic coupling between neighboring twisted-pairs. Near-end crosstalk (NEXT) comprises the signals originated in the same side of the received signal (due to the existence of downstream and upstream transmission) and far-end crosstalk (FEXT) includes the signal originated in the opposite side of the received signal. The impact of NEXT is generally suppressed by employing frequency division duplex (FDD) to separate downstream and upstream transmission.

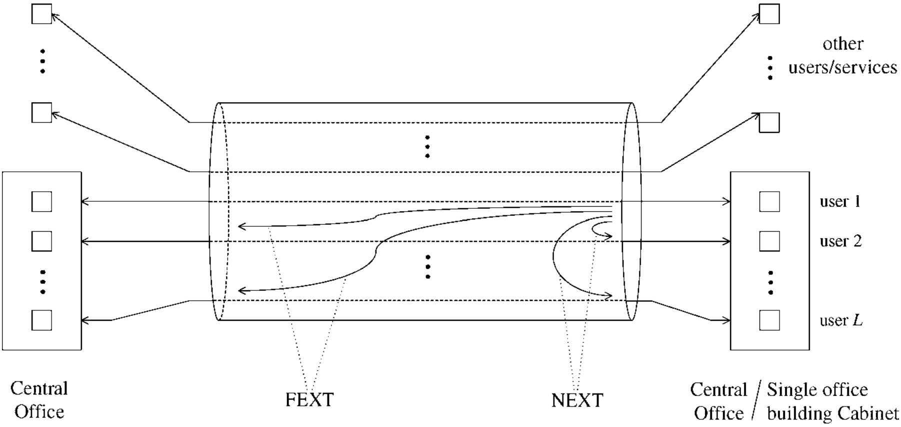


Fig. 1.3 Scheme of a bundle of twisted pairs of a DSL system.

The general case under analysis consists of a binder group composed of L users in the same physical location plus some other users that possibly belong to a different service provider and use different types of DSL systems (see Figure 1.3). The MIMO channel represents the communication of the L intended users while the others are treated as interference.

DSL channels are highly frequency-selective with a signal model as in (1.5); as a consequence, practical communication systems are based on the multicarrier MIMO signal model in (1.2).

1.1.2.5 CDMA Channel

Excess-bandwidth systems (the majority of practical systems) utilize a transmit bandwidth larger than the minimum (Nyquist) bandwidth. Examples are systems using spreading codes and systems using a root-raised cosine transmit shaping pulse (with a nonzero rolloff factor) [120]. For these systems, fractional-rate sampling (sampling at a rate higher than the symbol rate) has significant practical advantages compared to symbol-rate sampling such as the insensitivity with respect to the sampling phase and the possibility to implement in discrete time many of the operations performed at the receiver such as the matched-filtering operation (cf. [121]). Fractionally sampled systems can be modeled as a multirate convolution which can be easily converted into a more convenient vector convolution as in (1.5).

One relevant example of excess-bandwidth system is code division multiple access (CDMA) systems, where multiple users transmit overlapping in time and frequency but using different signature waveforms or spreading codes (which are excess-bandwidth shaping pulses). The discrete-time model for such systems is commonly obtained following a matched filtering approach by sampling at the symbol rate the output of a bank of filters where each filter is matched to one of the signature waveforms [157]. An alternative derivation of the discrete-time model for CDMA systems is based on a fractionally sampled scheme by sampling at the chip rate. Adding up the effect of U users, the final discrete-time (noiseless) signal model is

$$\mathbf{y}(n) = \sum_{u=1}^{U} \sum_{l=0}^{L} \mathbf{h}_u(l)\, s_u(n-l), \tag{1.6}$$

where $\mathbf{h}_u(n)$ is the equivalent chip-rate sampled channel of the uth user defined as $\mathbf{h}_u(n) \triangleq [h_u(nP), \ldots, h_u(nP+(P-1))]^T$, $h_u(n)$ corresponds to the continuous impulse response $h_u(t)$ sampled at time $t = nT/P$, P denotes the oversampling factor or spreading factor, and L is the support of the channel $\mathbf{h}_u(n)$.

1.2 MIMO Communication Systems

A plethora of communication techniques exists for transmission over MIMO channels which essentially depend on the degree of channel state information (CSI) available at the transmitter and at the receiver. Clearly, the more channel information, the better the performance of the system. The reader interested in space–time wireless communication systems is referred to the two 2003 textbooks [87, 115] and to the more extensive 2005 textbooks [16, 50, 151].

CSI at the receiver (CSIR) is traditionally acquired via the transmission of a training sequence (pilot symbols) that allows the estimation of the channel. It is also possible to use blind methods that do not require any training symbols but exploit knowledge of the structure of the transmitted signal or of the channel. CSI at the transmitter (CSIT) is typically obtained either via a feedback channel from the receiver (this technique requires the channel to be sufficiently slowly varying and has a loss in spectral efficiency due to the utilization of part of the bandwidth to transmit the channel state) or by exploiting (whenever possible) the channel reciprocity that allows to infer the channel from previous receive measurements (cf. [10]).

It is generally assumed that perfect CSIR is available. Regarding CSIT, there are two main families of transmission methods that consider either no CSIT or perfect CSIT. In practice, however, it is more realistic to consider imperfect or partial CSIT.

1.2.1 Schemes with No CSIT

Space–time coding generalizes the classical concept of coding on the temporal domain [22] to coding on both spatial and temporal dimensions [1, 146]. The idea is to introduce redundancy in the transmitted signal, both over space and time, to allow the receiver to recover the signal even in difficult propagation situations. The conventional space–time coding trades off spectral efficiency for improved communication reliability. Since the initial papers in 1998 [1, 146], an extraordinary number of publications has flourished in the literature (cf. [37, 38, 87, 105]). The recent space–time block codes proposed in [38] and [105] can achieve the optimum tradeoff between spectral efficiency and transmission reliability, or the diversity–multiplexing gain tradeoff as charted in [178]. The better performance of the advanced space–time codes come with high decoding complexity.

Layered architectures (also termed BLAST[2]) refer to a particular case of a space–time coding when a separate coding scheme is used for each spatial branch, i.e., they are constructed by assembling one-dimensional constituent codes. The diagonal BLAST originally proposed by Foschini in 1996 [45] can in principle achieve the optimal diversity–multiplexing gain tradeoff [178]. However it requires short and powerful coding to eliminate error propagation, which makes it difficult to implement. The simpler vertical BLAST proposed in [44] admits independent coding and decoding for each spatially multiplexed

[2] BLAST stands for Bell-labs LAyered Space–Time architecture [44, 45].

substream, but the simplicity in the equalization and decoding aspects comes with low reliability since vertical BLAST does not collect the diversity across different layers. Hybrid schemes combining layered architectures with constituent space–time codes have been proposed as a reasonable tradeoff between performance and complexity, e.g., [5].

1.2.2 Schemes with Perfect CSIT

When perfect CSIT is available, the transmission can be adapted to each channel realization using signal processing techniques. Historically speaking, there are two main scenarios that have motivated the development of communication methods for MIMO channels with CSIT: wireline channels, and wireless channels.

The initial motivation to design techniques for communication over MIMO channels can be found in wireline systems by treating all the links within a bundle of cables as a whole, e.g., [63, 90, 129, 170, 171]. Another more recent source of motivation to design methods for communication over MIMO channels follows from multi-antenna wireless systems e.g., [3, 122, 131]. A historical perspective on signal processing methods for MIMO systems is given in Section 1.4.

1.2.3 Schemes with Imperfect/Partial CSIT

In real scenarios, it is seldom the case that the CSIT is either inexistent or perfect; in general, its knowledge is partial or imperfect for which hybrid communication schemes are more appropriate.

One basic approach is to start with a space–time code, for which no CSIT is required, and combine it with some type of signal processing technique to take advantage of the partial CSIT, e.g., [76].

Another different philosophy is to start with a signal processing approach, for which typically perfect CSIT is assumed, and make it robust to imperfections in the CSIT, e.g., [10, 101, 123, 159, 165].

1.3 A First Glimpse at Linear Transceivers: Beamforming

Beamforming is a term traditionally associated with array processing or *smart antennas* in wireless communications where an array of antennas exists either at the transmitter or at the receiver [75, 85, 99, 150, 154].

The concept of linear MIMO transceiver is closely related to that of classical beamforming as shown next.

1.3.1 Classical Beamforming for SIMO and MISO Channels

We consider the concept of beamforming over any arbitrary dimension, generalizing the traditional meaning that refers only to the space (antenna) dimension.

Consider a SIMO channel:

$$\mathbf{y} = \mathbf{h}x + \mathbf{n}, \tag{1.7}$$

where one symbol x is transmitted (normalized such that $\mathbb{E}\left[|x|^2\right] = 1$) and a vector $\mathbf{y}$ is received (the noise is assumed zero mean and white $\mathbb{E}\left[\mathbf{n}\mathbf{n}^\dagger\right] = \mathbf{I}$). The classical receive beamforming approach estimates the transmitted symbol by linearly combining the received vector via the beamvector $\mathbf{w}$:

$$\hat{x} = \mathbf{w}^\dagger \mathbf{y} = \mathbf{w}^\dagger \left(\mathbf{h}x + \mathbf{n}\right). \tag{1.8}$$

We can now design the receive beamvector $\mathbf{w}$ to maximize the SNR given by

$$\mathrm{SNR} = \frac{|\mathbf{w}^\dagger \mathbf{h}|^2}{\mathbf{w}^\dagger \mathbf{w}}. \tag{1.9}$$

The solution follows easily from the Cauchy–Schwarz's inequality:

$$|\mathbf{w}^\dagger \mathbf{h}| \leq \|\mathbf{w}\| \, \|\mathbf{h}\|, \tag{1.10}$$

where equality is achieved when $\mathbf{w} \propto \mathbf{h}$, i.e., when the beamvector is aligned with the channel. This is commonly termed *matched filter* or *maximum ratio combining*. The resulting SNR is then given by the squared-norm of the channel $\|\mathbf{h}\|^2$, i.e., fully utilizing the energy of the channel.

Consider now a MISO channel:

$$y = \mathbf{h}^\dagger \mathbf{s} + n, \tag{1.11}$$

where the vector signal $\mathbf{s}$ is transmitted and the scalar y is received (the noise is assumed zero mean and normalized $\mathbb{E}\left[|n|^2\right] = 1$). The classical transmit beamforming approach transmits on each antenna a weighted version of the symbol to be conveyed x via the beamvector $\mathbf{p}$:

$$\mathbf{s} = \mathbf{p}x, \tag{1.12}$$

where the transmitted power is given by the squared-norm of the beamvector $\|\mathbf{p}\|^2$ (assuming $\mathbb{E}\left[|x|^2\right] = 1$). The overall signal model is then

$$y = (\mathbf{h}^\dagger \mathbf{p})x + n. \tag{1.13}$$

We can now design the transmit beamvector $\mathbf{p}$ to maximize the SNR

$$\mathrm{SNR} = |\mathbf{h}^\dagger \mathbf{p}|^2, \tag{1.14}$$

subject to a power constraint $\|\mathbf{p}\|^2 \leq P_0$. The solution again follows easily from the Cauchy–Schwarz's inequality:

$$|\mathbf{h}^\dagger \mathbf{p}| \leq \|\mathbf{h}\| \, \|\mathbf{p}\| \leq \|\mathbf{h}\| \sqrt{P_0}, \tag{1.15}$$

where both equalities are achieved when $\mathbf{p} = \sqrt{P_0}\,\mathbf{h}/\|\mathbf{h}\|$, i.e., when the beamvector is aligned with the channel and satisfies the power constraint with equality. An alternative way to derive this result is by rewriting the SNR as

$$\mathrm{SNR} = \mathbf{p}^\dagger (\mathbf{h}\mathbf{h}^\dagger)\mathbf{p}, \tag{1.16}$$

from which the maximum value follows straightforwardly as the eigenvector of matrix $\mathbf{h}\mathbf{h}^\dagger$ corresponding to the maximum eigenvalue, which is precisely $\mathbf{h}/\|\mathbf{h}\|$, properly normalized to satisfy the power constraint. The resulting SNR is then given by $P_0 \|\mathbf{h}\|^2$, i.e., fully utilizing the energy of the channel and the maximum power at the transmitter.

1.3.2 Single Beamforming for MIMO Channels

We are now ready to extend the previous treatment of classical beamforming only at the receiver or only at the transmitter to both sides of the link as illustrated in Figure 1.4 (e.g., [3, 113]). Consider now a MIMO channel:

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}, \tag{1.17}$$

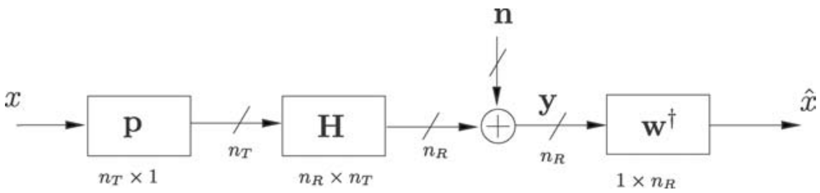


Fig. 1.4 Single beamforming scheme of a MIMO communication system.

where the vector signal $\mathbf{s}$ is transmitted and a vector $\mathbf{y}$ is received (the noise is assumed zero mean and white $\mathbb{E}\left[\mathbf{n}\mathbf{n}^\dagger\right] = \mathbf{I}$). The transmit beamforming generates the vector signal with beamvector $\mathbf{p}$ as

$$\mathbf{s} = \mathbf{p}x, \tag{1.18}$$

where one symbol x is transmitted (normalized such that $\mathbb{E}\left[|x|^2\right] = 1$), and the receive beamforming estimates the transmitted symbol by linearly combining the received vector with the beamvector $\mathbf{w}$:

$$\hat{x} = \mathbf{w}^\dagger\mathbf{y} = \mathbf{w}^\dagger\left(\mathbf{H}\mathbf{p}x + \mathbf{n}\right). \tag{1.19}$$

The SNR is given by

$$\mathrm{SNR} = \frac{|\mathbf{w}^\dagger\mathbf{H}\mathbf{p}|^2}{\mathbf{w}^\dagger\mathbf{w}}. \tag{1.20}$$

We can now maximize it with respect to the receive beamvector $\mathbf{w}$, for a given fixed $\mathbf{p}$, exactly as in the case of a classical receive beamforming. From the Cauchy–Schwarz's inequality we have that the optimum receiver is $\mathbf{w} \propto \mathbf{H}\mathbf{p}$, i.e., when the beamvector is aligned with the effective channel $\mathbf{h} = \mathbf{H}\mathbf{p}$. The resulting SNR is then given by

$$\mathrm{SNR} = \mathbf{p}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{p}. \tag{1.21}$$

The transmit beamvector $\mathbf{p}$ that maximizes this expression is, as in the classical transmit beamforming, the eigenvector of matrix $\mathbf{H}^\dagger\mathbf{H}$ corresponding to the maximum eigenvalue (or, equivalently, to the right singular vector of the channel matrix $\mathbf{H}$ corresponding to the maximum singular value, denoted by $\mathbf{v}_{H,\max}$), properly normalized to satisfy the power constraint with equality: $\mathbf{p} = \sqrt{P_0}\,\mathbf{v}_{H,\max}$. The final achieved SNR is $P_0\sigma_{H,\max}^2$, where $\sigma_{H,\max}$ denotes the maximum singular value.

Now that we know that the optimal transmitter is the best right singular vector, we can step back and elaborate on the optimal receiver $\mathbf{w} \propto \mathbf{Hp} = \sqrt{P_0}\,\mathbf{H}\mathbf{v}_{H,\max} = \sqrt{P_0}\,\sigma_{H,\max}\mathbf{u}_{H,\max}$ to realize that it is actually equal (up to an arbitrary scaling factor) to the best left singular vector of the channel matrix $\mathbf{H}$.

Summarizing, the best transmit–receive beamvectors correspond nicely to the right–left singular vectors of the channel matrix $\mathbf{H}$ associated to the largest singular value and the global communication process becomes

$$\hat{x} = \mathbf{w}^{\dagger}\left(\mathbf{Hp}x + \mathbf{n}\right) = (\sqrt{P_0}\,\sigma_{H,\max})x + n, \tag{1.22}$$

where n is an equivalent scalar noise with zero mean and unit variance.

1.3.3 Multiple Beamforming (Matrix Beamforming) for MIMO Channels: Problem Statement

As we have seen, obtaining the best transmit–receive beamvectors when transmitting one symbol over a MIMO channel is rather simple. However, precisely one of the interesting properties of MIMO channels is the multiplexing capability they exhibit. To properly take advantage of the potential increase in rate, we need to transmit more than one symbol simultaneously. We can easily extend the previous signal model to account for the simultaneous transmission of L symbols:

$$\mathbf{s} = \sum_{i=1}^{L} \mathbf{p}_i x_i = \mathbf{Px}, \tag{1.23}$$

where $\mathbf{P}$ is a matrix with columns equal to the transmit beamvectors $\mathbf{p}_i$ corresponding to the L transmitted symbols x_i stacked for convenience in vector $\mathbf{x}$ (normalized such that $\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\dagger}\right] = \mathbf{I}$). The power constraint in this case is

$$\sum_{i=1}^{L} \left\|\mathbf{p}_i\right\|^2 = \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \le P_0. \tag{1.24}$$

Similarly, each estimated symbol at the receiver is $\hat{x}_i = \mathbf{w}_i^{\dagger}\mathbf{y}$ or, more compactly,

$$\hat{\mathbf{x}} = \mathbf{W}^{\dagger}\mathbf{y}, \tag{1.25}$$

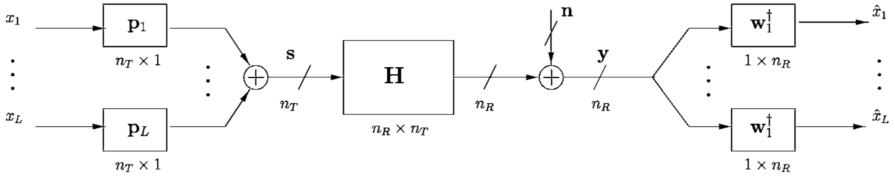


Fig. 1.5 Multiple beamforming interpretation of a MIMO communication system.

where $\mathbf{W}$ is a matrix with columns equal to the receive beamvectors $\mathbf{w}_i$ corresponding to the L transmitted symbols. We can either interpret this communication scheme as a *multiple beamforming* scheme as illustrated in Figure 1.5 or as a *matrix beamforming* scheme as shown in Figure 1.6. Both interpretations are actually natural extensions of the single beamforming scheme in Figure 1.4.

The design of the transmitter and receiver in the multiple beamforming case is fundamentally different from the single beamforming case. This happens because the L data streams are coupled and exhibit a tradeoff for two different reasons:

(i) The total power budget P_0 needs to be distributed among the different substreams.

(ii) Even for a given power allocation among the substreams, the design of the transmit "directions" is still coupled as the transmission of one symbol interferes the others, as can be seen from

$$\hat{x}_i = \mathbf{w}_i^\dagger \left(\mathbf{H}\mathbf{p}_i x_i + \mathbf{n}_i\right), \tag{1.26}$$

where $\mathbf{n}_i = \sum_{j \neq i} \mathbf{H}\mathbf{p}_j x_j + \mathbf{n}$ is the equivalent interference-plus-noise seen by the ith substream.

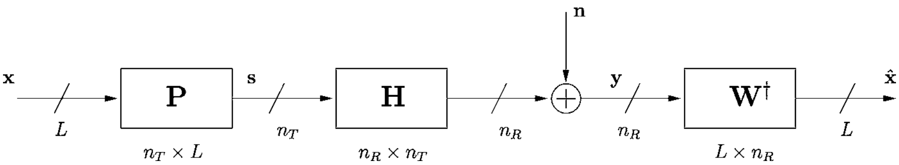


Fig. 1.6 Matrix beamforming interpretation of a MIMO communication system.

This inherent tradeoff among the substreams complicates the problem to the point that not even the problem formulation is clear: What objective should we consider to measure the system performance?

As a consequence, different authors have considered a variety of objective functions to design such systems (see the historical overview in Section 1.4). In some cases, deriving optimal solutions according to the selected objective and, in other cases, only giving suboptimal solutions due to the difficulty of the problem. It is important to mention that if we can assume the use of sufficiently long and good codes, then the problem formulation becomes rather simple as elaborated later in Section 1.4.

This text considers a general problem formulation based on an arbitrary objective function (alternatively, on individual constraints on the quality of each data stream) and develops a unified framework based on majorization theory that allows the simplification of the problem so that optimal solutions can be easily obtained.

1.3.4 Diagonal Transmission for MIMO Channels: A Heuristic Solution

Inspired by the solution in the single beamforming case, we can come up with a suboptimal strategy that simplifies the problem design a great deal. Recall that in the single beamforming scheme, the best transmit and receive beamvectors correspond to the right and left singular vectors of the channel matrix $\mathbf{H}$, respectively, associated to the largest singular value. In the multiple beamforming scheme, we can consider the natural extension and choose, for the ith substream, the right and left singular vectors of the channel matrix $\mathbf{H}$ associated to the ith largest singular value, $\mathbf{v}_{H,i}$ and $\mathbf{u}_{H,i}$, respectively:

$$\mathbf{p}_i = \sqrt{p_i}\,\mathbf{v}_{H,i} \quad \text{and} \quad \mathbf{w}_i = \mathbf{u}_{H,i}, \tag{1.27}$$

where p_i denotes the power allocated to the ith substream that must satisfy the power constraint $\sum_{i=1}^{L} p_i \leq P_0$.

With this choice of transmit–receive processing, the global communication process becomes diagonal or orthogonal (in the sense that the different substreams do not interfere with each other):

$$\begin{aligned} \hat{x}_i &= \mathbf{w}_i^\dagger \left(\mathbf{H}\mathbf{p}_i x_i + \mathbf{n}_i\right) && (1.28)\\ &= \sqrt{p_i}\,\sigma_{H,i}\, x_i + n_i \quad 1 \le i \le L, && (1.29) \end{aligned}$$

or, more compactly,

$$\begin{aligned} \hat{\mathbf{x}} &= \mathbf{W}^\dagger \left(\mathbf{H}\mathbf{P}\mathbf{x} + \mathbf{n}\right) && (1.30)\\ &= \mathrm{diag}\left(p_1,\ldots,p_L\right) \boldsymbol{\Sigma}_H \mathbf{x} + \mathbf{n}, && (1.31) \end{aligned}$$

where $\boldsymbol{\Sigma}_H$ is a diagonal matrix that contains the L largest singular values of $\mathbf{H}$ in decreasing order and $\mathbf{n}$ is an equivalent vector noise with zero mean and covariance matrix $\mathbb{E}\left[\mathbf{n}\mathbf{n}^\dagger\right] = \mathbf{I}$.

Since the substreams do not interfere with each other, we can nicely write the signal to interference-plus-noise ratios (SINRs) as

$$\mathrm{SINR}_i = p_i\,\sigma_{H,i}^2 \quad 1 \le i \le L \tag{1.32}$$

and the only remaining problem is to find the appropriate power allocation $\{p_i\}$ which will depend on the particular objective function chosen to measure the performance of the system.

Fortunately, as will be shown in this text, we do not need to content ourselves with suboptimal solutions and we can aim for the global solution.

1.4 Historical Perspective on MIMO Transceivers

The problem of jointly designing the transmit and receive signal processing is an old one. Already in the 1960s, we can easily find papers that jointly design transmit–receive filters for frequency-selective SISO channels to minimize the MSE (e.g., [11, 128] and references therein). The design of MIMO transceivers for communication systems dates back to the 1970s, where cable systems were the main application [90, 129].

The design of MIMO systems is generally quite involved since several substreams are typically established over MIMO channels (multiplexing property). Precisely, the existence of several substreams, each with its own performance, makes the definition of a global measure of the system performance not clear; as a consequence, a wide span of different design criteria has been explored in the literature as overviewed next.

At this point, it is important to emphasize that if one can assume the use of sufficiently long and good codes, i.e., if instead of a signal processing approach we adopt an information theoretic perspective, then the problem formulation simplifies drastically and the state of the art of the problem is very different. As first devised by Shannon in 1949 [140] for frequency-selective channels and rigorously formalized for a matrix channel in [21, 152, 153], the best transmission scheme that achieves the channel capacity consist of: (i) diagonalizing the channel matrix, (ii) using a waterfilling power allocation over the channel eigenmodes, and (iii) employing a Gaussian signaling (see also [33, 122, 148]). In many real systems, however, rather than with Gaussian codes, the transmission is done with practical discrete constellations and coding schemes.

In a more general setup, we can formulate the design of the MIMO system as the optimization of a global objective function based on the individual performance of each of the established substreams. Alternatively, we can consider the achievable set of individual performance of the substreams (e.g., in a CDMA system where each user has some minimum performance constraint). The classical aforementioned information-theoretic solution will be then a particular case of this more general setup.

The first linear designs (for cable systems) considered a mathematically tractable cost function as a measure of the system performance: the sum of the MSEs of all channel substreams or, equivalently, the trace of the MSE matrix [2, 90, 129, 171] (different papers explored variations of the problem formulation concerning the dimensions of the channel matrix, the channel frequency-selectivity, the excess-bandwidth, etc.). Decision-feedback schemes were also considered [128, 82, 170].

Due to the popularization of wireless multi-antenna MIMO systems in the late 1990s [43, 45, 122, 148], a new surge of interest on the design of MIMO transceivers appeared with a wireless rather than wired motivation. Different design criteria have been used by different authors as shown in the following. In [131] a unified signal model for block transmissions was presented as a MIMO system and different design criteria were considered such as the minimization of the trace of the MSE matrix and the maximization of the SINR with a zero-forcing (ZF) constraint. In [170], the minimization of the determinant of the MSE matrix was considered for decision-feedback (DF) schemes. In [130], a reverse-engineering approach was taken to obtain different known solutions as the minimization of the weighted trace of the MSE matrix with appropriate weights. In [3], the flat multi-antenna MIMO case was considered providing insights from the point of view of beamforming. Various criteria were considered in [132] under average power constraint as well as peak power constraint.

For the aforementioned design criteria, the problem is very complicated but fortunately it simplifies because the channel matrix turns out to be diagonalized by the optimal transmit–receive processing and the transmission is effectively performed on a diagonal or parallel fashion. Indeed, the diagonal transmission implies a *scalarization* of the problem (meaning that all matrix equations are substituted with scalar ones) with the consequent simplification (cf. Section 1.3.4). In light of the optimality of the diagonal structure for transmission in all the previous examples (including the capacity-achieving solution [33, 122, 148]), one might expect that the same would hold for any other criteria as well. However, as shown in [111], this is not the case.

More recently, the design of MIMO transceivers has been approached using the bit error rate (BER), rather than the MSE or the SINR, as basic performance measure. This approach is arguably more relevant as the ultimate performance of a system is measured by the (BER), but it is also more difficult to handle. In [106], the minimization of the BER (and also of the Chernoff upper bound) averaged over the channel substreams was treated in detail when a diagonal structure is imposed. The minimum BER design of a linear MIMO transceiver without the diagonal structure constraint was independently obtained in [36] and [111], resulting in an optimal *nondiagonal* structure. This result, however, only holds when the constellations used in all the substreams are equal.

In [111], a general unifying framework was developed that embraces a wide range of different design criteria for linear MIMO transceivers;

in particular, the optimal design was obtained for the family of Schur-concave and Schur-convex cost functions which arise in majorization theory [97]. Interestingly, this framework gives a clear answer to the question of whether the diagonal transmission is optimal: when the cost function is Schur-concave then the diagonal structure is optimal, but when the cost function is Schur-convex then the optimal structure is not diagonal anymore.

From the previous unifying framework based on majorization theory, it follows that the minimization of the BER averaged over the substreams, considered in [36, 111], is a Schur-convex function, provided that the constellations used on the substreams are equal, and therefore it can be optimally solved. The general case of different constellations, however, is much more involved (in such a case, the cost function is neither Schur-convex nor Schur-concave) and was solved in [110] via a primal decomposition approach, a technique borrowed from optimization theory [12, 88, 142].

An alternative way to formulate the design of MIMO transceivers is to consider an independent requirement of quality for each of the substreams rather than a global measure of quality. This was considered and optimally solved in [114], again based on majorization theory.

Interestingly, the unifying framework based on the majorization theory was later extended to nonlinear DF MIMO transceivers in [74] (see also [141]). The extension in [74] is based on a new matrix decomposition, namely, the generalized triangular decomposition [70]. While the linear transceiver design relies on the concept of *additive* majorization, the nonlinear decision-feedback transceiver invokes the *multiplicative* majorization. One can see an intriguing mathematical symmetry between the linear and nonlinear designs.

As evidenced by the previous results, majorization theory is a mathematical tool that plays a key role in transforming the originally complicated matrix-valued nonconvex problem into a simple scalar problem. Other recent successful applications of majorization theory in communication systems, from either an information-theoretic or a signal processing perspective, include the design of signature sequences in CDMA systems to maximize the sum-rate or to satisfy Quality-of-Service (QoS) requirements with minimum power by Viswanath *et al.* [161, 162, 163] and the study of the impact of correlation of the transmit antennas in MISO systems by Boche *et al.* [18, 79].

1.5 Outline

This text considers the design of point-to-point MIMO transceivers (this also includes multiuser CDMA systems) with CSI at both sides of the link according to an arbitrary cost function as a measure of the system performance. A unified framework is developed that hinges on majorization theory as a key tool to transform the originally complicated matrix-valued nonconvex problem into a simple scalar problem in most cases convex which can be addressed under the powerful framework of convex optimization theory [12, 13, 20]. The framework allows the choice of any cost function as a measure of the overall system performance and the design is based then on the minimization of the cost function subject to a power constraint or vice versa. In addition, the framework embraces the possibility of imposing a set of QoS constraints for the data streams with minimum required power.

This chapter has already given the basic background on MIMO channels and MIMO communication systems, including a natural evolution from classical beamforming to MIMO transceivers and a historical perspective on MIMO transceivers.

Chapter 2 introduces majorization theory on which the rest of the text is based.

Chapter 3 is fully devoted to *linear* MIMO transceivers composed of a linear precoder at the transmitter and a linear equalizer at the receiver. In particular, the key simplification relies on the *additive* majorization relation. Different types of design are considered in order of increasing conceptual and mathematical complexity: (i) based on a Schur-concave/convex cost function as a global measure of performance, (ii) based on individual QoS constraints, and (iii) based on an arbitrary cost function as a global measure of performance.

Then, Chapter 4 considers *nonlinear DF* MIMO transceivers, composed of a linear precoder at the transmitter and a decision-feedback equalizer (DFE) at the receiver (consisting of a feedforward stage and a feedback stage) or the dual form based on dirty paper coding by uplink–

downlink duality. Interestingly, the key simplification relies in this case on a *multiplicative* majorization relation. As in the linear case, different types of design are considered: (i) based on an arbitrary cost function as a global measure of performance (including Schur-concave/convex cost functions) and (ii) based on individual QoS constraints.

Hence, from Chapters 3 and 4, both the linear and nonlinear cases are nicely unified under an additive and multiplicative majorization relation. The basic design of point-to-point linear and nonlinear DF MIMO transceivers with CSI is thus well understood. This is not to say that the general problem of MIMO transceivers with CSI is fully solved. On the contrary, there are still many unanswered questions and future lines of research.

Chapter 5 precisely describes unanswered questions and future lines of research, namely, (i) the design of *multiuser* MIMO transceivers for networks with interfering users, (ii) the design of *robust* MIMO transceivers to imperfect CSI, (iii) the design of nonlinear MIMO transceivers with ML decoding, and (iv) the design of MIMO transceivers from an information-theoretic perspective with arbitrary constellations.

Notation. The following notation is used. Boldface upper-case letters denote matrices, boldface lower-case letters denote column vectors, and italics denote scalars. $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ represent the set of $m \times n$ matrices with real- and complex-valued entries, respectively. $\mathbb{R}_+$ and $\mathbb{R}_{++}$ stand for the set of nonnegative and positive real numbers, respectively. The super-scripts $(\cdot)^T$, $(\cdot)^*$, and $(\cdot)^\dagger$ denote matrix transpose, complex conjugate, and Hermitian operations, respectively. $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ denote the real and imaginary part, respectively. $\mathrm{Tr}(\cdot)$ and $\det(\cdot)$ (also $|\cdot|$) denote the trace and determinant of a matrix, respectively. $\|\mathbf{x}\|$ is the Euclidean norm of a vector $\mathbf{x}$ and $\|\mathbf{X}\|_F$ is the Frobenius norm of a matrix $\mathbf{X}$ (defined as $\|\mathbf{X}\|_F \triangleq \sqrt{\mathrm{Tr}(\mathbf{X}^\dagger\mathbf{X})}$). $[\mathbf{X}]_{i,j}$ (also $[\mathbf{X}]_{ij}$) denotes the (ith, jth) element of matrix $\mathbf{X}$. $\mathbf{d}(\mathbf{X})$ and $\boldsymbol{\lambda}(\mathbf{X})$ denote the diagonal elements and eigenvalues, respectively, of matrix $\mathbf{X}$. A block-diagonal matrix with diagonal blocks given by the set $\{\mathbf{X}_k\}$ is denoted by $\mathrm{diag}(\{\mathbf{X}_k\})$. The operator $(x)^+ \triangleq \max(0,x)$ is the projection onto the nonnegative orthant.

2

Majorization Theory

Many of the problems addressed in this text result in complicated nonconvex constrained optimization problems that involve matrix-valued variables. Majorization theory is a key tool that allows us to transform these problems into simple problems with scalar variables that can be easily solved.

In this chapter, we introduce the basic notion of majorization and state some basic results. A complete and superb reference on the subject is the 1979 book by Marshall and Olkin [97].[1] The 1997 book by Bhatia [15] contains significant material on majorization theory as well. Other textbooks on matrix and multivariate analysis may also include a section on majorization theory, e.g., [66, Sec. 4.3][2] and [4, Sec. 8.10].

2.1 Basic Definitions

Majorization makes precise the vague notion that the components of a vector $\mathbf{x}$ are "less spread out" or "more nearly equal" than the components of a vector $\mathbf{y}$.

[1] Interestingly, a revised version of the book, co-authored by Marshall, Olkin, and Arnold, will be published soon by Springer-Verlag.

[2] Be aware that the definition of majorization in [66, Def. 4.3.24] is the opposite of the more standard convention.

Definition 2.1. [97] For any vector $\mathbf{x} \in \mathbb{R}^n$, let

$$x_{[1]} \geq \cdots \geq x_{[n]} \tag{2.1}$$

denote its components in decreasing order, and let

$$x_{(1)} \leq \cdots \leq x_{(n)} \tag{2.2}$$

denote its components in increasing order.

Definition 2.2. [97, 1.A.1] For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we say $\mathbf{x}$ is majorized by $\mathbf{y}$ or $\mathbf{y}$ majorizes $\mathbf{x}$ (denoted by $\mathbf{x} \prec \mathbf{y}$ or $\mathbf{y} \succ \mathbf{x}$) if

$$\begin{aligned} \sum_{i=1}^{k} x_{[i]} &\leq \sum_{i=1}^{k} y_{[i]} \quad 1 \leq k < n \\ \sum_{i=1}^{n} x_{[i]} &= \sum_{i=1}^{n} y_{[i]}. \end{aligned} \tag{2.3}$$

Alternatively, the previous conditions can be rewritten as

$$\begin{aligned} \sum_{i=1}^{k} x_{(i)} &\leq \sum_{i=1}^{k} y_{(i)} \quad 1 \leq k < n \\ \sum_{i=1}^{n} x_{(i)} &= \sum_{i=1}^{n} y_{(i)}. \end{aligned} \tag{2.4}$$

There are several equivalent characterizations of the majorization relation $\mathbf{x} \prec \mathbf{y}$ in addition to the conditions given in Definition 2.2. One is actually the answer of a question posed and answered in 1929 by Hardy, Littlewood, and Pólya [59, 60]: $\mathbf{y}$ majorizes $\mathbf{x}$ if

$$\sum_{i=1}^{n} \phi(x_i) \leq \sum_{i=1}^{n} \phi(y_i) \tag{2.5}$$

for all continuous convex functions ϕ. Another interesting characterization of $\mathbf{y} \succ \mathbf{x}$, also by Hardy, Littlewood, and Pólya [59, 60], is that

$\mathbf{x} = \mathbf{P}\mathbf{y}$ for some doubly stochastic matrix[3] $\mathbf{P}$. In fact, the previous characterization implies that the set of vectors $\mathbf{x}$ that satisfy $\mathbf{x} \prec \mathbf{y}$ is the convex hull spanned by the $n!$ points formed from the permutations of the elements of $\mathbf{y}$.[4] Yet another interesting characterization of $\mathbf{y} \succ \mathbf{x}$ is in the form of waterfilling:

$$\sum_{i=1}^{n} (x_i - a)^+ \leq \sum_{i=1}^{n} (y_i - a)^+ \tag{2.6}$$

for all $a \in \mathbb{R}$ and $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$, where $(u)^+ \triangleq \max(u,0)$. The interested reader is referred to [97, Ch. 4] for more characterizations.

Definition 2.3. [97, 1.A.2] For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we say $\mathbf{x}$ is weakly majorized by $\mathbf{y}$ or $\mathbf{y}$ weakly majorizes $\mathbf{x}$ (denoted by $\mathbf{x} \prec^w \mathbf{y}$ or $\mathbf{y} \succ^w \mathbf{x}$) if[5]

$$\sum_{i=1}^{k} x_{(i)} \geq \sum_{i=1}^{k} y_{(i)} \quad 1 \leq k \leq n \tag{2.7}$$

or, equivalently,

$$\sum_{i=k}^{n} x_{[i]} \geq \sum_{i=k}^{n} y_{[i]} \quad 1 \leq k \leq n. \tag{2.8}$$

Note that $\mathbf{x} \prec \mathbf{y}$ implies $\mathbf{x} \prec^w \mathbf{y}$; in other words, majorization is a more restrictive definition than weakly majorization.

Observe that the original order of the elements of $\mathbf{x}$ and $\mathbf{y}$ plays no role in the definition of majorization. In other words,

$$\mathbf{x} \prec \mathbf{\Pi}\mathbf{x} \tag{2.9}$$

for all permutation matrices $\mathbf{\Pi}$.

[3] A square matrix $\mathbf{P}$ is said to be *stochastic* if either its rows or columns are probability vectors, i.e., if its elements are all nonnegative and either the rows or the columns sums are one. If both the rows and columns are probability vectors, then the matrix is called *doubly stochastic*. Stochastic matrices can be considered representations of the transition probabilities of a finite Markov chain.

[4] The permutation matrices are doubly stochastic and, in fact, the convex hull of the permutation matrices coincides with the set of doubly stochastic matrices.

[5] More specifically, $\mathbf{x}$ is said to be weakly supermajorized by $\mathbf{y}$ (as opposed to the submajorization relation denoted by $\mathbf{x} \prec_w \mathbf{y}$ [97, 1.A.2]).

Functions that preserve the ordering of majorization are said to be Schur-convex as defined next.

Definition 2.4. [97, 3.A.1] A real-valued function ϕ defined on a set $\mathcal{A} \subseteq \mathbb{R}^n$ is said to be Schur-convex on $\mathcal{A}$ if

$$\mathbf{x} \prec \mathbf{y} \quad \text{on } \mathcal{A} \Rightarrow \phi(\mathbf{x}) \leq \phi(\mathbf{y}). \tag{2.10}$$

If, in addition, $\phi(\mathbf{x}) < \phi(\mathbf{y})$ whenever $\mathbf{x} \prec \mathbf{y}$ but $\mathbf{x}$ is not a permutation of $\mathbf{y}$, then ϕ is said to be strictly Schur-convex on $\mathcal{A}$.

Similarly, ϕ is said to be Schur-concave on $\mathcal{A}$ if

$$\mathbf{x} \prec \mathbf{y} \quad \text{on } \mathcal{A} \Rightarrow \phi(\mathbf{x}) \geq \phi(\mathbf{y}) \tag{2.11}$$

and ϕ is strictly Schur-concave on A if strict inequality $\phi(\mathbf{x}) > \phi(\mathbf{y})$ holds when $\mathbf{x}$ is not a permutation of $\mathbf{y}$.

Of course, if ϕ is Schur-convex on $\mathcal{A}$ then $-\phi$ is Schur-concave on $\mathcal{A}$ and vice versa.

It is important to remark that the sets of Schur-convex and Schur-concave functions do not form a partition of the set of all functions from $\mathcal{A} \subseteq \mathbb{R}^n$ to $\mathbb{R}$. In fact, neither are the two sets disjoint (the intersection is not empty), unless we consider strictly Schur-convex/concave functions, nor do they cover the entire set of all functions as illustrated in Figure 2.1.

We now give some illustrative examples.

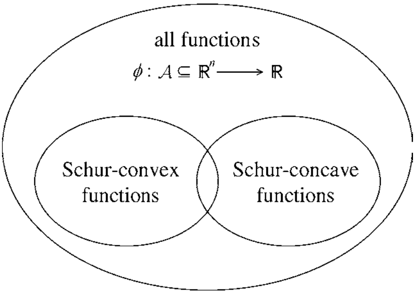

all functions
$\phi : \mathcal{A} \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}$
Schur-convex functions
Schur-concave functions

Fig. 2.1 Illustration of the sets of Schur-convex and Schur-concave functions within the set of all functions $\phi : \mathcal{A} \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}$.

Example 2.1. The function $\phi(\mathbf{x}) = \sum_{i=1}^{n} x_i$ is both Schur-convex and Schur-concave since $\phi(\mathbf{x}) = \phi(\mathbf{y})$ for any $\mathbf{x} \prec \mathbf{y}$. However, it is neither strictly Schur-convex nor strictly Schur-concave.

Example 2.2. The function $\phi(\mathbf{x}) = c$ is trivially both Schur-convex and Schur-concave, but not strictly.

Example 2.3. The function $\phi(\mathbf{x}) = x_1 + 2x_2 + x_3$ is neither Schur-convex nor Schur-concave as can be seen from the counterexample given by $\mathbf{x} = [2,1,1]^T$, $\mathbf{y} = [2,2,0]^T$, and $\mathbf{z} = [4,0,0]^T$, from which $\mathbf{x} \prec \mathbf{y} \prec \mathbf{z}$ but $\phi(\mathbf{x}) < \phi(\mathbf{y}) > \phi(\mathbf{z})$.

The following definition will be instrumental in the derivation of transformations that relate vectors that satisfy the majorization relation.

Definition 2.5. [97, p. 21] A *T-transform* is a matrix of the form:

$$\mathbf{T} = \alpha \mathbf{I} + (1-\alpha)\mathbf{\Pi} \tag{2.12}$$

for some $\alpha \in [0,1]$ and some $n \times n$ permutation matrix $\mathbf{\Pi}$ with $n-2$ diagonal entries equal to 1. Let $[\mathbf{\Pi}]_{ij} = [\mathbf{\Pi}]_{ji} = 1$ for some indices $i < j$, then

$$\mathbf{\Pi y} = [y_1, \ldots, y_{i-1}, y_j, y_{i+1}, \ldots, y_{j-1}, y_i, y_{j+1}, \ldots, y_n]^T$$

and hence

$$\begin{aligned}\mathbf{T y} = [&y_1, \ldots, y_{i-1}, \alpha y_i + (1-\alpha) y_j, y_{i+1}, \ldots, \\ &y_{j-1}, \alpha y_j + (1-\alpha) y_i, y_{j+1}, \ldots, y_n]^T.\end{aligned}$$

2.2 Basic Results

We start with some fundamental majorization relations.

Lemma 2.1. [97, p. 7] For any vector $\mathbf{x} \in \mathbb{R}^n$, let $\mathbf{1} \in \mathbb{R}^n$ denote the vector with equal elements given by $1_i \triangleq \sum_{j=1}^{n} x_j/n$. Then

$$\mathbf{1} \prec \mathbf{x}. \tag{2.13}$$

Lemma 2.1 is simply stating the obvious fact that a vector of equal components has the "least spread out" or the "most equal"components among all vectors. The following is a fundamental result on the application of majorization theory to matrices.

Lemma 2.2. [97, 9.B.1] Let $\mathbf{R}$ be a Hermitian matrix with diagonal elements denoted by the vector $\mathbf{d}$ and eigenvalues denoted by the vector $\boldsymbol{\lambda}$. Then

$$\boldsymbol{\lambda} \succ \mathbf{d}. \tag{2.14}$$

Lemmas 2.1 and 2.2 can be put together to nicely "bound" the diagonal elements of a matrix as

$$\mathbf{1} \prec \mathbf{d} \prec \boldsymbol{\lambda}, \tag{2.15}$$

or, in other words, as stated in the following result.

Corollary 2.1. Let $\mathbf{R}$ be a Hermitian matrix and $\mathbf{Q}$ a unitary matrix. Then,

$$\mathbf{1}(\mathbf{R}) \prec \mathbf{d}\big(\mathbf{Q}^\dagger \mathbf{R} \mathbf{Q}\big) \prec \boldsymbol{\lambda}(\mathbf{R}), \tag{2.16}$$

where $\mathbf{1}(\mathbf{A})$ denotes the vector with equal elements whose sum equal to the trace of $\mathbf{A}$, $\mathbf{d}(\mathbf{A})$ is the vector with the diagonal elements of $\mathbf{A}$, and $\boldsymbol{\lambda}(\mathbf{A})$ is the vector with the eigenvalues of $\mathbf{A}$.

Proof. It follows directly from Lemmas 2.1 and 2.2, and by noticing that $\mathbf{1}\left(\mathbf{Q}^\dagger \mathbf{R} \mathbf{Q}\right) = \mathbf{1}(\mathbf{R})$ and $\boldsymbol{\lambda}\left(\mathbf{Q}^\dagger \mathbf{R} \mathbf{Q}\right) = \boldsymbol{\lambda}(\mathbf{R})$. □

Corollary 2.1 “bounds”the diagonal elements of $\mathbf{Q}^{\dagger}\mathbf{R}\mathbf{Q}$ for any unitary matrix $\mathbf{Q}$; however, it does not specify what can be achieved. The following result will be instrumental for that purpose.

Lemma 2.3. [97, 9.B.2] For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ satisfying $\mathbf{x} \prec \mathbf{y}$, there exists a real symmetric (and therefore Hermitian) matrix with diagonal elements given by $\mathbf{x}$ and eigenvalues given by $\mathbf{y}$.

Corollary 2.2. For any vector $\boldsymbol{\lambda} \in \mathbb{R}^n$, there exists a real symmetric (and therefore Hermitian) matrix with equal diagonal elements and eigenvalues given by $\boldsymbol{\lambda}$.

Proof. The proof is straightforward from Lemmas 2.1 and 2.3. □

A restatement of Lemma 2.3 more convenient for our purpose is given next.

Corollary 2.3. Let $\mathbf{R}$ be a Hermitian matrix and $\mathbf{x} \in \mathbb{R}^n$ be a vector satisfying $\mathbf{x} \prec \boldsymbol{\lambda}(\mathbf{R})$. Then, there exists a unitary matrix $\mathbf{Q}$ such that

$$\mathbf{d}\left(\mathbf{Q}^{\dagger}\mathbf{R}\mathbf{Q}\right) = \mathbf{x}. \tag{2.17}$$

Lemma 2.3 is the converse of Lemma 2.2 (in fact it is stronger than the converse since it guarantees the existence of a real symmetric matrix instead of just a Hermitian matrix). Now, we can complete the characterization of Corollary 2.1.

Corollary 2.4. Let $\mathbf{R}$ be a Hermitian matrix. There exists a unitary matrix $\mathbf{Q}$ such that

$$\mathbf{d}\left(\mathbf{Q}^{\dagger}\mathbf{R}\mathbf{Q}\right) = \mathbf{1}(\mathbf{R}) \tag{2.18}$$

and also a unitary matrix $\mathbf{Q}$ such that

$$\mathbf{d}\left(\mathbf{Q}^{\dagger}\mathbf{R}\mathbf{Q}\right) = \boldsymbol{\lambda}(\mathbf{R}). \tag{2.19}$$

The next result relates the majorization and weak majorization relations in a useful way.

Lemma 2.4. [97, 5.A.9.a] For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ satisfying $\mathbf{y} \succ^w \mathbf{x}$, there exists a vector $\mathbf{u}$ such that

$$\mathbf{u} \leq \mathbf{x} \quad \text{and} \quad \mathbf{y} \succ \mathbf{u}. \tag{2.20}$$

The following results may be useful in determining whether a function is Schur-convex or Schur-concave (in addition, to using directly Definition 2.4).

Lemma 2.5. [97, 3.B.1] An increasing function of a Schur-convex (Schur-concave) function is Schur-convex (Schur-concave). Similarly, a decreasing function of a Schur-convex (Schur-concave) function is Schur-concave (Schur-convex).

Lemma 2.6. [97, 3.B.2] The composite function $f_0(g(x_1), \ldots, g(x_n))$ is Schur-convex if $f_0 : \mathbb{R}^n \to \mathbb{R}$ is Schur-convex and increasing in each argument and $g : \mathbb{R} \to \mathbb{R}$ is convex.

Corollary 2.5. Let $\phi(\mathbf{x}) = \sum_i g(x_i)$ where g is convex. Then ϕ is Schur-convex.

Lemma 2.7. [97, 3.H.2] Let $\phi(\mathbf{x}) = \sum_i g_i(x_i)$, where $x_i \geq x_{i+1}$ and each g_i is differentiable. Then ϕ is Schur-convex if and only if

$$g_i'(a) \geq g_{i+1}'(b) \quad \text{whenever} \quad a \geq b, \quad i = 1, \ldots, n-1.$$

Lemma 2.8. [97, 3.A.3] Let $\phi : \mathcal{D}_n \to \mathbb{R}$ be a real-valued function continuous on $\mathcal{D}_n \triangleq \{\mathbf{x} \in \mathbb{R}^n : x_1 \geq \cdots \geq x_n\}$ and continuously differentiable on the interior of $\mathcal{D}_n$. Then ϕ is Schur-convex (Schur-concave) on $\mathcal{D}_n$ if and only if $\frac{\partial \phi(\mathbf{x})}{\partial x_i}$ is decreasing (increasing) in $i = 1, \ldots, n$.

We now turn to the important algorithmic aspect of majorization theory which is necessary, for example, to compute a matrix with given diagonal elements and eigenvalues.

Lemma 2.9. [97, 2.B.1] For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ satisfying $\mathbf{x} \prec \mathbf{y}$, there exists a sequence of T-transforms $\mathbf{T}^{(1)}, \ldots, \mathbf{T}^{(K)}$ such that $\mathbf{x} = \mathbf{T}^{(K)} \cdots \mathbf{T}^{(1)} \mathbf{y}$ and $K < n$.

We now give an algorithm to obtain such a sequence of T-transforms from [97, 2.B.1].

Algorithm 2.1. [97, 2.B.1] Algorithm to obtain a sequence of T-transforms such that $\mathbf{x} = \mathbf{T}^{(K)} \cdots \mathbf{T}^{(1)} \mathbf{y}$.

Input: Vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ satisfying $\mathbf{x} \prec \mathbf{y}$ (it is assumed that the components of $\mathbf{x}$ and $\mathbf{y}$ are in decreasing order and that $\mathbf{x} \neq \mathbf{y}$).

Output: Set of T-transforms $\mathbf{T}^{(1)}, \ldots, \mathbf{T}^{(K)}$.

```
0. Let $\mathbf{y}^{(0)} = \mathbf{y}$ and $k = 1$ be the iteration index.
1. Find the largest index $i$ such that $y_i^{(k-1)} > x_i$ and the smallest index $j$ greater than $i$ such that $y_j^{(k-1)} < x_j$.
2. Let $\delta = \min\left(x_j - y_j^{(k-1)}, y_i^{(k-1)} - x_i\right)$ and $\alpha = 1 - \delta / \left(y_i^{(k-1)} - y_j^{(k-1)}\right)$.
3. Use $\alpha$ to compute $\mathbf{T}^{(k)}$ as in (2.12) and let $\mathbf{y}^{(k)} = \mathbf{T}^{(k)} \mathbf{y}^{(k-1)}$.
4. If $\mathbf{y}^{(k)} \neq \mathbf{x}$, then set $k = k + 1$ and go to step 1; otherwise, finish.
```

A recursive algorithm to obtain a matrix with a given vector of eigenvalues and vector of diagonal elements is indicated in [97, 9.B.2] and [161, Sec. IV-A]. We consider the practical and simple method obtained in [161, Sec. IV-A] and reproduce it here for completeness.

Algorithm 2.2. [161, Sec. IV-A] Algorithm to obtain a real symmetric matrix $\mathbf{R}$ with diagonal values given by $\mathbf{x}$ and eigenvalues given by $\mathbf{y}$ provided that $\mathbf{x} \prec \mathbf{y}$.

Input: Vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ satisfying $\mathbf{x} \prec \mathbf{y}$ (it is assumed that the components of $\mathbf{x}$ and $\mathbf{y}$ are in decreasing order and that $\mathbf{x} \neq \mathbf{y}$).

Output: Matrix $\mathbf{R}$.

1. Using Algorithm 2.1, obtain a sequence of T-transforms such that $\mathbf{x} = \mathbf{T}^{(K)} \cdots \mathbf{T}^{(1)} \mathbf{y}$.
2. Define the Givens rotation $\mathbf{U}^{(k)}$ as
$$\left[\mathbf{U}^{(k)}\right]_{ij} = \begin{cases} \sqrt{\left[\mathbf{T}^{(k)}\right]_{ij}} & \text{for} \quad i < j \\ -\sqrt{\left[\mathbf{T}^{(k)}\right]_{ij}} & \text{otherwise.} \end{cases}$$
3. Let $\mathbf{R}^{(0)} = \operatorname{diag}(\mathbf{y})$ and $\mathbf{R}^{(k)} = \mathbf{U}^{(k)T} \mathbf{R}^{(k-1)} \mathbf{U}^{(k)}$. The desired matrix is given by $\mathbf{R} = \mathbf{R}^{(K)}$. Alternatively, define the unitary matrix $\mathbf{Q} = \mathbf{U}^{(1)} \cdots \mathbf{U}^{(K)}$ and the desired matrix is given by $\mathbf{R} = \mathbf{Q}^T \operatorname{diag}(\mathbf{y})\, \mathbf{Q}$.

Algorithm 2.2 obtains a real symmetric matrix $\mathbf{R}$ with given eigenvalues and diagonal elements. For the interesting case in which the diagonal elements must be equal and allowing the desired matrix to be complex, it is possible to obtain an alternative much simpler solution in closed form as given next.

Lemma 2.10. Let $\mathbf{Q}$ a unitary matrix satisfying the condition $|[\mathbf{Q}]_{ik}| = |[\mathbf{Q}]_{il}|\ \forall i, k, l$. Then, the matrix $\mathbf{R} = \mathbf{Q}^{\dagger} \operatorname{diag}(\boldsymbol{\lambda})\, \mathbf{Q}$ has equal diagonal elements (and eigenvalues given by $\boldsymbol{\lambda}$). Two examples of $\mathbf{Q}$ are the unitary Discrete Fourier Transform (DFT) matrix and the Hadamard matrix (when the dimensions are appropriate such as a power of two [116, Sec. 5.6] [157, p. 66]) .

Nevertheless, Algorithm 2.2 has the nice property that the obtained matrix $\mathbf{Q}$ is real-valued and can be naturally decomposed (by construction) as the product of Givens rotations (where each term performs a single rotation [143]). This simple structure plays a key role for practical implementation. Interestingly, an iterative approach to construct a matrix with equal diagonal elements and with a given set of eigenvalues was obtained in [100], based also on a sequence of rotations.

2.3 Multiplicative Majorization

Parallel to the concept of *additive* majorization is the notion of *multiplicative* majorization (also termed log-majorization).

Definition 2.6. The vector $\mathbf{x} \in \mathbb{R}_+^n$ is multiplicatively majorized by $\mathbf{y} \in \mathbb{R}_+^n$, denoted by $\mathbf{x} \prec_\times \mathbf{y}$, if

$$\begin{aligned} \prod_{i=1}^{k} x_{[i]} &\leq \prod_{i=1}^{k} y_{[i]} \quad 1 \leq k < n \\ \prod_{i=1}^{n} x_{[i]} &= \prod_{i=1}^{n} y_{[i]}. \end{aligned} \tag{2.21}$$

To differentiate the two types of majorization, we sometimes use the symbol $\prec_+$ rather than $\prec$ to denote (additive) majorization.

Similar to the definition of Schur convex/concave function, it is natural to define a multiplicatively Schur-convex/concave function.

Definition 2.7. A function $\phi : \mathcal{A} \to \mathbb{R}$ is said to be multiplicatively Schur-convex on $\mathcal{A} \in \mathbb{R}^n$ if

$$\mathbf{x} \prec_\times \mathbf{y} \quad \text{on } \mathcal{A} \Rightarrow \phi(\mathbf{x}) \leq \phi(\mathbf{y}) \tag{2.22}$$

and multiplicatively Schur-concave on $\mathcal{A}$ if

$$\mathbf{x} \prec_\times \mathbf{y} \quad \text{on } \mathcal{A} \Rightarrow \phi(\mathbf{x}) \geq \phi(\mathbf{y}). \tag{2.23}$$

However, it is not necessary to use the notion of multiplicatively Schur-convex/concave functions since $\mathbf{x} \prec_+ \mathbf{y}$ if and only if $\exp(\mathbf{x}) \prec_\times \exp(\mathbf{y})$.[6] Hence, the so-called multiplicatively Schur-convex/concave functions defined in (2.22) and (2.23) can be equivalently referred to as functions such that $\phi \circ \exp$ is Schur-convex and Schur-concave, respectively, where the composite function is defined as $\phi \circ \exp(\mathbf{x}) \triangleq \phi(e^{x_1}, \ldots, e^{x_n})$.

The following two lemmas relate the Schur-convexity/concavity of a function f with that of the composite function $f \circ \exp$.

[6] Indeed, using the language of group theory, we say that the groups $(\mathbb{R}, +)$ and $(\mathbb{R}_+, \times)$ are isomorphic since there is a bijection function $\exp : \mathbb{R} \to \mathbb{R}_+$ such that $\exp(x + y) = \exp(x) \times \exp(y)$ for $\forall x, y \in \mathbb{R}$.

Lemma 2.11. If f is Schur-convex and increasing in each argument, then $f \circ \exp$ is Schur-convex.

Proof. This lemma is an immediate corollary of Lemma 2.6 since the function $\exp(x)$ is convex. □

Lemma 2.12. For a composite function $f \circ \exp$ which is Schur-concave on $\mathcal{D}_n \triangleq \{\mathbf{x} \in \mathbb{R}^n : x_1 \geq \cdots \geq x_n\}$, f is Schur-concave on $\mathcal{D}_n$ if it is increasing in each argument.

Proof. To prove a function is Schur-convex/concave, without loss of generality one only needs to check the two-argument case [97, 3.A.5]. Denote $g(x_1,x_2) = f(e^{x_1},e^{x_2})$. With $g(x_1,x_2)$ being Schur-concave on $\mathcal{D}_2$, it follows from Lemma 2.8 that

$$\frac{\partial g(x_1,x_2)}{\partial x_1} \leq \frac{\partial g(x_1,x_2)}{\partial x_2}, \tag{2.24}$$

or equivalently

$$\left.\frac{\partial f(y_1,e^{x_2})}{\partial y_1}\right|_{y_1=e^{x_1}} e^{x_1} \leq \left.\frac{\partial f(e^{x_1},y_2)}{\partial y_2}\right|_{y_2=e^{x_2}} e^{x_2}. \tag{2.25}$$

Because f is increasing in each argument, the two derivatives at the both sides of (2.25) are positive. Moreover $e^{x_1} \geq e^{x_2} > 0$ in $\mathbf{x} \in \mathcal{D}_2$. It follows from (2.25) that

$$\frac{\partial f(x_1,x_2)}{\partial x_1} \leq \frac{\partial f(x_1,x_2)}{\partial x_2}, \tag{2.26}$$

which means that f is Schur-concave on $\mathcal{D}_2$. □

The following two examples show that the implication in Lemmas 2.11 and 2.12 does not hold in the opposite direction.

Example 2.4. The function $f(\mathbf{x}) = \prod_{i=1}^{n} x_i$ is Schur-concave on $\mathcal{D}_n$ since $\frac{\partial f(\mathbf{x})}{\partial x_i} = \frac{f(\mathbf{x})}{x_i}$ is increasing in i on $\mathcal{D}_n$ (see Lemma 2.8). However, the composite function $f \circ \exp(\mathbf{x}) = \exp(\sum_i x_i)$ is Schur-convex (and Schur-concave as well).

Example 2.5. The function $f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i x_i$, where $\alpha_1 \leq \cdots \leq \alpha_n$, is Schur-concave on $\mathcal{D}_n$. The composite function is $f \circ \exp(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \exp(x_i)$. For $\alpha_i \leq \alpha_{i+1}$, the derivative $\frac{\partial f \circ \exp(\mathbf{x})}{\partial x_i} = \alpha_i \exp(x_i)$ is *not* always monotonous in $i = 1, \ldots, n$ for any $\mathbf{x} \in \mathcal{D}_n$. Hence according to Lemma 2.8, although f is Schur-concave, $f \circ \exp$ is not Schur-concave on $\mathcal{D}_n$ (neither Schur-convex).

3

Linear MIMO Transceivers

This chapter deals with the design of point-to-point linear MIMO transceivers (commonly referred to as linear precoder at the transmitter and linear equalizer at the receiver) with perfect CSI at both sides of the link from a signal processing perspective. The design of MIMO transceivers has been studied for three decades (the first papers dating back to the 1970s), initially motivated by cable systems and more recently by multi-antenna wireless systems. A MIMO transceiver can be designed according to different design criteria that will tradeoff differently the performance of each of the substreams into the global performance of the system. The choice of the appropriate measure of the system performance typically depends on the application at hand and on the mathematical tractability of the resulting problem. As a consequence, different criteria have been pursued in the literature.

Most designs are based on tractable cost functions such as the sum of the MSEs of all substreams or, equivalently, the trace of the MSE matrix [2, 90, 129, 171]. Other considered criteria include the maximization of the signal to interference-plus-noise ratio (SINR) [131], the minimization of the weighted sum of MSEs [130], and the minimization of the determinant of the MSE matrix [170].

These problems can be optimally solved even though they are very complicated. The reason is that the channel matrix turns out to be diagonalized by the optimal transmit–receive processing and the transmission is effectively performed on a diagonal or parallel fashion. Indeed, the diagonal transmission implies a *scalarization* of the problem (meaning that all matrix equations are substituted with scalar ones) with the consequent simplification.

If more elaborated measures of performance are considered, e.g., based on the bit error rate (BER) rather than on the MSE or SINR, the problem becomes even more complicated due to the nonconvexity of the cost function and the difficulty of manipulating the matrix-valued variables. In an attempt to simplify these problems and in light of the optimality of the diagonal structure for transmission in all the previous examples (including the capacity-achieving solution [33, 122, 148]), one might expect that the same would hold for any other criteria as well. However, as shown in [111], this is not the case.

In [106], the minimization of the BER (and also of the Chernoff upper bound) averaged over the channel substreams was treated in detail when a diagonal structure is imposed. The minimum BER design of a linear MIMO transceiver without the diagonal structure constraint was independently obtained in [36] and [111], resulting in an optimal nondiagonal structure. This result, however, only applies when the constellations used in all the substreams are equal. The general case of different constellations,[1] however, is much more involved and was solved in [110] via a primal decomposition approach.

In [111], a general unifying framework was developed that embraces a wide range of different design criteria for linear MIMO transceivers; in particular, the optimal design was obtained for the family of Schur-concave and Schur-convex cost functions which arise in majorization theory [97]. Interestingly, this framework gives a clear answer to the question of whether the diagonal transmission is optimal: when the cost function is Schur-concave then the diagonal structure is optimal, but when the cost function is Schur-convex then the optimal structure is not

[1] Different constellations are typically obtained when some kind of bit allocation strategy is used such as the gap-approximation method [28, Part II] which chooses the constellations as a function of the channel realization.

diagonal anymore. For arbitrary cost functions, not necessarily Schur-concave/convex, a more general framework still based on majorization theory can be developed.

An alternative way to formulate the design of MIMO transceivers is to consider an independent requirement of quality for each of the substreams rather than a global measure of quality [114]. This may be a useful formulation, for example, if different services are being supported each with a different Quality-of-Service (QoS) requirement. In fact, this approach provides a complete characterization of the problem as a multi-objective optimization and, in particular, of the Pareto-optimal points.[2]

This chapter fully explores the framework based on majorization theory for the optimal design of linear MIMO transceivers. Using this framework, the original complicated nonconvex problem with matrix-valued variables is reformulated as a simple convex problem with scalar-valued variables. Then, the simplified problem can be addressed under the powerful framework of convex optimization theory to obtain closed-form solutions or numerical algorithms with worst-case polynomial convergence. In particular, three problem formulations are considered in increasing order of complexity:

(1) Design based on a Schur-concave/convex cost function as a global measure of performance.
(2) Design based on individual QoS constraints.
(3) Design based on an arbitrary cost function as a global measure of performance.

After introducing the signal model in Section 3.1 and formulating the problem designs in Section 3.2, Section 3.3 obtains the optimum receiver and then Sections 3.4–3.6 derive the optimum transmitter according to the three different problem formulations listed above. Section 3.7 briefly explores the extension to multicarrier systems. Finally, Section 3.8 summarizes the chapter.

[2] A Pareto optimal solution is an optimal solution to a multi-objective optimization problem; it is defined as any solution that cannot be improved with respect to any component without worsening the others [20].

3.1 System Model

The baseband signal model corresponding to a transmission through a general MIMO communication channel with n_T transmit and n_R receive dimensions is

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}, \tag{3.1}$$

where $\mathbf{s} \in \mathbb{C}^{n_T \times 1}$ is the transmitted vector, $\mathbf{H} \in \mathbb{C}^{n_R \times n_T}$ is the channel matrix, $\mathbf{y} \in \mathbb{C}^{n_R \times 1}$ is the received vector, and $\mathbf{n} \in \mathbb{C}^{n_R \times 1}$ is a zero-mean circularly symmetric complex Gaussian interference-plus-noise vector with arbitrary covariance matrix $\mathbf{R}_n$. For the sake of notation and without loss of generality, we will assume the noise to be white, i.e., $\mathbf{R}_n = \mathbf{I}$.[3]

We consider the use of linear transceivers, composed of a linear precoder at the transmitter and a linear equalizer at the receiver (the case of nonlinear decision-feedback (DF) transceivers is fully covered in Chapter 4). The transmitted vector can be written as (see Figure 3.1)

$$\mathbf{s} = \mathbf{P}\mathbf{x}, \tag{3.2}$$

where $\mathbf{P} \in \mathbb{C}^{n_T \times L}$ is the transmit matrix or linear $\mathbf{P}$ recoder, and $\mathbf{x} \in \mathbb{C}^{L \times 1}$ is the data vector that contains the L symbols to be transmitted (zero-mean,[4] normalized, and uncorrelated such that $\mathbb{E}\left[\mathbf{x}\mathbf{x}^\dagger\right] = \mathbf{I}$) drawn from a set of constellations. For the sake of notation, it is

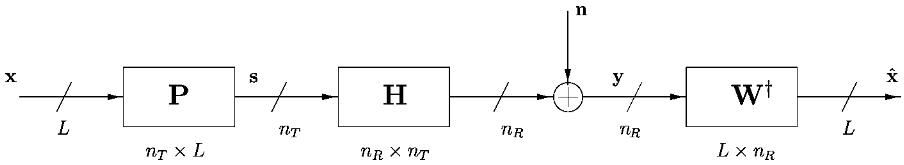


Fig. 3.1 Scheme of a general MIMO communication system with a linear transceiver.

[3] We can always assume that the first stage at the receiver is a whitening stage, producing the equivalent signal model $\bar{\mathbf{y}} = \mathbf{R}_n^{-1/2}\mathbf{y} = \bar{\mathbf{H}}\mathbf{s} + \bar{\mathbf{n}}$, where $\bar{\mathbf{H}} = \mathbf{R}_n^{-1/2}\mathbf{H}$ is the whitened channel and $\bar{\mathbf{n}}$ is a white noise. Observe that any additional linear processing at the receiver $\bar{\mathbf{W}}^\dagger\bar{\mathbf{y}}$ will be equivalent to a linear processing on the original received signal $\mathbf{W}^\dagger\mathbf{y}$ with $\mathbf{W} = \mathbf{R}_n^{-1/2}\bar{\mathbf{W}}$ (in particular, we have $\bar{\mathbf{W}}^\dagger\bar{\mathbf{W}} = \mathbf{W}^\dagger\mathbf{R}_n\mathbf{W}$ and $\bar{\mathbf{W}}^\dagger\bar{\mathbf{H}} = \mathbf{W}^\dagger\mathbf{H}$).

[4] The mean of the signal does not carry any information and can always be set to zero saving power at the transmitter.

assumed in this chapter that $L \leq \min(n_R, n_T)$. The total average transmitted power (in units of energy per transmission) is

$$P_T = \mathbb{E}[\,\|\mathbf{s}\|^2] = \operatorname{Tr}\left(\mathbf{P}\mathbf{P}^\dagger\right). \tag{3.3}$$

Similarly, the estimated data vector at the receiver is (see Figure 3.1)

$$\hat{\mathbf{x}} = \mathbf{W}^\dagger \mathbf{y}, \tag{3.4}$$

where $\mathbf{W}^\dagger \in \mathbb{C}^{L \times n_R}$ is the receive matrix or linear equalizer (as will be seen later, the optimal receiver is the **W**iener filter).

It is interesting to observe that the ith column of $\mathbf{P}$ and $\mathbf{W}$, $\mathbf{p}_i$ and $\mathbf{w}_i$, respectively, can be interpreted as the transmit and receive beamvectors associated to the ith transmitted symbol x_i:

$$\hat{x}_i = \mathbf{w}_i^\dagger \left(\mathbf{H}\mathbf{p}_i x_i + \mathbf{n}_i\right), \tag{3.5}$$

where $\mathbf{n}_i = \sum_{j \neq i} \mathbf{H}\mathbf{p}_j x_j + \mathbf{n}$ is the equivalent noise seen by the ith substream, with covariance matrix $\mathbf{R}_{n_i} = \sum_{j \neq i} \mathbf{H}\mathbf{p}_j \mathbf{p}_j^\dagger \mathbf{H}^\dagger + \mathbf{I}$. Therefore, the linear MIMO transceiver scheme (see Figure 3.1) can be equivalently interpreted as a multiple beamforming transmission (see Figure 1.5).

The previously introduced complex-valued signal model could have been similarly written with an augmented real-valued notation, simply by augmenting the n-dimensional complex vectors to $2n$-dimensional real vectors (stacking the real and imaginary parts). However, the use of a complex-valued notation is always preferred since it models the system in a simpler and more compact way. Interestingly, it turns out that complex linear filtering is equivalent to (augmented) real linear filtering if the random vectors involved are proper or circular [102, 117]; otherwise, complex linear filtering is suboptimal and it is necessary to consider either real linear filtering or widely complex linear filtering [117, 118]. Fortunately, many of the commonly employed constellations, such as the family of QAM constellations, are proper [134], which allows the use of a nice complex notation (although some other constellations, such as BPSK and GMSK, are improper and a complex notation is not adequate anymore).

3.1.1 Measures of Performance: MSE, SINR, and BER

The performance of the ith established substream or link in (3.5) can be conveniently measured, among others, in terms of the MSE, SINR, or BER, defined, respectively, as

$$\mathrm{MSE}_i \triangleq \mathbb{E}[|\hat{x}_i - x_i|^2] = |\mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_i - 1|^2 + \mathbf{w}_i^\dagger \mathbf{R}_{n_i} \mathbf{w}_i, \tag{3.6}$$

$$\mathrm{SINR}_i \triangleq \frac{\text{desired component}}{\text{undesired component}} = \frac{|\mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_i|^2}{\mathbf{w}_i^\dagger \mathbf{R}_{n_i} \mathbf{w}_i}, \tag{3.7}$$

$$\mathrm{BER}_i \triangleq \frac{\#\text{ bits in error}}{\#\text{ transmitted bits}} \approx \tilde{g}_i\left(\mathrm{SINR}_i\right), \tag{3.8}$$

where $\tilde{g}_i$ is a decreasing function that relates the BER to the SINR at the ith substream. Any properly designed system should attempt to minimize the MSEs, maximize the SINRs, or minimize the BERs, as is mathematically formulated in the next section.

It will be notationally convenient to define the following MSE matrix as the covariance matrix of the error vector between the transmitted signal and the estimated one ($\mathbf{e} \triangleq \hat{\mathbf{x}} - \mathbf{x}$):

$$\begin{aligned}\mathbf{E} &\triangleq \mathbb{E}[(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^\dagger] \\ &= \left(\mathbf{W}^\dagger \mathbf{H}\mathbf{P} - \mathbf{I}\right)\left(\mathbf{P}^\dagger \mathbf{H}^\dagger \mathbf{W} - \mathbf{I}\right) + \mathbf{W}^\dagger \mathbf{W}\end{aligned} \tag{3.9}$$

from which the MSE of the ith link is obtained as the ith diagonal element of $\mathbf{E}$, i.e.,

$$\mathrm{MSE}_i = [\mathbf{E}]_{ii}. \tag{3.10}$$

Regarding (3.8), the BER can indeed be analytically expressed for most types of modulations as a function of the SINR when the interference-plus-noise term follows a Gaussian distribution [8, 157] (see, for example, [24] for exact BER expressions of amplitude modulations).[5] For the case of a zero forcing (ZF) receiver, each of the established links contains only Gaussian noise and, therefore, the analytical BER characterization can be exact. For the case of an MMSE

[5] Note that the BER function is valid for the MMSE receiver only when the decision regions of the detector are scaled to account for the bias in the MMSE receiver [28, Part I].

receiver, however, there is crosstalk among the established links with a non-Gaussian distribution. The computation of the BER involves then a summation over all the possible values of the interfering signals which is exponential in the size of the constellations. In order to reduce the complexity of evaluating these expressions, it is customary to obtain an approximate statistical model for the crosstalk. In fact, the central limit theorem can be invoked to show that the distribution converges almost surely to a Gaussian distribution as the number of interfering signals increases [119] (even when the central limit theorem cannot be invoked, it is in general possible to obtain some approximate expression for the BER as a function of the SINR [28, Part I, Sec. III.B]).

It is worth pointing out that expressing the BER as in (3.8) implicitly assumes that the different links are independently detected after the linear receiver $\mathbf{W}$ that processes jointly all the substreams (see Figure 3.2). This reduces the complexity drastically compared to a joint ML detection and is indeed the main advantage of using the receive matrix $\mathbf{W}$ (see Section 5.3 for existing results on transceiver design with an ML receiver).

An Important Example: QAM Constellations

To illustrate the previous characterization of the BER, we now consider a particular example of great interest: QAM constellations. As previously mentioned, for QAM constellations under Gaussian noise, the BER can be analytically expressed in an exact way [24]. However, for simplicity of exposition, we approximate the expression with the most significative term.

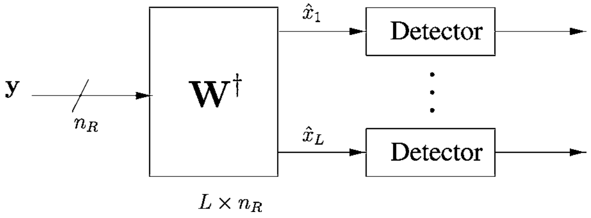


Fig. 3.2 Independent detection of the substreams after the joint linear processing with receive matrix $\mathbf{W}$.

The BER corresponding to an M-ary QAM constellation (assuming that a Gray encoding is used to map the bits into the constellation points) is [8, 24]

$$\mathrm{BER}\left(\mathrm{SINR}\right) \approx \frac{\alpha}{\log_2 M}\mathcal{Q}\left(\sqrt{\beta\,\mathrm{SINR}}\right), \tag{3.11}$$

where $\mathcal{Q}$ is defined as $\mathcal{Q}(x) \triangleq \left(1/\sqrt{2\pi}\right)\int_x^\infty e^{-\lambda^2/2}d\lambda$ [157],[6] $\alpha = 4\left(1 - 1/\sqrt{M}\right)$, and $\beta = 3/\left(M-1\right)$ are parameters that depend on the constellation size.[7] It is sometimes convenient to use the Chernoff upper bound of the tail of the Gaussian distribution function $\mathcal{Q}(x) \leq (1/2)\,e^{-x^2/2}$ [157] to approximate the BER (which becomes a reasonable approximation for high values of the SINR) as

$$\mathrm{BER}\left(\mathrm{SINR}\right) \approx \frac{\alpha}{2\log_2 M}\,e^{-\beta/2\,\mathrm{SINR}}. \tag{3.12}$$

(See [27] for better approximations for M-QAM and M-PSK constellations based on curve fitting and [123] for approximations in the neighborhood of some nominal point.)

3.2 Problem Formulation

This section formulates the design of the linear MIMO transceiver (matrices $\mathbf{P}$ and $\mathbf{W}$) as a tradeoff between the power transmitted and the performance achieved. First, using a global measure of performance either in the objective or in the constraints and, then, using individual constraints on the performance of each substream.

3.2.1 Global Measure of Performance

Measuring the global performance of a MIMO system with several substreams is tricky as there is an inherent tradeoff among the performance of the different substreams. Different applications may require a different balance on the performance of the substreams (although one may also argue that minimizing the BER averaged over the substreams is

[6] The $\mathcal{Q}$-function and the commonly used complementary error function "erfc" are related as $\mathrm{erfc}\,(x) = 2\,\mathcal{Q}(\sqrt{2}x)$ [157].

[7] For $I \times J$ rectangular constellations, the parameters are $\alpha = 2\left((I-1)/I + (J-1)/J\right)$ and $\beta = 6/\left(I^2 + J^2 - 2\right)$ [24].

the best criterion). The most common criterion in the literature is the minimization of the sum of the MSEs [2, 90, 129, 130, 131, 171]. Other considered criteria include the maximization of the SINR [131], the minimization of the determinant of the MSE matrix [170], the maximization of the mutual information [33, 122, 148], and, more recently, the minimization of the BER [36, 110, 111]. A general problem formulation was considered in [111] as we formulate next.

Consider that the performance of the system is measured by an arbitrary global cost function of the MSEs: $f_0\big(\{\mathrm{MSE}_i\}_{i=1}^{L}\big)$. The problem can then be formulated as the minimization of the cost function subject to the power constraint

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W}}{\text{minimize}} & f_0\left(\{\mathrm{MSE}_i\}\right) \\ \text{subject to} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \le P_0 \end{array} \tag{3.13}$$

or, conversely, as the minimization of the transmit power subject to a constraint on the global performance or quality of the system

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W}}{\text{minimize}} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \\ \text{subject to} & f_0\left(\{\mathrm{MSE}_i\}\right) \le \alpha_0, \end{array} \tag{3.14}$$

where P_0 and α_0 denote the maximum values for the power and for the cost function, respectively.

The cost function f_0 is an indicator of how well the system performs and should be properly selected for the problem at hand. In principle, any function can be used to measure the system performance as long as it is increasing in each argument. Indeed, the increasingness of f_0 is a mild and completely reasonable assumption: if the performance of one of the substream improves while the rest remain unchanged, any reasonable function should not increase the cost.

The problem formulations in (3.13) and (3.14) are in terms of a cost function of the MSEs; however, similar design problems can be straightforwardly formulated with cost functions of the SINRs and of the BERs (when using cost functions of the BERs, it is implicitly assumed that the constellations have already been chosen such that (3.8) can be employed). Interestingly, as will be elaborated in Section 3.3, cost functions of the SINRs and of the BERs can always be rewritten in terms of the MSEs. Therefore, the design with a global measure of performance will be based on (3.13) without loss of generality.

Both formulations (3.13) and (3.14) are essentially equivalent since they describe the same tradeoff curve of performance versus power. Any analytical or numerical solution for one of the problems can be catered for the other problem as well. Alternatively, either problem can be numerically solved by iteratively solving the other one, combined with the bisection method [20, Algorithm 4.1] as is shown in Algorithm 3.1. This observation follows straightforwardly by noting that both problems characterize the same strictly monotonic curve of performance versus power: $f(P)$.

Algorithm 3.1. Bisection method to solve the quality-constrained problem in (3.14) by repeatedly solving the power-constrained problem in (3.13).

Input: Quality required α_0 and desired resolution Δ.

Output: Minimum required power and (implicitly) the solution that achieves it.

```
0. Initialization: Find $P_{\min}$ and $P_{\max}$ such that $f(P_{\min}) > \alpha_0$
   and $f(P_{\max}) < \alpha_0$.
1. Set $P = (P_{\min} + P_{\max})/2$.
2. If $f(P) < \alpha_0$, then set $P_{\max} = P$. Otherwise, set $P_{\min} = P$.
3. If $(P_{\max} - P_{\min}) > \Delta$, then go to step 1.
```

3.2.2 Individual QoS Constraints

Measuring the performance of a MIMO systems with several substreams with a global cost function is in fact a partial view of the problem. A more complete picture is provided by a full characterization of the region of performance or quality that can be achieved by each of the substreams for a given power budget. This would indeed characterize the design as a multi-objective optimization problem along with the Pareto-optimal boundary. This may be a useful formulation, for example, if different services are being supported each with a different QoS requirement and was considered in [114] as formulated next.

Consider a formulation in terms of individual QoS constraints for each of the substreams; for example, in terms of MSE constraints $\mathrm{MSE}_i \leq \rho_i$, where ρ_i denotes the maximum MSE value for the ith substream. The problem can then be formulated as the minimization of the transmitted power subject to the constraints:

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W}}{\text{minimize}} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \\ \text{subject to} & \mathrm{MSE}_i \leq \rho_i \quad 1 \leq i \leq L. \end{array} \tag{3.15}$$

Constraints in terms of SINR and BER can be similarly considered; interestingly, as will be elaborated in Section 3.3, they can always be rewritten in terms of MSEs. Therefore, the design with individual QoS constraints will be based on (3.15) without loss of generality.

Observe that, in a way, the formulation in (3.15) with individual QoS constraints allows a more detailed characterization of the fundamental multi-objective nature of the problem than the formulation in (3.14) with a global measure of quality. For example, the formulation in (3.15) allows the computation of the achievable region of MSEs for a given power budget P_0 (a given set of constraints $\{\rho_i\}$ is achievable if and only if the minimum required power is not greater than P_0). However, at the same time, the formulation in (3.15) with individual constraints can also be seen as a particular case of the formulation in (3.14) with the following global constraint: $\max_i \{\mathrm{MSE}_i / \rho_i\} \leq 1$.

3.3 Optimum Linear Receiver

The problem formulations in (3.13), (3.14), and (3.15) are very difficult problems because they are nonconvex. To see this simply consider the term $|\mathbf{w}_i^{\dagger}\mathbf{H}\mathbf{p}_i - 1|^2$ in (3.6) particularized to the simple real-valued scalar case with trivial channel $h = 1$: the obtained function $(wp-1)^2$ is easily verified to be nonconvex in (w,p).

Alternatively, we can perform the optimization in two stages, since the following holds for any function ϕ (not necessarily convex) [20, Sec. 4.1.3] (see also [12, Sec. 6.4.2]):

$$\inf_{\mathbf{P},\mathbf{W}} \phi(\mathbf{P},\mathbf{W}) = \inf_{\mathbf{P}} \left(\inf_{\mathbf{W}} \phi(\mathbf{P},\mathbf{W}) \right). \tag{3.16}$$

This is commonly called *concentration* in the literature of estimation theory [83]. Fortunately, for a fixed transmit matrix $\mathbf{P}$, the considered problems become convex quadratic in the receive matrix $\mathbf{W}$; in other words, the inner minimization in (3.16) with respect to $\mathbf{W}$ is an easy problem and a closed-form optimal solution can be obtained as shown next.

For the problem formulations (3.13) and (3.14) based on a global measure of performance, in principle, the optimal receive matrix $\mathbf{W}$ may depend on the specific choice of the function f_0. Similarly, for the problem formulation (3.15) with individual QoS constraints, the optimal $\mathbf{W}$ might depend on the choice of the constraints. However, it turns out that the optimal solution is independent of f_0 or the QoS constraints (cf. [111, 114]). The reason is that the design of the receivers $\mathbf{w}_i$ (ith column of $\mathbf{W}$) for the different substreams is completely uncoupled and there is no tradeoff among the MSEs, i.e., the minimization of the MSE of a substream with respect to $\mathbf{w}_i$ (for a fixed transmit matrix $\mathbf{P}$) does not incur any penalty on the other substreams (see, for example, (3.5) where $\hat{x}_i$ is affected only by $\mathbf{w}_i$ and not $\mathbf{w}_j$ for $j \neq i$). This means that we can simultaneously minimize all MSEs, obtaining then the well-known linear minimum MSE (LMMSE) receiver, also termed *Wiener filter* [83] (see also [111, 114]):

$$\begin{aligned} \mathbf{W} &= \left(\mathbf{H}\mathbf{P}\mathbf{P}^\dagger\mathbf{H}^\dagger + \mathbf{I}\right)^{-1}\mathbf{H}\mathbf{P} && (3.17) \\ &= \mathbf{H}\mathbf{P}\left(\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}, && (3.18) \end{aligned}$$

where the second expression follows from the application of the matrix inversion lemma (see (B.12)).

An elegant proof of the optimality of the Wiener filter follows by "completing the squares" (from (3.9)):

$$\begin{aligned} \mathbf{E} &= \left(\mathbf{W}^\dagger\mathbf{H}\mathbf{P} - \mathbf{I}\right)\left(\mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{W} - \mathbf{I}\right) + \mathbf{W}^\dagger\mathbf{W} \\ &= \left(\mathbf{W} - \left(\mathbf{H}\mathbf{P}\mathbf{P}^\dagger\mathbf{H}^\dagger + \mathbf{I}\right)^{-1}\mathbf{H}\mathbf{P}\right)^\dagger\left(\mathbf{H}\mathbf{P}\mathbf{P}^\dagger\mathbf{H}^\dagger + \mathbf{I}\right) \\ &\quad \times \left(\mathbf{W} - \left(\mathbf{H}\mathbf{P}\mathbf{P}^\dagger\mathbf{H}^\dagger + \mathbf{I}\right)^{-1}\mathbf{H}\mathbf{P}\right) \\ &\quad + \mathbf{I} - \mathbf{P}^\dagger\mathbf{H}^\dagger\left(\mathbf{H}\mathbf{P}\mathbf{P}^\dagger\mathbf{H}^\dagger + \mathbf{I}\right)^{-1}\mathbf{H}\mathbf{P} \\ &\geq \mathbf{I} - \mathbf{P}^\dagger\mathbf{H}^\dagger\left(\mathbf{H}\mathbf{P}\mathbf{P}^\dagger\mathbf{H}^\dagger + \mathbf{I}\right)^{-1}\mathbf{H}\mathbf{P}, && (3.19) \end{aligned}$$

where we have used the fact that $\mathbf{X}+\mathbf{Y} \geq \mathbf{X}$ when $\mathbf{Y}$ is positive semidefinite. The lower bound is clearly achieved by (3.17). The concentrated MSE matrix is obtained by plugging (3.17) into (3.9) (or directly from (3.19)) as

$$\begin{aligned} \mathbf{E} &= \mathbf{I} - \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\left(\mathbf{H}\mathbf{P}\mathbf{P}^{\dagger}\mathbf{H}^{\dagger} + \mathbf{I}\right)^{-1}\mathbf{H}\mathbf{P} && (3.20) \\ &= \left(\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}\right)^{-1}, && (3.21) \end{aligned}$$

where the second expression follows from the matrix inversion lemma.

As mentioned in Section 3.2, we also want to consider problem formulations like in (3.13) and (3.14) but in terms of the SINRs or BERs of the substreams rather than the MSEs. It turns out that the Wiener filter (3.18) also maximizes the SINRs and minimizes the BERs of all the substreams, although an additional arbitrary scaling can be included. The SINR can be upper-bounded as

$$\mathrm{SINR}_i = \frac{|\mathbf{w}_i^{\dagger}\mathbf{H}\mathbf{p}_i|^2}{\mathbf{w}_i^{\dagger}\mathbf{R}_{n_i}\mathbf{w}_i} \leq \mathbf{p}_i^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}_{n_i}^{-1}\mathbf{H}\mathbf{p}_i, \tag{3.22}$$

where $\mathbf{R}_{n_i} = \sum_{j \neq i} \mathbf{H}\mathbf{p}_j\mathbf{p}_j^{\dagger}\mathbf{H}^{\dagger} + \mathbf{I}$ is the interference-plus-noise covariance matrix seen by the ith substream. The inequality comes from the Cauchy–Schwarz's inequality (see (B.14)), with vectors $\mathbf{R}_{n_i}^{-1/2}\mathbf{H}\mathbf{p}_i$ and $\mathbf{R}_{n_i}^{1/2}\mathbf{w}_i$, and the upper bound is achieved by

$$\mathbf{w}_i \propto \mathbf{R}_{n_i}^{-1}\mathbf{H}\mathbf{p}_i \propto \left(\mathbf{H}\mathbf{P}\mathbf{P}^{\dagger}\mathbf{H}^{\dagger} + \mathbf{I}\right)^{-1}\mathbf{H}\mathbf{p}_i, \tag{3.23}$$

which is the Wiener filter up to a scaling factor.[8] An alternative derivation of the receiver that maximizes the SINR is by rewriting the SINR as the generalized Rayleigh quotient

$$\mathrm{SINR}_i = \frac{\mathbf{w}_i^{\dagger}\left(\mathbf{H}\mathbf{p}_i\mathbf{p}_i^{\dagger}\mathbf{H}^{\dagger}\right)\mathbf{w}_i}{\mathbf{w}_i^{\dagger}\mathbf{R}_{n_i}\mathbf{w}_i}, \tag{3.24}$$

[8] From the matrix inversion lemma (B.12) or the Woodbury's Identity (B.13), it follows that

$$(\mathbf{H}\mathbf{P}\mathbf{P}^{\dagger}\mathbf{H}^{\dagger} + \mathbf{R}_n)^{-1}\mathbf{H}\mathbf{p}_i = \frac{1}{\mathbf{p}_i^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}_{n_i}^{-1}\mathbf{H}\mathbf{p}_i + 1}\mathbf{R}_{n_i}^{-1}\mathbf{H}\mathbf{p}_i.$$

which is maximized by the generalized eigenvector of the matrix pencil $\left(\mathbf{H}\mathbf{p}_i\mathbf{p}_i^\dagger\mathbf{H}^\dagger, \mathbf{R}_{n_i}\right)$ corresponding to the maximum generalized eigenvalue of $\left(\mathbf{H}\mathbf{p}_i\mathbf{p}_i^\dagger\mathbf{H}^\dagger\right)\mathbf{w}_i = \lambda_{\max}\mathbf{R}_{n_i}\mathbf{w}_i$ [52]:

$$\mathbf{w}_i \propto \mathbf{R}_{n_i}^{-1}\mathbf{H}\mathbf{p}_i. \tag{3.25}$$

Since the BER can be written (or approximated) as a decreasing function of the SINR, as in (3.8), then it follows that the Wiener filter also minimizes all the BERs of the substreams.

The Wiener filter does not guarantee the absence of crosstalk among the substreams. For that purpose, one may want to include the ZF constraint $\mathbf{W}^\dagger\mathbf{H}\mathbf{P} = \mathbf{I}$ in the formulation, which guarantees no crosstalk and substreams normalized to unit gain, obtaining the ZF receiver [83] (see also [108]):

$$\mathbf{W} = \mathbf{H}\mathbf{P}\left(\mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}. \tag{3.26}$$

As before, an elegant proof of the optimality of the ZF receiver follows by "completing the squares" (from (3.9) and $\mathbf{W}^\dagger\mathbf{H}\mathbf{P} = \mathbf{I}$):

$$\begin{aligned}\mathbf{E} &= \mathbf{W}^\dagger\mathbf{W} \\ &= \left(\mathbf{W} - \mathbf{H}\mathbf{P}\left(\mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}\right)^\dagger\left(\mathbf{W} - \mathbf{H}\mathbf{P}\left(\mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}\right) \\ &\quad + \left(\mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1} \\ &\geq \left(\mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}.\end{aligned} \tag{3.27}$$

The same result is obtained by maximizing the SINR:

$$\mathrm{SINR}_i = \frac{1}{\mathbf{w}_i^\dagger\mathbf{w}_i} \tag{3.28}$$

and, as before, by minimizing the BER.

3.3.1 Summary of MMSE and ZF Receivers

Summarizing, the MMSE receiver (or Wiener filter) and the ZF receiver (with the additional ZF constraint) are optimum in that they (i) minimize simultaneously all the MSEs, (ii) maximize simultaneously all the SINRs, and (iii) minimize simultaneously all the BERs. Both receivers can be compactly written as

$$\mathbf{W} = \mathbf{H}\mathbf{P}\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}, \tag{3.29}$$

where

$$\nu \triangleq \begin{cases} 1 & \text{for the MMSE receiver} \\ 0 & \text{for the ZF receiver} \end{cases} . \tag{3.30}$$

The MSE matrix reduces then to the following concentrated expression:

$$\mathbf{E} = \left(\nu\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}\right)^{-1}. \tag{3.31}$$

As a last observation, note that the optimum receive matrix in (3.29) can be interpreted as the concatenation of three stages: (1) an implicit noise-whitening stage $\mathbf{R}_n^{-1/2}$ (recall that we assume that the noise has already been pre-whitened), (2) a matched filter stage $\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}$ (where $\mathbf{H}$ denotes the whitened channel matrix), and (3) an MSE or ZF stage $\left(\nu\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}\right)^{-1}$.

3.3.2 Relation Among Different Measures of Performance

It is convenient now to relate the different measures of performance, namely, MSE, SINR, and BER, to the concentrated MSE matrix in (3.31).

From the definition of MSE matrix, the individual MSEs are given by the diagonal elements of the MSE matrix:

$$\mathrm{MSE}_i = \left[\left(\nu\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}\right)^{-1}\right]_{ii}. \tag{3.32}$$

It turns out that the SINRs and the MSEs are trivially related when using the MMSE or ZF receivers by [108, 111, 114, 157][9]

$$\mathrm{SINR}_i = \frac{1}{\mathrm{MSE}_i} - \nu. \tag{3.33}$$

Finally, the BERs can also be written as a function of the MSEs:

$$\mathrm{BER}_i = g_i\left(\mathrm{MSE}_i\right) \triangleq \tilde{g}_i\left(\mathrm{SINR}_i = \mathrm{MSE}_i^{-1} - \nu\right), \tag{3.34}$$

where $\tilde{g}_i$ was defined in (3.8).

[9] The relation is obvious for the ZF receiver as $\mathrm{MSE}_i = \mathbf{w}_i^{\dagger}\mathbf{R}_n\mathbf{w}_i$ and $\mathrm{SINR}_i = 1/(\mathbf{w}_i^{\dagger}\mathbf{R}_n\mathbf{w}_i)$. For the MMSE receiver, it follows from the matrix inversion lemma in (B.12) that (see (3.20) and (3.22)) $\mathrm{MSE}_i = \left[\mathbf{I} - \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\left(\mathbf{H}\mathbf{P}\mathbf{P}^{\dagger}\mathbf{H}^{\dagger} + \mathbf{R}_n\right)^{-1}\mathbf{H}\mathbf{P}\right]_{ii} = 1/\left(1 + \mathbf{p}_i^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}_{n_i}^{-1}\mathbf{H}\mathbf{p}_i\right)$.

Since the SINR and the BER can be expressed as a function of the MSE, given in (3.33) and (3.34), it suffices to focus on cost functions of the MSEs without loss of generality as formulated in Section 3.2.

Characterization of the BER Function for QAM Constellations

For the important case of QAM constellation, the BER as a function of the SINR (i.e., the function $\tilde{g}_i$) in (3.11) and (3.12) is a convex decreasing function (see Appendix 3.A). Most importantly, as can be seen from Figure 3.3 (and is proved in Appendix 3.A), the BER as a function of the MSE (i.e., the function g_i) given by

$$\mathrm{BER}(\mathrm{MSE}) \approx \frac{\alpha}{\log_2 M}\mathcal{Q}\left(\sqrt{\beta\left(\mathrm{MSE}^{-1}-\nu\right)}\right) \tag{3.35}$$

and, similarly, the Chernoff approximation

$$\mathrm{BER}(\mathrm{MSE}) \approx \frac{\alpha}{2\log_2 M}\,e^{-\beta/2\left(\mathrm{MSE}^{-1}-\nu\right)} \tag{3.36}$$

are convex increasing functions for the following range of sufficiently small MSE values[10]:

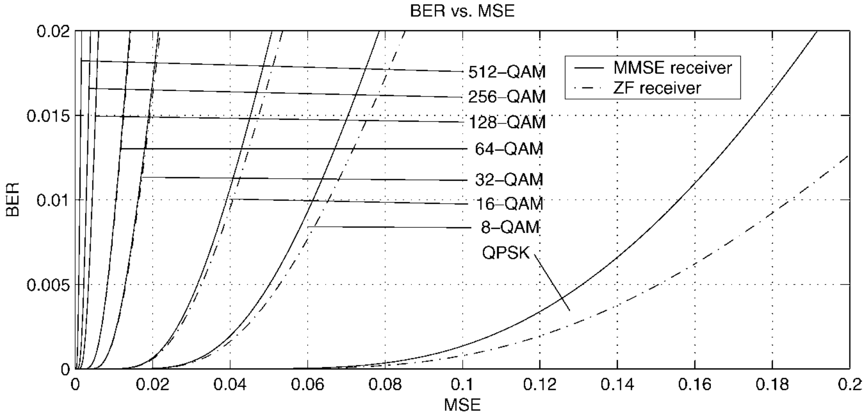


Fig. 3.3 BER as a function of the MSE for different QAM constellations.

[10] The same convexity result holds for the exact BER expression as given in [24], as opposed to using just the first term in the approximation as in (3.11).

- MMSE receiver [108, 111]:

$$\text{MSE} \leq \begin{cases} \left(\beta + 3 - \sqrt{\beta^2 - 10\beta + 9}\right)/8 & \text{for Q-function approx.} \\ \beta/4 & \text{for Chernoff approx.} \end{cases} \tag{3.37}$$

- ZF receiver [108]:

$$\text{MSE} \leq \begin{cases} \beta/3 & \text{for Q-function approx.} \\ \beta/4 & \text{for Chernoff approx.} \end{cases} \tag{3.38}$$

As a rule-of-thumb, the BER as a function of the MSE is convex for a BER less than 2×10^{-2} (this is a mild assumption, since practical systems have in general a smaller uncoded BER[11]); interestingly, for BPSK and QPSK constellations the BER function is always convex [108, 111] (see Appendix 3.A).

3.3.3 Diagonal versus Nondiagonal Transmission

To better understand the underlying structure of the communication when using an MMSE/ZF receiver, consider a transmitter of the form:

$$\mathbf{P} = \mathbf{V}_H \boldsymbol{\Sigma} \mathbf{Q}, \tag{3.39}$$

where $\mathbf{V}_H$ is a (semi-)unitary matrix with columns equal to the right singular vectors of the channel matrix $\mathbf{H}$ corresponding to the L largest singular values, $\boldsymbol{\Sigma} = \text{diag}\left(\sqrt{\mathbf{p}}\right)$ is a diagonal matrix containing the square-root of the power allocation $\mathbf{p}$ over the channel eigenmodes, and $\mathbf{Q}$ is a unitary matrix (also termed "rotation" matrix). The global transmit–receive process $\hat{\mathbf{x}} = \mathbf{W}^\dagger \left(\mathbf{HPx} + \mathbf{n}\right)$ can then be rewritten explicitly as

$$\hat{\mathbf{x}} = \mathbf{Q}^\dagger \left(\nu \mathbf{I} + \boldsymbol{\Sigma}^\dagger \mathbf{D}_H \boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\Sigma}^\dagger \mathbf{D}_H^{1/2} \left(\mathbf{D}_H^{1/2} \boldsymbol{\Sigma} \mathbf{Q} \mathbf{x} + \bar{\mathbf{n}}\right), \tag{3.40}$$

where $\bar{\mathbf{n}}$ is an equivalent normalized white noise and $\mathbf{D}_H = \mathbf{V}_H^\dagger \mathbf{H}^\dagger \mathbf{H} \mathbf{V}_H$ is the diagonalized squared channel matrix. For the ZF receiver ($\nu = 0$), the previous expression simplifies to

$$\hat{\mathbf{x}} = \mathbf{x} + \mathbf{Q} \left(\boldsymbol{\Sigma}^\dagger \mathbf{D}_H \boldsymbol{\Sigma}\right)^{-1/2} \bar{\mathbf{n}}, \tag{3.41}$$

[11] Given an uncoded bit error probability of at most 10^{-2} and using a proper coding scheme, coded bit error probabilities with acceptable low values such as 10^{-6} can be obtained.

which clearly satisfies the condition $\mathbf{W}^{\dagger}\mathbf{H}\mathbf{P} = \mathbf{I}$ (by definition) but has, in general, a correlated noise among the subchannels. In other words, when using the ZF receiver, the global transmission is not really diagonal or parallel since the noise is colored.

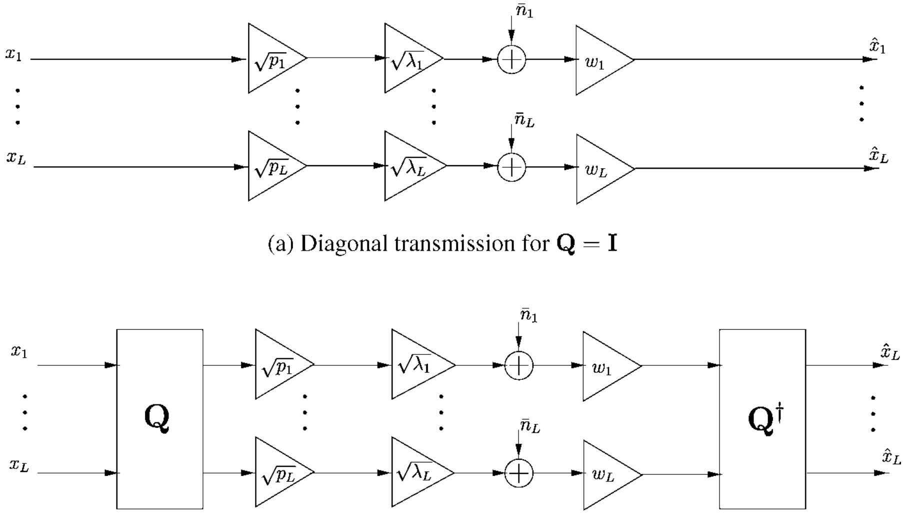
In fact, the fully diagonal or parallel transmission does not depend on whether the ZF or the MMSE receivers are used, but on the choice of the "rotation" $\mathbf{Q}$. Indeed, by setting $\mathbf{Q} = \mathbf{I}$, the global transmit–receive process (3.40) is fully diagonalized:

$$\hat{\mathbf{x}} = \left(\nu\mathbf{I} + \mathbf{\Sigma}^{\dagger}\mathbf{D}_H\mathbf{\Sigma}\right)^{-1}\mathbf{\Sigma}^{\dagger}\mathbf{D}_H^{1/2}\left(\mathbf{D}_H^{1/2}\mathbf{\Sigma}\mathbf{x} + \bar{\mathbf{n}}\right), \qquad (3.42)$$

which can be rewritten as

$$\hat{x}_i = \omega_i\left(\sqrt{p_i\lambda_i}\,x_i + \bar{n}_i\right) \quad 1 \leq i \leq L, \qquad (3.43)$$

where $\omega_i = \sqrt{p_i\lambda_i}/\left(\nu + p_i\lambda_i\right)$ (see Figure 3.4). Hence, by choosing $\mathbf{Q} = \mathbf{I}$, the MMSE receiver also results in a diagonal transmission (which is never the case in the traditional approach, where only the receiver is optimized). This is all summarized in the following.



(a) Diagonal transmission for $\mathbf{Q} = \mathbf{I}$

(b) Nondiagonal transmission (diagonal + rotation)

Fig. 3.4 Scheme of diagonal and nondiagonal (due to the rotation) transmissions.

Remark 3.1. The transmitter in (3.39) leads to a diagonal transmission (in the sense of no crosstalk and independent noise) if $\mathbf{Q} = \mathbf{I}$ regardless of whether the receiver is an MMSE or ZF. Otherwise, the transmission structure is not diagonal and is composed of an inner diagonal structure placed between a "pre-rotation" and a "post-rotation" operators, as shown in Figure 3.4.

3.4 Optimum Linear Transmitter with Global Measure of Performance: Schur-Convex/Concave Cost Functions

This section deals with problem formulations with a global measure of performance, either as the minimization of a global cost function subject to a power constraint as in (3.13) or the minimization of the power subject to a constraint on the global performance as in (3.14). For the sake of concreteness, we will focus on (3.13), although equivalent results hold for (3.14) as well.

All the results in this chapter are based on majorization theory for which the reader is referred to Chapter 2 and to the 1979 textbook by Marshall and Olkin [97].

The optimal receiver $\mathbf{W}$ has already been obtained in Section 3.3 (see (3.29)) as the Wiener filter or MMSE receiver, and also as the ZF receiver under the ZF constraint. The MSE matrix is then given by (3.31) and the MSEs by (3.32): $\mathrm{MSE}_i = \left[\left(\nu \mathbf{I} + \mathbf{P}^\dagger \mathbf{H}^\dagger \mathbf{H} \mathbf{P} \right)^{-1} \right]_{ii}$, where $\nu = 0$ for the ZF receiver and $\nu = 1$ for the MMSE receiver. Therefore, the problem of minimizing a cost function of the MSEs as a function of the linear precoder $\mathbf{P}$ at the transmitter can be formulated as (recall that cost functions of the SINRs and BERs can always be reformulated as functions of the MSEs):

$$\begin{array}{ll} \underset{\mathbf{P}}{\text{minimize}} & f_0\left(\left\{ \left[\left(\nu \mathbf{I} + \mathbf{P}^\dagger \mathbf{H}^\dagger \mathbf{H} \mathbf{P} \right)^{-1} \right]_{ii} \right\} \right) \\ \text{subject to} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^\dagger \right) \le P_0. \end{array} \tag{3.44}$$

This is a nonconvex problem as can be easily seen by ignoring the cost function f_0 and considering the simple real-valued scalar case with trivial channel $h = 1$ and unit noise variance: the obtained function $1/\left(\nu + p^2 \right)$ is easily verified to be nonconvex in p.

We will start by obtaining a suboptimal solution based on imposing a diagonal structure on the transmission. As shown in Section 3.3.2, this is accomplished by imposing the following form on the transmit matrix:

$$\mathbf{P} = \mathbf{V}_H \boldsymbol{\Sigma}, \tag{3.45}$$

where $\mathbf{V}_H$ is a (semi-)unitary matrix with columns equal to the right singular vectors of the channel matrix $\mathbf{H}$ corresponding to the L largest singular values and $\boldsymbol{\Sigma} = \mathrm{diag}\left(\sqrt{\mathbf{p}}\right)$ is a diagonal matrix containing the square-root of the power allocation $\mathbf{p}$ over the channel eigenmodes. Under such a diagonal structure, the expression for the MSEs simplifies to the scalar and convex expression:

$$\mathrm{MSE}_i = \frac{1}{\nu + p_i \lambda_{H,i}} \qquad 1 \leq i \leq L, \tag{3.46}$$

where the $\lambda_{H,i}$'s denote the L largest eigenvalues of matrix $\mathbf{H}^{\dagger}\mathbf{H}$. Problem (3.44) becomes then

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & f_0\left(\left\{\frac{1}{\nu + p_i \lambda_{H,i}}\right\}\right) \\ \text{subject to} & \mathbf{1}^T\mathbf{p} \leq P_0 \\ & \mathbf{p} \geq \mathbf{0}, \end{array} \tag{3.47}$$

which is convex provided that f_0 is convex (recall that f_0 is increasing, cf. Section 3.2) [20].

However, the simple reformulation in (3.47) need not be optimal in the sense that its solution need not be an optimal solution of the original problem formulation in (3.44). In the following, we provide a truly equivalent simple reformulation of the original complicated nonconvex problem (3.44) based on majorization theory as was originally derived in [111].[12]

Theorem 3.1. An optimal solution $\mathbf{P}$ to the complicated nonconvex matrix-valued problem in (3.44), where $f_0 : \mathbb{R}^L \to \mathbb{R}$ is a function increasing in each variable, can be characterized as follows:

[12] Recall that for the sake of notation we assume $L \leq \min(n_R, n_T)$. Otherwise, in Theorem 3.1, only the $\check{L} \triangleq \min(L, n_R, n_T)$ largest singular values of $\mathbf{H}$ are used and the term $\mathrm{diag}\left(\sqrt{\mathbf{p}}\right)$ appearing in (3.48) and (3.49) must be replaced by $[\mathbf{0}, \mathrm{diag}(\sqrt{\mathbf{p}})]$, where $\mathbf{p}$ has dimension $\check{L}$ [111].

- If the function f_0 is Schur-concave:

$$\mathbf{P} = \mathbf{V}_H \operatorname{diag}\left(\sqrt{\mathbf{p}}\right), \tag{3.48}$$

where $\mathbf{V}_H$ is a (semi-)unitary matrix with columns equal to the right singular vectors of matrix $\mathbf{H}$ corresponding to the L largest singular values in increasing order and the power allocation $\mathbf{p}$ is the solution to the simple problem (3.47) where the $\lambda_{H,i}$'s denote the L largest eigenvalues of matrix $\mathbf{H}^\dagger\mathbf{H}$ in increasing order.

- If the function f_0 is Schur-convex:

$$\mathbf{P} = \mathbf{V}_H \operatorname{diag}\left(\sqrt{\mathbf{p}}\right)\mathbf{Q}, \tag{3.49}$$

where $\mathbf{V}_H$ is as before, the power allocation $\mathbf{p}$ is given by the waterfilling expression

$$p_i = \left(\mu\lambda_{H,i}^{-1/2} - \lambda_{H,i}^{-1}\right)^+, \quad 1 \le i \le L \tag{3.50}$$

with μ chosen to satisfy $\sum_{i=1}^{L} p_i = P_0$, and $\mathbf{Q}$ is a unitary matrix such that $\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}$ has identical diagonal elements. This "rotation" matrix $\mathbf{Q}$ can be chosen as any unitary matrix that satisfies $\left|[\mathbf{Q}]_{ik}\right| = \left|[\mathbf{Q}]_{il}\right|$, $\forall i,k,l$, such as the unitary Discrete Fourier Transform (DFT) matrix or the unitary Hadamard matrix (when the dimensions are appropriate such as a power of two [157, p. 66]), as well as with Algorithm 2.2.

If, in addition, the function f_0 is *strictly* Schur-concave/convex, then the previous characterization must necessarily be satisfied by *any* optimal solution.

Furthermore, the previous characterization follows verbatim for functions that may not be Schur-concave/convex on $\mathbb{R}^L$ but are minimized when the arguments are sorted in decreasing order (or any fixed ordering) and then they become Schur-concave/convex on $\left\{\mathbf{X} \in \mathbb{R}^L \mid x_1 \ge x_2 \ge \cdots \ge x_L\right\}$.

Proof. The proof hinges on majorization theory for which the reader is referred to Chapter 2 or [97]. The key result on which the proof is based is given in Corollaries 2.1 and 2.4 and outlined next for convenience. For a Hermitian matrix $\mathbf{M}$ and a unitary matrix $\mathbf{Q}$, it follows that

$$\mathbf{1}(\mathbf{M}) \prec \mathbf{d}\left(\mathbf{Q}^\dagger\mathbf{M}\mathbf{Q}\right) \prec \boldsymbol{\lambda}(\mathbf{M}), \tag{3.51}$$

where $\mathbf{d}(\mathbf{A})$ and $\boldsymbol{\lambda}(\mathbf{A})$ denote the diagonal elements and eigenvalues of $\mathbf{A}$, respectively, and $\mathbf{1}$ denotes the vector with identical components equal to the average of the diagonal elements of $\mathbf{A}$. More interestingly, $\mathbf{Q}$ can always be chosen such that the diagonal elements are equal to one extreme or the other. To achieve $\mathbf{d}\left(\mathbf{Q}^\dagger\mathbf{M}\mathbf{Q}\right) = \mathbf{1}(\mathbf{M})$, $\mathbf{Q}$ has to be chosen such that $\mathbf{Q}^\dagger\mathbf{M}\mathbf{Q}$ has equal diagonal elements; whereas to achieve $\mathbf{d}\left(\mathbf{Q}^\dagger\mathbf{M}\mathbf{Q}\right) = \boldsymbol{\lambda}(\mathbf{M})$, $\mathbf{Q}$ has to be chosen to diagonalize matrix $\mathbf{M}$, i.e., equal to the eigenvectors of $\mathbf{M}$. Observe that $\mathbf{1}\left(\mathbf{Q}^\dagger\mathbf{M}\mathbf{Q}\right) = \mathbf{1}(\mathbf{M})$ and $\boldsymbol{\lambda}\left(\mathbf{Q}^\dagger\mathbf{M}\mathbf{Q}\right) = \boldsymbol{\lambda}(\mathbf{M})$.

Another critical observation is that the MSE matrix $\mathbf{E}(\mathbf{P}) = \left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}$ satisfies

$$\mathbf{E}(\mathbf{P}\mathbf{Q}) = \mathbf{Q}^\dagger\mathbf{E}(\mathbf{P})\,\mathbf{Q} \tag{3.52}$$

for any unitary matrix $\mathbf{Q}$. Now we can proceed with the actual proof.

If f_0 is Schur-concave, it follows from Definition 2.4 that $f_0(\mathbf{d}(\mathbf{E}(\mathbf{P}))) \geq f_0(\boldsymbol{\lambda}(\mathbf{E}(\mathbf{P})))$. As pointed out before, for any given $\mathbf{P}$, it is always possible to achieve the lower bound by using instead the transmit matrix $\tilde{\mathbf{P}} = \mathbf{P}\mathbf{Q}$ such that $\mathbf{E}(\mathbf{P}\mathbf{Q}) = \mathbf{Q}^\dagger\mathbf{E}(\mathbf{P})\,\mathbf{Q}$ is diagonal, in which case $f_0\big(\mathbf{d}\big(\mathbf{E}\big(\tilde{\mathbf{P}}\big)\big)\big) = f_0\big(\boldsymbol{\lambda}\big(\mathbf{E}\big(\tilde{\mathbf{P}}\big)\big)\big) = f_0\big(\boldsymbol{\lambda}\big(\mathbf{E}(\mathbf{P})\big)\big)$ (observe that the power remains the same: $\mathrm{Tr}\left(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^\dagger\right) = \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^\dagger\right)$). This implies that for Schur-concave functions there is an optimal $\mathbf{P}$ such that $\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}$ is diagonal or, equivalently, such that $\mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}$ is diagonal. For strictly Schur-concave functions, the term $\mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}$ must necessarily be diagonal as the cost function could be strictly decreased otherwise. We will further assume without loss of generality that the diagonal elements of $\mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}$ are in increasing order to cope with the case in which the arguments of f_0 are restricted to be in decreasing order.

If f_0 is Schur-convex the opposite happens, it follows from Definition 2.4 that $f_0(\mathbf{d}(\mathbf{E}(\mathbf{P}))) \geq f_0(\mathbf{1}(\mathbf{E}(\mathbf{P})))$. As pointed out before, for any given $\mathbf{P}$, it is always possible to achieve the lower bound by using instead the transmit matrix $\tilde{\mathbf{P}} = \mathbf{P}\mathbf{Q}$ such that $\mathbf{E}(\mathbf{P}\mathbf{Q}) = \mathbf{Q}^\dagger\mathbf{E}(\mathbf{P})\,\mathbf{Q}$

has equal diagonal elements, in which case $f_0(\mathbf{d}(\mathbf{E}(\tilde{\mathbf{P}}))) = f_0(\mathbf{1}(\mathbf{E}(\tilde{\mathbf{P}}))) = f_0(\mathbf{1}(\mathbf{E}(\mathbf{P})))$ (observe that the power remains the same). This implies that for Schur-convex functions there is an optimal $\mathbf{P}$ such that $(\nu\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P})^{-1}$ has equal diagonal elements given by the average: $\left[(\nu\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P})^{-1}\right]_{ii} = (1/L)\,\mathrm{Tr}\,(\nu\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P})^{-1}$. For strictly Schur-convex functions, the MSE matrix must necessarily have equal diagonal elements as the cost function could be strictly decreased otherwise. At this point, we can focus on the minimization of $\mathrm{Tr}\,(\nu\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P})^{-1}$, for which we can assume without loss of generality that $\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}$ is diagonal with diagonal elements in increasing order (the order in this case is completely irrelevant). Of course, the optimal transmit matrix will then be given by $\tilde{\mathbf{P}} = \mathbf{P}\mathbf{Q}$, where $\mathbf{P}$ is such that $\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}$ is diagonal and $\mathbf{Q}$ makes the diagonal elements of the MSE matrix $\mathbf{Q}^{\dagger}(\nu\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P})^{-1}\mathbf{Q}$ equal. The "rotation" matrix $\mathbf{Q}$ can be computed with the general Algorithm 2.2. However, since $(\nu\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P})^{-1}$ is a diagonal matrix, it follows that any $\mathbf{Q}$ that satisfies the following conditions will do the trick: $|[\mathbf{Q}]_{ik}| = |[\mathbf{Q}]_{il}|$, $\forall i,k,l$ (cf. Lemma 2.10).

Now, given that $\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}$ is a diagonal matrix with diagonal elements in increasing order, we can invoke Lemma 3.16 in Appendix 3.B to conclude that the optimal $\mathbf{P}$ can be written as $\mathbf{P} = \mathbf{V}_H\boldsymbol{\Sigma}$, where $\mathbf{V}_H$ and $\boldsymbol{\Sigma} = \mathrm{diag}\,(\sqrt{\mathbf{p}})$ are defined in the theorem statement. The simplified problem reformulation in (3.47) follows then straightforwardly for the case of f_0 Schur-concave. For the case of f_0 Schur-convex, the simplified reformulation is (cf. Section 3.4.3)

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & \sum_{i=1}^{L} \frac{1}{\nu + p_i\,\lambda_{H,i}} \\ \text{subject to} & \mathbf{1}^T\mathbf{p} \leq P_0 \\ & \mathbf{p} \geq \mathbf{0} \end{array} \tag{3.53}$$

with solution given by (3.50).

A last observation is required to complete the proof. Since we have assumed that the elements of the diagonal matrix $\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}$ are in increasing order, we should include the constraints $p_i\,\lambda_{H,i} \leq p_{i+1}\,\lambda_{H,i+1}$, for $1 \leq i \leq L-1$, in the simplified problems. However, this is not really necessary as argued next. For Schur-convex functions, the problem simplifies to the minimization of $\mathrm{Tr}\,(\nu\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P})^{-1}$ where the ordering is clearly irrelevant. For Schur-concave functions on $\mathbb{R}^L$, the ordering is also irrelevant as they are insensitive to permutations of the arguments. Finally, for functions that are minimized when the arguments are in decreasing order and only then they become Schur-concave, it is obvious that the optimal solution to the simplified problem in (3.47) will satisfy the ordering as, otherwise, we could simply swap the terms $p_i \lambda_{H,i}$ and $p_{i+1} \lambda_{H,i+1}$ further decreasing the objective value. In addition to the previous justification, it is interesting to point out that, whenever $\lambda_{H,i} \neq \lambda_{H,i+1}$, we cannot possibly have $p_i \lambda_{H,i} > p_{i+1} \lambda_{H,i+1}$ at an optimal point as we could otherwise swap the terms with two consequences: (i) the objective value would not increase, and (ii) the required power could be strictly decreased by noticing that the optimum ordering of the eigenvalues is increasing $\lambda_{H,i} < \lambda_{H,i+1}$ as follows from Lemma 3.17. □

Observe that Theorem 3.1 gives a clear answer to the question of whether the diagonal transmission is optimal: when the cost function is Schur-concave then the diagonal structure is optimal, but when the cost function is Schur-convex then the optimal structure is not diagonal anymore due to the rotation matrix (see Figure 3.4 and Section 3.3.2 for more details). It also answers the question of what cost functions we can choose and still be able to optimally solve the problem design: as long as the cost function is Schur-concave or Schur-convex, we can solve the original complicated design problem in (3.44) by solving the simple problem (3.47) or with (3.50), respectively.

Remarkably, for Schur-convex functions, the optimal power allocation is independent of the cost function f_0 as is given by (3.50).

The class of Schur-concave/convex functions is actually extremely wide and embraces most common functions one can think of as cost functions. Tables 3.1 and 3.2 contain a list of examples of Schur-concave and Schur-convex functions, respectively, along with the optimal power allocation $\mathbf{p}$ and rotation matrix $\mathbf{Q}$. Sections 3.4.2 and 3.4.3 consider in detail the list of examples.

3.4.1 An Illustrative Example: Minimax Design

To illustrate the usefulness of the characterization in Theorem 3.1, we now consider a minimax design or, in other words, the optimization of

Table 3.1 Examples of problem formulations that can be rewritten as minimization of a Schur-concave cost function of the MSEs.

Problem	Optimal solution
Minim. sum of MSEs or (**E**) [2, 90, 129, 131, 171]	$p_i = \left(\mu \lambda_{H,i}^{-1/2} - \nu \lambda_{H,i}^{-1}\right)^+$ $\mathbf{Q} = \mathbf{I}$
Minim. weighted sum of MSEs [90, 130]	$p_i = \left(\mu \alpha_i^{1/2} \lambda_{H,i}^{-1/2} - \nu \lambda_{H,i}^{-1}\right)^+$ $\mathbf{Q} = \mathbf{I}$
Minim. weighted product of MSEs	$p_i = \left(\mu \alpha_i - \nu \lambda_{H,i}^{-1}\right)^+$ $\mathbf{Q} = \mathbf{I}$
Minim. det (**E**) [170]	$p_i = \left(\mu - \nu \lambda_{H,i}^{-1}\right)^+$ $\mathbf{Q} = \mathbf{I}$
Maxim. mutual information [33]	$p_i = \left(\mu - \lambda_{H,i}^{-1}\right)^+$ $\mathbf{Q} = \mathbf{I}$
Maxim. weighted sum of SINRs	$p_i = \begin{cases} P_0 & \text{if } \left(\alpha_i \lambda_{H,i}\right) \text{ is max.} \\ 0 & \text{otherwise} \end{cases}$ $\mathbf{Q} = \mathbf{I}$
Maxim. product of SINRs	$p_i = P_0/L$ (unif. power alloc.) $\mathbf{Q} = \mathbf{I}$
Maxim. weighted product of SINRs	$p_i = P_0 \alpha_i / \sum_j \alpha_j$ $\mathbf{Q} = \mathbf{I}$

Table 3.2 Examples of problem formulations that can be rewritten as minimization of a Schur-convex cost function of the MSEs.

Problem	Optimal solution
Minim. maximum of MSEs Maxim. minimum of SINRs[13] Maxim. harmonic mean of SINRs Minim. average BER (equal constellations) Minim. maximum of BERs	$p_i = \left(\mu \lambda_{H,i}^{-1/2} - \nu \lambda_{H,i}^{-1}\right)^+$ $\mathbf{Q} =$ Fourier/Hadamard

the worst substream:

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W}}{\text{minimize}} & \max_i \{\mathrm{MSE}_i\} \\ \text{subject to} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^\dagger\right) \le P_0. \end{array} \tag{3.54}$$

From Section 3.3, we know that the optimal receiver is the Wiener filter of MMSE receiver:

$$\mathbf{W} = \mathbf{H}\mathbf{P}\left(\mathbf{I} + \mathbf{P}^\dagger \mathbf{H}^\dagger \mathbf{H}\mathbf{P}\right)^{-1} \tag{3.55}$$

[13] For the ZF receiver, the maximization of the harmonic mean of the SINRs is equivalent to the minimization of the unweighted sum of the MSEs, which can be classified as both Schur-concave and Schur-convex (since it is invariant to rotations).

and the problem reduces then to

$$\begin{array}{ll} \underset{\mathbf{P}}{\text{minimize}} & \max_i \left\{ \left[\left(\nu \mathbf{I} + \mathbf{P}^\dagger \mathbf{H}^\dagger \mathbf{H} \mathbf{P} \right)^{-1} \right]_{ii} \right\} \\ \text{subject to} & \text{Tr}\left(\mathbf{P}\mathbf{P}^\dagger \right) \leq P_0. \end{array} \tag{3.56}$$

Now, without invoking Theorem 3.1, we can only opt for a suboptimal choice that makes the problem easy to solve. For example, we can impose a diagonal structure by using a transmitter of the form $\mathbf{P} = \mathbf{V}_H \mathbf{\Sigma}$ with $\mathbf{\Sigma} = \text{diag}\left(\sqrt{\mathbf{p}}\right)$. The MSE expression simplifies to $\text{MSE}_i = 1/\left(1 + p_i \lambda_{H,i}\right)$ and the minimax problem can be finally written as the following simple convex optimization problem:

$$\begin{array}{lll} \underset{t,\mathbf{p}}{\text{minimize}} & t & \\ \text{subject to} & t \geq \frac{1}{\nu + p_i \lambda_{H,i}} & 1 \leq i \leq L \\ & \mathbf{1}^T \mathbf{p} \leq P_0 & \\ & \mathbf{p} \geq \mathbf{0}, & \end{array} \tag{3.57}$$

with solution given by

$$p_i = P_0 \frac{\lambda_{H,i}^{-1}}{\sum_j \lambda_{H,j}^{-1}} \quad 1 \leq i \leq L. \tag{3.58}$$

However, this solution is in principle suboptimal because we have imposed a diagonal structure. With the aid of Theorem 3.1, we can solve the original problem in an optimal way by proceeding as follows:

(1) Notice that $f_0(\mathbf{x}) = \max_i \{x_i\}$ is a Schur-convex function (cf. Section 3.4.3.1).

(2) Invoke Theorem 3.1 to obtain the form of the optimal solution: $\mathbf{P} = \mathbf{V}_H \,\text{diag}\left(\sqrt{\mathbf{p}}\right) \mathbf{Q}$.

(3) Use the optimal power allocation $p_i = \left(\mu \lambda_{H,i}^{-1/2} - \lambda_{H,i}^{-1}\right)^+$.

In fact, we can now realize that the previous solution imposing a diagonal structure is indeed suboptimal because a rotation is required for this particular problem.

3.4.2 Examples of Schur-Concave Cost Functions

From Theorem 3.1, the optimal solution for Schur-concave cost functions is $\mathbf{P} = \mathbf{V}_H \,\text{diag}\left(\sqrt{\mathbf{p}}\right)$, where $\mathbf{V}_H$ contains the best right singular vectors of the channel matrix $\mathbf{H}$ and $\mathbf{p}$ is the power allocation over the channel eigenmodes. The MSEs are given by

$$\text{MSE}_i = \frac{1}{\nu + p_i \lambda_{H,i}} \quad 1 \leq i \leq L \tag{3.59}$$

and the original optimization problem (3.44) reduces to (3.47), whose solution clearly depends on the particular choice of f_0.

In many cases, as we will see, the optimal power allocation follows a waterfilling form: $p_i = (\mu a_i - b_i)^+$, where the a_i's and b_i's are some fixed numbers and μ is a waterlevel that has to be found to satisfy the power constraint with equality $\sum_i p_i = P_0$ or, more generally, some condition $g(\mu) = 0$. The numerical evaluation of such waterfilling solutions can be done efficiently in practice either by bisection or by hypothesis testing as explored in detail in [108, 112]. For convenience, we reproduce in Appendix 3.E a general algorithm based on hypothesis testing with a worst-case complexity of L iterations.

In the following, we consider several important examples of Schur-concave cost functions either on $\mathbb{R}^L$ or on $\left\{\mathbf{x} \in \mathbb{R}^L \mid x_1 \geq x_2 \geq \cdots \geq x_L\right\}$ (for functions that are minimized with arguments in decreasing order). Table 3.1 summarizes the list of examples giving the optimal power allocation $\mathbf{p}$ and rotation matrix $\mathbf{Q} = \mathbf{I}$.

3.4.2.1 Minimization of the Weighted Sum of MSEs

The minimization of the sum of MSEs was considered in [90, 130, 131, 171]. We consider the weighted version as in [90, 130]. The cost function is

$$f_0(\{\text{MSE}_i\}) = \sum_{i=1}^{L} (\alpha_i \, \text{MSE}_i), \tag{3.60}$$

which happens to be minimized with arguments in decreasing order (with weights in increasing order) and it becomes then Schur-concave.

Lemma 3.2. The function $f_0(\mathbf{x}) = \sum_i (\alpha_i x_i)$ (assuming $\alpha_i \leq \alpha_{i+1}$) is minimized when the arguments are in decreasing order $x_i \geq x_{i+1}$ and it is then a Schur-concave function on $\left\{\mathbf{x} \in \mathbb{R}^L \mid x_1 \geq x_2 \geq \cdots \geq x_L\right\}$.

Proof. See Appendix 3.D. □

Since the function is Schur-concave, by Theorem 3.1 the original problem (3.44) can be rewritten as the simple convex problem:

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & \sum_i \alpha_i \frac{1}{\nu+p_i\lambda_{H,i}} \\ \text{subject to} & \mathbf{1}^T\mathbf{p}\le P_0 \\ & \mathbf{p}\ge \mathbf{0}, \end{array} \tag{3.61}$$

with optimal solution given (from the KKT optimality conditions) by the following waterfilling expression:

$$p_i = \left(\mu\alpha_i^{1/2}\lambda_{H,i}^{-1/2} - \nu\lambda_{H,i}^{-1}\right)^+ \quad 1\le i\le L, \tag{3.62}$$

where μ is the waterlevel chosen such that $\sum_i p_i = P_0$.

3.4.2.2 Minimization of the Exponentially Weighted Product of MSEs

The cost function corresponding to the minimization of the exponentially weighted product of MSEs is

$$f_0\left(\{\text{MSE}_i\}\right) = \prod_{i=1}^{L}\left(\text{MSE}_i\right)^{\alpha_i}, \tag{3.63}$$

which happens to be minimized with arguments in decreasing order (with weights in increasing order) and it becomes then Schur-concave.

Lemma 3.3. The function $f_0(\mathbf{x}) = \prod_i x_i^{\alpha_i}$ (assuming $\alpha_i \le \alpha_{i+1}$) is minimized when the arguments are in decreasing order $x_i \ge x_{i+1}$ and it is then a Schur-concave function on $\left\{\mathbf{x}\in\mathbb{R}^L \mid x_1 \ge x_2 \ge \cdots \ge x_L > 0\right\}$.

Proof. See Appendix 3.D. □

Since the function is Schur-concave, by Theorem 3.1 the original problem (3.44) can be rewritten as the simple convex problem (since the objective is log-convex, it is also convex [20]):

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & \prod_i \left(\frac{1}{\nu+p_i\lambda_{H,i}}\right)^{\alpha_i} \\ \text{subject to} & \mathbf{1}^T\mathbf{p}\le P_0 \\ & \mathbf{p}\ge \mathbf{0}, \end{array} \tag{3.64}$$

with optimal solution given (from the KKT optimality conditions) by the following waterfilling expression:

$$p_i = \left(\mu\alpha_i - \nu\lambda_{H,i}^{-1}\right)^+ \quad 1 \le i \le L, \tag{3.65}$$

where μ is the waterlevel chosen such that $\sum_i p_i = P_0$. Note that for the unweighted case $\alpha_i = 1$, (3.65) becomes the classical capacity-achieving waterfilling solution [33, 122].

3.4.2.3 Minimization of the Determinant of the MSE Matrix

The minimization of the determinant of the MSE matrix was considered in [170]. We now show how this particular criterion is easily accommodated in our framework as a Schur-concave function of the diagonal elements of the MSE matrix $\mathbf{E}$.

Using the fact that $\mathbf{X} \ge \mathbf{Y} \Rightarrow |\mathbf{X}| \ge |\mathbf{Y}|$, it follows that $|\mathbf{E}|$ is minimized with the Wiener filter (3.29). From the expression of the MSE matrix $\mathbf{E}$ in (3.31), it is clear that $|\mathbf{E}|$ does not change if the transmit matrix $\mathbf{P}$ is post-multiplied by a unitary matrix. Therefore, we can always choose a rotation matrix so that $\mathbf{E}$ is diagonal without loss of optimality (as already known from [170]) and then

$$|\mathbf{E}| = \prod_j \lambda_j(\mathbf{E}) = \prod_j [\mathbf{E}]_{jj}. \tag{3.66}$$

Therefore, the minimization of $|\mathbf{E}|$ can be seen as the minimization of the (unweighted) product of the MSEs treated in Section 3.4.2.2.

3.4.2.4 Maximization of Mutual Information

The maximization of the mutual information leads to a capacity-achieving solution [33] and is formulated as

$$\max_{\mathbf{Q}} \quad I = \log|\mathbf{I} + \mathbf{H}\mathbf{Q}\mathbf{H}^\dagger|, \tag{3.67}$$

where $\mathbf{Q}$ is the transmit covariance matrix. Using the fact that $|\mathbf{I} + \mathbf{X}\mathbf{Y}| = |\mathbf{I} + \mathbf{Y}\mathbf{X}|$ and that $\mathbf{Q} = \mathbf{P}\mathbf{P}^\dagger$ (from (3.2)), the mutual information can be written as $I = \log\left|\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right|$. Comparing this with (3.31), it follows that the mutual information can be expressed (see [29] for detailed connections between the mutual information and the MSE matrix) as

$$I = -\log |\mathbf{E}| \tag{3.68}$$

and, therefore, the maximization of I is equivalent to the minimization of $|\mathbf{E}|$ treated in Section 3.4.2.3.

Hence, the minimization of the unweighted product of the MSEs, the minimization of the determinant of the MSE matrix, and the maximization of the mutual information are all equivalent criteria with solution given by a channel-diagonalizing structure and the classical capacity-achieving waterfilling for the power allocation:

$$p_i = \left(\mu - \lambda_{H,i}^{-1}\right)^+ . \tag{3.69}$$

3.4.2.5 Maximization of the Weighted Sum of SINRs

The objective function to be maximized is

$$\tilde{f}_0\left(\{\mathrm{SINR}_i\}\right) = \sum_{i=1}^{L} (\alpha_i \, \mathrm{SINR}_i) \tag{3.70}$$

from which we can define, via (3.33), the following function of the MSEs to be minimized:

$$f_0\left(\{\mathrm{MSE}_i\}\right) \triangleq -\tilde{f}_0(\{\mathrm{MSE}_i^{-1} - \nu\}) = -\sum_{i=1}^{L} \alpha_i \left(\mathrm{MSE}_i^{-1} - \nu\right), \tag{3.71}$$

which happens to be minimized with arguments in decreasing order (with weights in increasing order) and it becomes then Schur-concave.

Lemma 3.4. The function $f_0(\mathbf{x}) = -\sum_i \alpha_i \left(x_i^{-1} - \nu\right)$ (assuming $\alpha_i \leq \alpha_{i+1}$) is minimized when the arguments are in decreasing order $x_i \geq x_{i+1} > 0$ and it is then a Schur-concave function on $\left\{\mathbf{x} \in \mathbb{R}^L \mid x_1 \geq x_2 \geq \cdots \geq x_L > 0\right\}$.

Proof. See Appendix 3.D. □

Since the function is Schur-concave, by Theorem 3.1 the original problem (3.44) can be rewritten as the simple convex problem (linear problem):

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{maximize}} & \sum_i p_i \left(\alpha_i \, \lambda_{H,i}\right) \\ \text{subject to} & \mathbf{1}^T \mathbf{p} \le P_0 \\ & \mathbf{p} \ge \mathbf{0}, \end{array} \tag{3.72}$$

with optimal solution given by allocating all the available power to the eigenmode with maximum weighted gain $(\alpha_i \, \lambda_{H,i})$ (otherwise the objective value could be increased by transferring power from other eigenmodes to this eigenmode). Although this solution maximizes indeed the weighted sum of the SINRs, it is a terrible solution in practice due to the extremely poor spectral efficiency (only one substream would be conveying information). This criterion gives a pathological solution and should not be used.

3.4.2.6 Maximization of the Exponentially Weighted Product of SINRs

The objective function to be maximized is

$$\tilde{f}_0\left(\{\mathrm{SINR}_i\}\right) = \prod_{i=1}^{L} \left(\mathrm{SINR}_i\right)^{\alpha_i} \tag{3.73}$$

from which we can define, via (3.33), the following function of the MSEs to be minimized:

$$f_0\left(\{\mathrm{MSE}_i\}\right) \triangleq -\tilde{f}_0(\{\mathrm{MSE}_i^{-1} - \nu\}) = -\prod_{i=1}^{L} \left(\mathrm{MSE}_i^{-1} - \nu\right)^{\alpha_i}, \tag{3.74}$$

which happens to be minimized with arguments in decreasing order (with weights in increasing order) and it becomes then Schur-concave. Observe that the maximization of the product of the SINRs is equivalent to the maximization of the sum of the SINRs expressed in dB.

Lemma 3.5. The function $f_0(\mathbf{x}) = -\prod_i \left(x_i^{-1} - \nu\right)^{\alpha_i}$ (assuming $\alpha_i \le \alpha_{i+1}$) is minimized when the arguments are in decreasing order $1 \ge x_i \ge x_{i+1} > 0$ and it is then a Schur-concave function on $\left\{\mathbf{x} \in \mathbb{R}^L \mid 0.5 \ge x_1 \ge x_2 \ge \cdots \ge x_L > 0\right\}$.

Proof. See Appendix 3.D. □

By Lemma 3.5, the objective function (3.74) is Schur-concave provided that $\mathrm{MSE}_i \leq 0.5 \; \forall i$, which is a mild assumption since an MSE greater than 0.5 is unreasonable for a practical communication system. Therefore, by Theorem 3.1 the original problem (3.44) can be rewritten as the simple convex problem (the weighted geometric mean is a concave function[14] [20, 124]):

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{maximize}} & \prod_i \left(p_i \, \lambda_{H,i}\right)^{\tilde{\alpha}_i} \\ \text{subject to} & \mathbf{1}^T \mathbf{p} \leq P_0 \\ & \mathbf{p} \geq \mathbf{0}, \end{array} \tag{3.75}$$

where $\tilde{\alpha}_i \triangleq \alpha_i / (\sum_j \alpha_j)$. The optimal solution is given (from the KKT optimality conditions) by:

$$p_i = \tilde{\alpha}_i \, P_0. \tag{3.76}$$

For the unweighted case $\alpha_i = 1$, the solution reduces to a uniform power allocation

$$p_i = P_0 / L. \tag{3.77}$$

Interestingly, for the unweighted case, the problem can be reformulated as the maximization of the geometric mean subject to the arithmetic mean:

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{maximize}} & \prod_i p_i^{1/L} \\ \text{subject to} & (1/L) \sum_i p_i \leq P_0 / L \\ & \mathbf{p} \geq \mathbf{0}. \end{array} \tag{3.78}$$

From the arithmetic–geometric mean inequality $\left(\prod_k x_k\right)^{1/N} \leq \frac{1}{N} \sum_k x_k$ (with equality if and only if $x_k = x_l \; \forall k, l$) [96, p. 202][66], it follows that the optimal solution is indeed the uniform power allocation.

[14] The concavity of the geometric mean is easily verified by showing that the Hessian matrix is positive semi-definite for positive values of the arguments. The extension to include boundary points (points with zero-valued arguments) is straightforward either by using a continuity argument to show that $f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \geq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$ for $0 \leq \theta \leq 1$ or by considering the epigraph of the function and using [93, Prop. 2.7.4].

3.4.2.7 Maximization of the Weighted Product of 1+SINRs

The objective function to be maximized is

$$\tilde{f}_0\left(\{\mathrm{SINR}_i\}\right) = \prod_{i=1}^{L}\left(1 + \mathrm{SINR}_i\right) \tag{3.79}$$

from which we can define, via (3.33), the following function of the MSEs to be minimized (for the MMSE receiver):

$$f_0\left(\{\mathrm{MSE}_i\}\right) \triangleq 1/\tilde{f}_0(\{\mathrm{MSE}_i^{-1} - 1\}) = \prod_{i=1}^{L}\mathrm{MSE}_i, \tag{3.80}$$

which is equivalent to the minimization of the unweighted product of MSEs in (3.63), to the minimization of the determinant of the MSE matrix in (3.66), and to the maximization of the mutual information in (3.67), with solution given by the capacity-achieving expression (3.69). This result is completely natural since maximizing the logarithm of (3.79) is tantamount to maximizing the mutual information $I = \sum_i \log\left(1 + \mathrm{SINR}_i\right)$.

3.4.2.8 Minimization of the Product of BERs

In terms of BER, the minimization of the arithmetic mean (cf. Section 3.4.3.4) is more meaningful than the geometric mean, but we will also consider the product of the BERs for completeness. The objective function to minimize is

$$\tilde{f}_0\left(\{\mathrm{BER}_i\}\right) = \prod_{i=1}^{L}\mathrm{BER}_i \tag{3.81}$$

from which we can define, via (3.33) and (3.34), the following function of the MSEs to be minimized:

$$f_0\left(\{\mathrm{MSE}_i\}\right) = \prod_{i=1}^{L} g_i\left(\mathrm{MSE}_i\right), \tag{3.82}$$

where functions g_i is defined in (3.34). As stated next, this function is Schur-concave when the constellations used on the different substreams are equal.

Lemma 3.6. The function $f_0(\mathbf{x}) = \prod_i g(x_i)$, where g is given by (3.35) or (3.36), is a Schur-concave function on $\left\{\mathbf{x} \in \mathbb{R}^L \mid 0 < x_i \leq \theta \quad \forall i\right\}$ for sufficiently small θ such that $\left(\frac{\partial g(x)}{\partial x}\right)^2 \geq g(x) \frac{\partial^2 g(x)}{\partial x^2}$ for $0 < x \leq \theta$.

Proof. See Appendix 3.D. □

Since the function is Schur-concave, by Theorem 3.1 the original problem (3.44) can be rewritten as a simple problem.

3.4.3 Examples of Schur-Convex Cost Functions

From Theorem 3.1, the optimal solution for Schur-convex cost functions is $\mathbf{P} = \mathbf{V}_H \operatorname{diag}(\sqrt{\mathbf{p}})\, \mathbf{Q}$, where $\mathbf{V}_H$ contains the best eigenvectors of the channel matrix, $\mathbf{p}$ is the power allocation over the channel eigenmodes, and $\mathbf{Q}$ is an additional rotation which can be chosen equal to the DFT or Hadamard matrices. In that case, the MSE matrix $\mathbf{E}$ is nondiagonal and has equal diagonal elements given by

$$\mathrm{MSE}_i = \frac{1}{L} \operatorname{Tr}(\mathbf{E}) = \frac{1}{L} \sum_{j=1}^{L} \frac{1}{\nu + p_j \lambda_{H,j}} \quad 1 \leq i \leq L \tag{3.83}$$

and the original optimization problem (3.44) reduces to (3.53), whose solution does not depend on the particular choice of f_0 and is given (from the KKT optimality conditions) by the following waterfilling expression:

$$p_i = \left(\mu \lambda_{H,i}^{-1/2} - \nu \lambda_{H,i}^{-1}\right)^+ \quad 1 \leq i \leq L, \tag{3.84}$$

where μ is the waterlevel chosen such that $\sum_i p_i = P_0$ (see Appendix 3.E for a practical algorithm to evaluate the waterfilling expression). Note that for the ZF receiver ($\nu = 0$), the waterfilling solution (3.84) simplifies to

$$p_i = P_0 \frac{\lambda_{H,i}^{-1/2}}{\sum_j \lambda_{H,j}^{-1/2}} \quad 1 \leq i \leq L. \tag{3.85}$$

It is interesting to remark that problem (3.53) is equivalent to the minimization of the trace of the MSE matrix. Hence, among the infinite solutions that minimize $\mathrm{Tr}\left(\mathbf{E}\right)$, only that which yields equal diagonal elements in $\mathbf{E}$ is the optimal solution for a Schur-convex objective function (which is obtained in fact with the waterfilling solution in (3.84) and the rotation $\mathbf{Q}$ as described in Theorem 3.1).

In the following, we consider several important examples of Schur-convex cost functions either on $\mathbb{R}^L$ or on $\left\{\mathbf{x} \in \mathbb{R}^L \mid x_1 \geq x_2 \geq \cdots \geq x_L\right\}$ (for functions that are minimized with arguments in decreasing order). Table 3.2 summarizes the list of examples giving the optimal power allocation $\mathbf{p}$ and rotation matrix $\mathbf{Q}$.

3.4.3.1 Minimization of the Maximum of the MSEs

In general, the overall performance of a system (average BER) is dominated by the substream with highest MSE. It makes sense then to minimize the maximum of the MSEs. The cost function is

$$f_0\left(\{\mathrm{MSE}_i\}\right) = \max_{1 \leq i \leq L} \left\{\mathrm{MSE}_i\right\}, \tag{3.86}$$

which is Schur-convex.

Lemma 3.7. The function $f_0\left(\mathbf{x}\right) = \max_i \left\{x_i\right\}$ is a Schur-convex function.

Proof. See Appendix 3.D. □

Since the function is Schur-convex, by Theorem 3.1 the solution to the original problem (3.44) is given by the precoder (3.49) and the waterfilling power allocation (3.50).

3.4.3.2 Maximization of the Harmonic Mean of the SINRs

The objective function to be maximized is

$$\tilde{f}_0\left(\{\mathrm{SINR}_i\}\right) = \left(\sum_{i=1}^{L} \frac{1}{\mathrm{SINR}_i}\right)^{-1} \tag{3.87}$$

from which we can define, via (3.33), the following function of the MSEs to be minimized:

$$f_0\left(\{\mathrm{MSE}_i\}\right) \triangleq 1/\tilde{f}_0(\{\mathrm{MSE}_i^{-1} - \nu\}) = \sum_{i=1}^{L} \frac{\mathrm{MSE}_i}{1 - \nu\,\mathrm{MSE}_i}, \tag{3.88}$$

which is Schur-convex.

Lemma 3.8. The function $f_0(\mathbf{x}) = \sum_i \frac{x_i}{1-\nu x_i}$ is a Schur-convex function on $\left\{\mathbf{x} \in \mathbb{R}^L \mid 0 \le x_i < 1 \quad \forall i\right\}$.

Proof. See Appendix 3.D. □

Since the function is Schur-convex, by Theorem 3.1 the solution to the original problem (3.44) is given by the precoder (3.49) and the waterfilling power allocation (3.50).

3.4.3.3 Maximization of the Minimum of the SINRs

The objective function to be maximized is

$$\tilde{f}_0\left(\{\mathrm{SINR}_i\}\right) = \min_{1 \le i \le L} \{\mathrm{SINR}_i\} \tag{3.89}$$

from which we can define, via (3.33), the following function of the MSEs to be minimized:

$$f_0\left(\{\mathrm{MSE}_i\}\right) \triangleq -\tilde{f}_0(\{\mathrm{MSE}_i^{-1} - \nu\}) = \max_{1 \le i \le L} \left\{\nu - \mathrm{MSE}_i^{-1}\right\}, \tag{3.90}$$

which is equivalent to the minimization of $\max_i \{\mathrm{MSE}_i\}$ treated with detail in Section 3.4.3.1. In [130], the same criterion was used imposing a channel diagonal structure.

3.4.3.4 Minimization of the Average BER

Since the ultimate measure of a digital communication system is the BER (cf. Section 3.1.1), the average BER over the substreams is a meaningful criterion. The problem of minimizing the average BER in MIMO systems has recently been receiving a considerable attention.

In [106], the problem was treated in detail imposing a diagonal structure in the transmission (the approximation of the BER by the Chernoff upper bound was also considered). The minimum BER solution without imposing any structural constraint has been independently obtained in [36] and in [111], where it has been shown that the optimal solution consists of a nondiagonal transmission (nondiagonal MSE matrix).

The objective function to minimize is

$$\tilde{f}_0\left(\{\text{BER}_i\}\right) = \sum_{i=1}^{L} \text{BER}_i \tag{3.91}$$

from which we can define, via (3.33) and (3.34), the following function of the MSEs to be minimized:

$$f_0\left(\{\text{MSE}_i\}\right) = \sum_{i=1}^{L} g_i\left(\text{MSE}_i\right), \tag{3.92}$$

where the functions g_i's are defined as in (3.34). As stated next, the average BER is a Schur-convex function of the MSEs when the constellations used on the different substreams are equal.

Lemma 3.9. The function $f_0\left(\mathbf{x}\right) = \sum_i g\left(x_i\right)$, where g is given by (3.35) or (3.36), is a Schur-convex function on $\left\{\mathbf{x} \in \mathbb{R}^L \mid 0 < x_i \leq \theta \quad \forall i\right\}$ for sufficiently small θ such that $g\left(x_i\right) \leq 2 \times 10^{-2}\ \forall i$.

Proof. See Appendix 3.D. □

Since the function is Schur-convex, by Theorem 3.1 the solution to the original problem (3.44) is given by the precoder (3.49) and the waterfilling power allocation (3.50).

3.4.3.5 Minimization of the Maximum of the BERs

In general, the overall performance of a system (average BER) is dominated by the substream with highest BER. It makes sense then to minimize the maximum of the BERs. The cost function is

$$f_0\left(\{\text{BER}_i\}\right) = \max_{1\leq i\leq L} \left\{\text{BER}_i\right\}. \tag{3.93}$$

Since the BER is an increasing function of the MSE (cf. Section 3.3.2), minimizing the maximum BER is tantamount to minimizing the maximum MSE as considered in detail in Section 3.4.3.1, which happens to be a Schur-convex function.

3.4.4 Numerical Results

We now show some numerical results for different designs in terms of BER as a function of the SNR at the transmitter (i.e., transmitted power normalized with the noise variance). The results are averaged over 10^7 realizations of the channel matrix $\mathbf{H}$ with i.i.d. elements following a Gaussian distribution with zero mean and unit variance.

Figure 3.5 compares the performance of the design based on the minimization of the maximum MSE (MAX-MSE) in the cases of suboptimal diagonal solution and optimal nondiagonal solution (cf. Section 3.4.1). For the case of $L = 3$ over a 4×4 MIMO channel, the

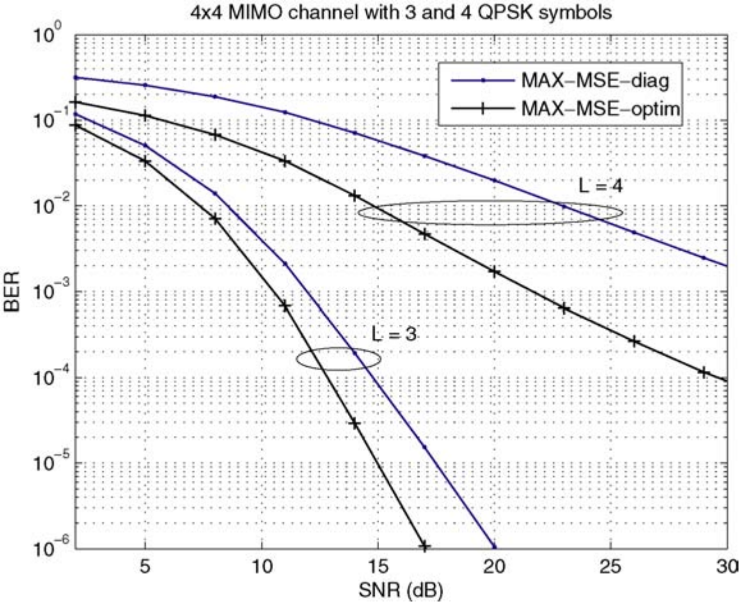


Fig. 3.5 BER versus SNR for a 4×4 MIMO channel with $L = 3$ transmitted QPSK symbols for the MAX-MSE design according to a suboptimal diagonal solution and the optimal nondiagonal solution.

savings in SNR are on the order of 2 dB, whereas for the case of $L = 4$ (fully loaded system), the savings in SNR are much larger on the order of 7 dB.

Figure 3.6 shows the performance of different designs based on three Schur-concave functions (cf. Section 3.4.2): minimization of the product of the MSEs (PROD-MSE), maximization of the product of the SINRs (PROD-SINR), minimization of the sum of the MSEs (SUM-MSE); and also on three Schur-convex functions (cf. Section 3.4.3): maximization of the harmonic SINR (HARM-SINR), minimization of the maximum MSE (MAX-MSE), minimization of the sum (or average) or BERs (SUM-BER). Observe that, as expected, the performance of the three designs based on Schur-convex functions is identical as they have the same solution (cf. Theorem 3.1). Notice that Schur-convex designs are better than Schur-concave ones as the transmission of the symbols is performed distributedly over the channel eigenvalues, which provides robustness against realizations with small eigenvalues.

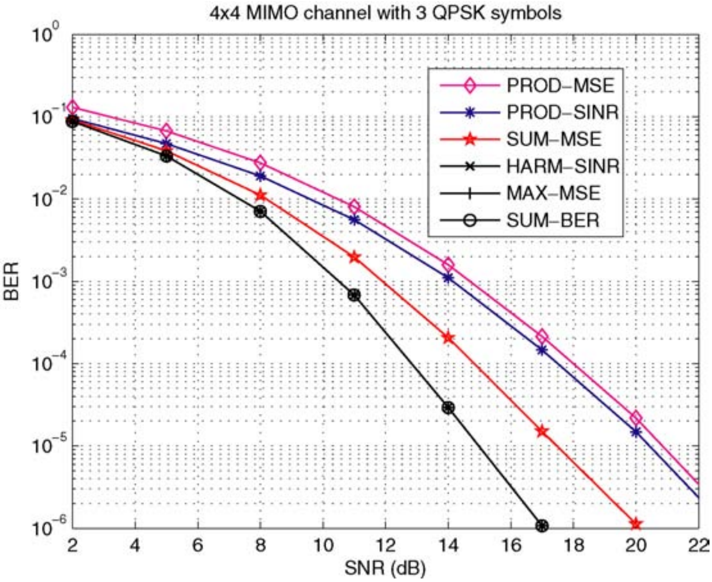


Fig. 3.6 BER versus SNR for a 4 × 4 MIMO channel with $L = 3$ transmitted QPSK symbols for designs based on three Schur-concave functions (PROD-MSE, PROD-SINR, and SUM-MSE) and three Schur-convex functions (HARM-SINR, MAX-MSE, and SUM-BER).

The performance when using a ZF receiver is extremely close to that obtained with an MMSE receiver (on the order of 0.5 dB) and has not been included in the plots for the sake of clarity. (Note that in the classical equalization setup with no transmitter optimization, the difference in performance between the ZF and MMSE receivers is typically more significant.)

3.5 Optimum Linear Transmitter with Individual QoS Constraints

This section deals with the problem formulation with individual QoS constraints and minimum transmit power as in (3.15).

The optimal receiver $\mathbf{W}$ has already been obtained in Section 3.3 (see (3.29)) as the Wiener filter or MMSE receiver, and also as the ZF receiver under the ZF constraint. The MSE matrix is then given by (3.31) and the MSEs by (3.32): $\mathrm{MSE}_i = \left[\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}\right]_{ii}$, where $\nu = 0$ for the ZF receiver and $\nu = 1$ for the MMSE receiver. Therefore, the problem of minimizing the transmit power subject to individual MSE QoS constraints as a function of the linear precoder $\mathbf{P}$ at the transmitter can be formulated as (recall that QoS constraints in terms of the SINRs and BERs can always be reformulated as MSE QoS constraints):

$$\begin{array}{ll} \underset{\mathbf{P}}{\text{minimize}} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^\dagger\right) \\ \text{subject to} & \left[\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}\right]_{ii} \le \rho_i \quad 1 \le i \le L, \end{array} \tag{3.94}$$

which is a nonconvex problem for the same reasons as (3.44).

We will start by obtaining a suboptimal solution based on imposing a diagonal structure on the transmission. As shown in Section 3.3.2, this is accomplished by imposing the following form on the transmit matrix:

$$\mathbf{P} = \mathbf{V}_H \boldsymbol{\Sigma}, \tag{3.95}$$

where $\mathbf{V}_H$ is a (semi-)unitary matrix with columns equal to the right singular vectors of the channel matrix $\mathbf{H}$ corresponding to the L largest singular values and $\boldsymbol{\Sigma} = \mathrm{diag}\left(\sqrt{\mathbf{p}}\right)$ is a diagonal matrix containing the square-root of the power allocation $\mathbf{p}$ over the channel eigenmodes. Under such a diagonal structure, the expression for the MSEs simplifies to the scalar and convex expression:

$$\mathrm{MSE}_i = \frac{1}{\nu + p_i \lambda_{H,i}} \quad 1 \le i \le L, \tag{3.96}$$

where the $\lambda_{H,i}$'s denote the L largest eigenvalues of matrix $\mathbf{H}^\dagger \mathbf{H}$. Problem (3.94) becomes then

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & \mathbf{1}^T \mathbf{p} \\ \text{subject to} & \frac{1}{\nu + p_i \lambda_{H,i}} \le \rho_i \quad 1 \le i \le L, \\ & \mathbf{p} \ge \mathbf{0}, \end{array} \tag{3.97}$$

which is a convex problem.

However, the simple reformulation in (3.97) need not be optimal in the sense that its solution need not be an optimal solution to the original problem formulation in (3.94). In the following, we provide a truly equivalent simple reformulation of the original complicated nonconvex problem (3.94) based on majorization theory (cf. Chapter 2 and [97]) as was originally derived in [114].[15]

We will start with the simple case where the MSE QoS constraints are equal, $\rho_i = \rho$ for all i, and then we will proceed with the general case.

3.5.1 Equal MSE QoS Constraints

For equal QoS constraints, the solution turns out to be equivalent to that for Schur-convex functions in Theorem 3.1. The reason is that we can equivalently write the QoS constraints in (3.94) as $f_0(\{\mathrm{MSE}_i\}) \le \rho$, where $f_0(\mathbf{x}) = \max_i \{x_i\}$ which is a Schur-convex function.

[15] Recall that for the sake of notation we assume $L \le \min(n_R, n_T)$. Otherwise, in Theorems 3.10 and 3.11, only the $\check{L} \triangleq \min(L, n_R, n_T)$ largest singular values of $\mathbf{H}$ are used, the term $\mathrm{diag}(\sqrt{\mathbf{p}})$ appearing in (3.98) and (3.105) must be replaced by $[\mathbf{0}, \mathrm{diag}(\sqrt{\mathbf{p}})]$, where $\mathbf{p}$ has dimension $\check{L}$, the constraint in (3.99) must be replaced by $\sum_{i=1}^{\check{L}} \frac{1}{\nu + p_i \lambda_{H,i}} \le L\rho - (L - \check{L})$, and the constraints in (3.106) must be replaced by $\sum_{j=i}^{\check{L}} \frac{1}{\nu + p_j \lambda_{H,j}} \le \sum_{j=i}^{\check{L}} \tilde{\rho}_j$, where $\tilde{\rho}_1 \triangleq \sum_{j=1}^{(L-\check{L})+1} \rho_j - (L - \check{L})$ and $\tilde{\rho}_{i(>1)} \triangleq \rho_{i+(L-\check{L})}$ [114].

Theorem 3.10. The optimal solution $\mathbf{P}$ to the complicated nonconvex matrix-valued problem in (3.94) with equal QoS constraints, i.e., $\rho_i = \rho$ for all i, satisfies all the constraints with equality and is given by

$$\mathbf{P} = \mathbf{V}_H \,\mathrm{diag}\left(\sqrt{\mathbf{p}}\right)\mathbf{Q}, \tag{3.98}$$

where $\mathbf{V}_H$ is a (semi-)unitary matrix with columns equal to the right singular vectors of matrix $\mathbf{H}$ corresponding to the L largest singular values in increasing order, $\mathbf{Q}$ is a unitary matrix such that $\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}$ has identical diagonal elements (e.g., the unitary DFT matrix or the unitary Hadamard matrix), and the power allocation $\mathbf{p}$ is the solution to the following simple convex problem:

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & \mathbf{1}^T\mathbf{p} \\ \text{subject to} & \sum_{i=1}^{L} \frac{1}{\nu + p_i\,\lambda_{H,i}} \leq L\rho, \\ & \mathbf{p} \geq \mathbf{0}, \end{array} \tag{3.99}$$

where the $\lambda_{H,i}$'s denote the L largest eigenvalues of matrix $\mathbf{H}^\dagger\mathbf{H}$ in increasing order.

Proof. Start by rewriting the original problem (3.94) as

$$\begin{array}{ll} \underset{\mathbf{P}}{\text{minimize}} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^\dagger\right) \\ \text{subject to} & \max_i \left[\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}\right]_{ii} \leq \rho. \end{array} \tag{3.100}$$

Now, for a given $\mathbf{P}$, choose a unitary matrix $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\mathbf{Q}^\dagger$ is diagonal (with diagonal elements in increasing order). Then, matrix $\left(\nu\mathbf{I} + \tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}$, where $\tilde{\mathbf{P}} = \mathbf{P}\mathbf{Q}^\dagger$, is diagonal and the original MSE matrix is given by $\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1} = \mathbf{Q}^\dagger\left(\nu\mathbf{I} + \tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}\mathbf{Q}$. We can then rewrite the problem in terms of $\tilde{\mathbf{P}}$ and $\mathbf{Q}$ as

$$\begin{array}{lll} \underset{\tilde{\mathbf{P}},\mathbf{Q}}{\text{minimize}} & \mathrm{Tr}\left(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^\dagger\right) & \\ \text{subject to} & \tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}} & \text{diagonal (increasing diag. elements)} \\ & \max_i \left[\mathbf{Q}^\dagger\left(\nu\mathbf{I} + \tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}\mathbf{Q}\right]_{ii} \leq \rho. & \end{array} \tag{3.101}$$

Observe now that $(1/L)\,\mathrm{Tr}\,(\mathbf{E}) \leq \max_i\,[\mathbf{E}]_{ii}$ with equality if and only if all the diagonal elements are equal. This means that

$$\max_i \left[\mathbf{Q}^\dagger \left(\nu\mathbf{I} + \tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}\mathbf{Q}\right]_{ii} \geq \frac{1}{L}\,\mathrm{Tr}\left(\left(\nu\mathbf{I} + \tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}\right) \tag{3.102}$$

and the lower bound is achieved if and only if the diagonal elements of $\mathbf{Q}^\dagger\left(\nu\mathbf{I} + \tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}\mathbf{Q}$ are equal. From majorization theory (see Corollary 2.4), we know that we can always find such a $\mathbf{Q}$ as stated in the theorem (see Lemma 2.10 for details).

Thus, we can now reformulate the problem without $\mathbf{Q}$ as

$$\begin{array}{ll} \underset{\tilde{\mathbf{P}}}{\text{minimize}} & \mathrm{Tr}\left(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^\dagger\right) \\ \text{subject to} & \tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}} \quad \text{diagonal (increasing diag. elements)} \\ & \mathrm{Tr}\left(\left(\nu\mathbf{I} + \tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}\right) \leq L\rho. \end{array} \tag{3.103}$$

At this point, given that matrix $\tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}$ is diagonal with diagonal elements in increasing order, we can invoke Lemma 3.16 in Appendix 3.B (as in the proof of Theorem 3.1) to conclude that the optimal $\mathbf{P}$ can be written as $\mathbf{P} = \mathbf{V}_H\mathbf{\Sigma}$, where $\mathbf{V}_H$ and $\mathbf{\Sigma} = \mathrm{diag}\left(\sqrt{\mathbf{p}}\right)$ are defined in the theorem statement. The simple convex reformulation in (3.99) follows then straightforwardly.

As argued in the proof of Theorem 3.1, it is important to point out that, even though we have assumed the elements of the diagonal matrix $\mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}$ in increasing order, it is not necessary to explicitly include the constraints $p_i\,\lambda_{H,i} \leq p_{i+1}\,\lambda_{H,i+1}$, for $1 \leq i \leq L-1$, in the simplified problem (3.99) as the optimal solution to (3.99) already satisfies those constraints. □

It is important to remark that the optimal solution to the simplified problem in (3.99) is given (from the KKT optimality conditions) by the waterfilling solution:

$$p_i = \left(\mu\,\lambda_{H,i}^{-1/2} - \nu\,\lambda_{H,i}^{-1}\right)^+ \quad 1 \leq i \leq L, \tag{3.104}$$

where μ is the waterlevel chosen such that $\sum_{i=1}^{L}\frac{1}{\nu+p_i\,\lambda_{H,i}} \leq L\rho$ (see Appendix 3.E for a practical algorithm to evaluate the waterfilling expression).

3.5.2 Arbitrary MSE QoS Constraints

Having seen the simple case of equal QoS constraints, we are now ready to see the general case with arbitrary QoS constraints.

Theorem 3.11. The optimal solution $\mathbf{P}$ to the complicated nonconvex matrix-valued problem in (3.94) with QoS constraints in decreasing order w.l.o.g., i.e., $\rho_i \geq \rho_{i+1}$ for $1 \leq i < L$, satisfies all the constraints with equality and is given by

$$\mathbf{P} = \mathbf{V}_H \,\mathrm{diag}\left(\sqrt{\mathbf{p}}\right) \mathbf{Q}, \tag{3.105}$$

where $\mathbf{V}_H$ is a (semi-)unitary matrix with columns equal to the right singular vectors of matrix $\mathbf{H}$ corresponding to the L largest singular values in increasing order, $\mathbf{Q}$ is a unitary matrix such that $\left[\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}\right]_{ii} = \rho_i$ for all i (which can be computed with Algorithm 2.2), and the power allocation $\mathbf{p}$ is the solution to the following simple convex problem:

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & \mathbf{1}^T\mathbf{p} \\ \text{subject to} & \sum_{j=i}^{L} \frac{1}{\nu + p_j\,\lambda_{H,j}} \leq \sum_{j=i}^{L} \rho_j \quad 1 \leq i \leq L, \\ & \mathbf{p} \geq \mathbf{0}, \end{array} \tag{3.106}$$

where the $\lambda_{H,i}$'s denote the L largest eigenvalues of matrix $\mathbf{H}^\dagger\mathbf{H}$ in increasing order.

Proof. Start by rewriting the original problem (3.94) as

$$\begin{array}{ll} \underset{\mathbf{P}}{\text{minimize}} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^\dagger\right) \\ \text{subject to} & \mathbf{d}\left(\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}\right) \leq \boldsymbol{\rho}, \end{array} \tag{3.107}$$

where $\mathbf{d}\left(\cdot\right)$ denotes a vector with the diagonal elements of a matrix.

For the moment we can proceed as in the proof for equal QoS constraints by choosing, for a given $\mathbf{P}$, a unitary matrix $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\mathbf{Q}^\dagger$ is diagonal (with diagonal elements in increasing order). Then, matrix $\left(\nu\mathbf{I} + \tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}$, where $\tilde{\mathbf{P}} = \mathbf{P}\mathbf{Q}^\dagger$, is diagonal and the original MSE matrix is given by $\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1} =$

$\mathbf{Q}^\dagger\big(\nu\mathbf{I}+\tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\big)^{-1}\mathbf{Q}$. We can then rewrite the problem in terms of $\tilde{\mathbf{P}}$ and $\mathbf{Q}$ as

$$\begin{array}{ll} \underset{\tilde{\mathbf{P}},\mathbf{Q}}{\text{minimize}} & \operatorname{Tr}\big(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^\dagger\big) \\ \text{subject to} & \tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}} \quad \text{diagonal (increasing diag. elements)} \\ & \mathbf{d}\Big(\mathbf{Q}^\dagger\big(\nu\mathbf{I}+\tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\big)^{-1}\mathbf{Q}\Big)\le\boldsymbol{\rho}. \end{array} \tag{3.108}$$

From majorization theory (see Lemma 2.4 and Corollary 2.3), we know that, for a given $\tilde{\mathbf{P}}$, we can always find a $\mathbf{Q}$ satisfying the QoS constraints if and only if

$$\boldsymbol{\lambda}\Big(\big(\nu\mathbf{I}+\tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\big)^{-1}\Big)\succ^w\boldsymbol{\rho}, \tag{3.109}$$

where $\succ^w$ denotes the weakly majorization relation (see Definition 2.3)[16] and $\boldsymbol{\lambda}(\cdot)$ denotes a vector with the eigenvalues of a matrix. Therefore, we can use the weakly majorization relation in lieu of the QoS constraints:

$$\begin{array}{ll} \underset{\tilde{\mathbf{P}}}{\text{minimize}} & \operatorname{Tr}\big(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^\dagger\big) \\ \text{subject to} & \tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}} \quad \text{diagonal (increasing diag. elements)} \\ & \mathbf{d}\Big(\big(\nu\mathbf{I}+\tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\big)^{-1}\Big)\succ^w\boldsymbol{\rho}. \end{array} \tag{3.110}$$

At this point, given that matrix $\tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}$ is diagonal with diagonal elements in increasing order, we can invoke Lemma 3.16 in Appendix 3.B (as in the proof of Theorem 3.1) to conclude that the optimal $\tilde{\mathbf{P}}$ can be written as $\tilde{\mathbf{P}}=\mathbf{V}_H\boldsymbol{\Sigma}$, where $\mathbf{V}_H$ and $\boldsymbol{\Sigma}=\operatorname{diag}\big(\sqrt{\mathbf{p}}\big)$ are defined in the theorem statement. The simple convex reformulation in (3.106) follows then by using the structure $\tilde{\mathbf{P}}=\mathbf{V}_H\boldsymbol{\Sigma}$ and by rewriting the weakly majorization relation explicitly as in Definition 2.3 (observe that both the elements of $\boldsymbol{\rho}$ and of $\mathbf{d}\Big(\big(\nu\mathbf{I}+\tilde{\mathbf{P}}^\dagger\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{P}}\big)^{-1}\Big)$ are already in decreasing order).

As argued in the proof of Theorem 3.1, it is important to point out that, even though we have assumed the elements of the diagonal

[16] The weakly majorization relation $\mathbf{y}\succ^w\mathbf{x}$ is defined as $\sum_{j=i}^n y_j\le\sum_{j=i}^n x_j$ for $1\le i\le n$, where the elements of $\mathbf{y}$ and $\mathbf{x}$ are assumed in decreasing order.

matrix $\tilde{\mathbf{P}}^\dagger \mathbf{H}^\dagger \mathbf{H} \tilde{\mathbf{P}}$ in increasing order, it is not necessary to explicitly include the constraints $p_i \lambda_{H,i} \leq p_{i+1} \lambda_{H,i+1}$, for $1 \leq i \leq L-1$, in the simplified problem (3.106) as the optimal solution to (3.106) already satisfies those constraints. □

The optimal solution to the simplified problem in (3.106) is given (from the KKT optimality conditions) by the waterfilling solution:

$$p_i = \left(\mu_i \lambda_{H,i}^{-1/2} - \nu \lambda_{H,i}^{-1}\right)^+ \quad 1 \leq i \leq L, \tag{3.111}$$

where the multiple waterlevels μ_i's are chosen to satisfy:

$$\begin{aligned}
&\textstyle\sum_{j=i}^{L} \frac{1}{\nu + p_j \lambda_{H,j}} \leq \sum_{j=i}^{L} \rho_j \quad 1 < i \leq L \\
&\textstyle\sum_{j=1}^{L} \frac{1}{\nu + p_j \lambda_{H,j}} = \sum_{j=1}^{L} \rho_j \\
&\mu_i \geq \mu_{i-1} \quad (\mu_0 \triangleq 0) \\
&(\mu_i - \mu_{i-1}) \left(\textstyle\sum_{j=i}^{L} \frac{1}{\nu + p_j \lambda_{H,j}} - \sum_{j=i}^{L} \rho_j\right) = 0.
\end{aligned} \tag{3.112}$$

Evaluating numerically this waterfilling solution is not as simple as for the previous waterfillings with a single waterlevel, but it can still be done efficiently as shown in [114] (see Algorithm 3.4 in Appendix 3.E for a practical implementation).

3.5.3 Characterization of the Optimum Value as a Function

It will be convenient for Section 3.6.1 as well as Section 3.7 to characterize the optimum value of problem (3.94) as a function of the QoS constraints.

In particular, consider the following more general problem:

$$\begin{array}{ll}
\underset{\mathbf{P}}{\text{minimize}} & \operatorname{Tr}\left(\mathbf{P}\mathbf{P}^\dagger\right) \\
\text{subject to} & g_i\left(\left[\left(\nu \mathbf{I} + \mathbf{P}^\dagger \mathbf{H}^\dagger \mathbf{H} \mathbf{P}\right)^{-1}\right]_{ii}\right) \leq b_i \qquad 1 \leq i \leq L,
\end{array} \tag{3.113}$$

where g_i is an invertible increasing function (it is assumed w.l.o.g. that $g_i^{-1}(b_i) \geq g_{i+1}^{-1}(b_{i+1})$), e.g., the relation between the BER and the MSE as given in (3.34). Observe that this problem is actually equivalent to (3.94) simply by rewriting the QoS constraints as

$$\left[\left(\nu \mathbf{I} + \mathbf{P}^\dagger \mathbf{H}^\dagger \mathbf{H} \mathbf{P}\right)^{-1}\right]_{ii} \leq g_i^{-1}(b_i) \triangleq \rho_i. \tag{3.114}$$

We can then invoke Theorem 3.11 to write the following equivalent simple convex problem:

$$\begin{array}{lll} \underset{\mathbf{p}}{\text{minimize}} & \mathbf{1}^T\mathbf{p} & \\ \text{subject to} & \sum_{j=i}^{L} \frac{1}{\nu + p_j \lambda_{H,j}} \leq \sum_{j=i}^{L} g_i^{-1}(b_i) \quad 1 \leq i \leq L \\ & \mathbf{p} \geq \mathbf{0}, & \end{array} \tag{3.115}$$

whose optimal value its characterized next.

Proposition 3.12. [110] Let $P^{\star}(\mathbf{b})$ denote the minimum cost value of problem (3.115) (or (3.113)) as a function of the QoS requirements $\mathbf{b}$ (for any ordering of the $g_i^{-1}(b_i)$'s and $\lambda_{H,i}$'s). Then, if the g_i's are convex and increasing:

(a) The function $P^{\star}(\mathbf{b})$ is convex.
(b) A subgradient[17] of $P^{\star}(\mathbf{b})$ at some feasible point $\mathbf{b}$ is given by $\mathbf{s}(\mathbf{b})$ with components $s_i(\mathbf{b}) = -\mu_i^2(\mathbf{b})\left(g_i^{-1}\right)'(b_i)$, where $\mu_i(\mathbf{b})$ is the waterlevel in the solution (3.111) to problem (3.115).
(c) The function $P^{\star}(\mathbf{b})$ is differentiable on the set of $\mathbf{b}$ corresponding to the region of strict convexity of the g_i's.

Proof. (a) follows easily from the fact that $\phi(\mathbf{x}) = \inf_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y})$ is convex if $\phi(\mathbf{x}, \mathbf{y})$ is jointly convex [20, Sec. 3.2.5],[12, Sec. 5.4.4]. Indeed, define the function $f(\mathbf{p}, \mathbf{b}) = \mathbf{1}^T\mathbf{p}$ if the constraints in (3.115) are satisfied and $+\infty$ otherwise, which is a convex function as the functions g_i^{-1} are concave. Then $P^{\star}(\mathbf{b}) = \inf_{\mathbf{p}} f(\mathbf{p}, \mathbf{b})$ is convex.

(b) A subgradient of $P^{\star}(\boldsymbol{\rho})$, as a function of $\rho_i = g_i^{-1}(b_i)$, is given by the optimal Lagrange multipliers λ_i of problem (3.115) for the given value of $\boldsymbol{\rho}$ [12, Sec. 5.4.4]: $\mathbf{s}(\boldsymbol{\rho}) = -\tilde{\boldsymbol{\lambda}}(\boldsymbol{\rho})$, where $\tilde{\lambda}_i = \sum_{j=1}^{i} \lambda_j$, as it satisfies the condition:

$$P^{\star}(\boldsymbol{\rho}) \geq P^{\star}(\boldsymbol{\rho}_0) - \tilde{\boldsymbol{\lambda}}^T(\boldsymbol{\rho}_0)(\boldsymbol{\rho} - \boldsymbol{\rho}_0) \quad \forall \boldsymbol{\rho},$$

[17] For a convex (concave) function f, a subgradient at point $\mathbf{x}_0$ is defined as any vector $\mathbf{s}$ that satisfies $f(\mathbf{x}) \geq (\leq) f(\mathbf{x}_0) + \mathbf{s}^T(\mathbf{x} - \mathbf{x}_0)$ for all $\mathbf{x}$ (cf. Section A.7).

which, combined with the concavity of $g_i^{-1}(b_i)$, gives

$$P^{\star}(\mathbf{b}) \geq P^{\star}(\mathbf{b}_0) + \mathbf{s}^T(\mathbf{b}_0)(\mathbf{b} - \mathbf{b}_0) \quad \forall \mathbf{b},$$

where $s_i(\mathbf{b}_0) = -\tilde{\lambda}_i(\mathbf{b}_0)\left(g_i^{-1}\right)'(b_{0,i})$. To complete the proof, note that the waterlevels in the solution (3.111) are given by $\mu_i = \tilde{\lambda}_i^{1/2}$.

(c) This is technically involved and the interested reader is referred to [110]. □

3.5.4 Numerical Results

Figure 3.7 gives the required power at the transmitter versus the different QoS constraints in terms of BER, for a random 4 × 4 channel matrix $\mathbf{H}$ with i.i.d. elements Gaussian distributed with zero mean and unit variance. In particular, the diagonal suboptimal solution is

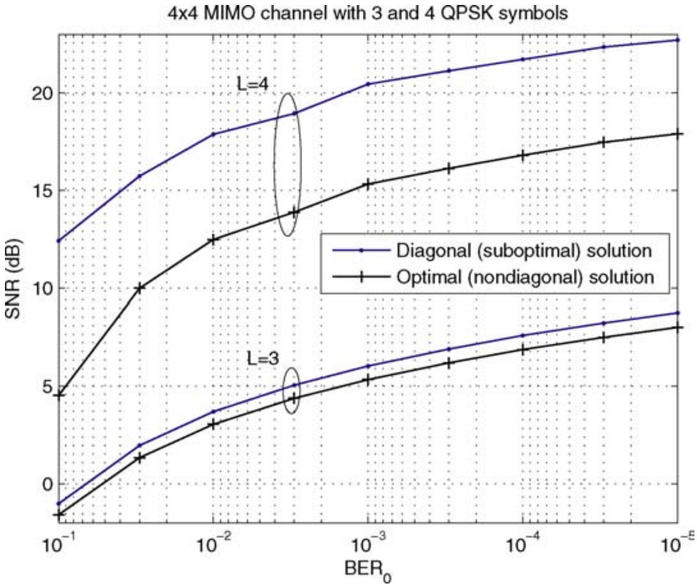


Fig. 3.7 Power versus the different QoS constraints in terms of BER for a 4 × 4 MIMO channel with $L = 3$ and $L = 4$ transmitted QPSK symbols. (The BER of the first substream is along the x-axis, and the BER of the second, third, and possibly fourth substreams are given by scaling with the factors 0.5, 0.5, and 0.1).

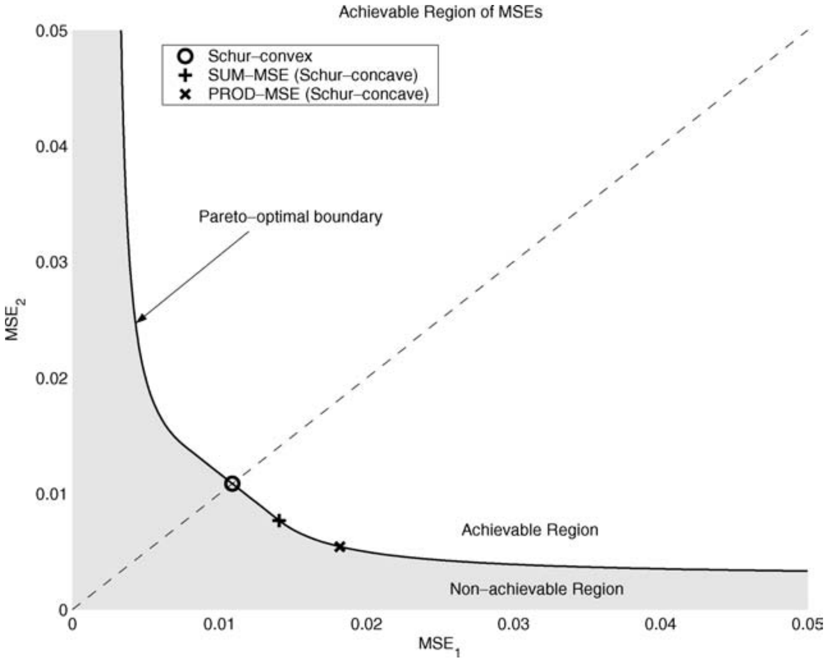


Fig. 3.8 Achievable region of the MSEs for a 4 × 4 MIMO channel with $L = 2$ and SNR = 15 dB, along with the location of the design based on two Schur-concave functions (PROD-MSE and SUM-MSE) and two Schur-convex functions (MAX-MSE and SUM-BER).

compared with the optimal nondiagonal solution with a difference of 1–2 dB for $L = 3$ and of 5 dB for $L = 4$ (fully loaded system).

Figure 3.8 shows the achievable region in terms of MSEs for a given realization of a 4 × 4 MIMO channel with $L = 2$ and SNR = 15 dB. The achievable region was computed by specifying QoS constraints and then checking whether the minimum required power was above or below the threshold. The boundary between the achievable and non-achievable regions corresponds to the Pareto-optimal designs, characterized by not being outperformed by any other solution simultaneously in all substreams. The solutions corresponding to designs based on global performance measures lie on the Pareto-optimal boundary, although each in a different point. In particular, the solutions for two Schur-concave functions (PROD-MSE and SUM-MSE) and two Schur-convex functions (MAX-MSE and SUM-BER) are also indicated. Observe that, since

Schur-convex methods have equal MSEs on all substreams, they all correspond to the intersection of the Pareto-optimal boundary with the line $\mathrm{MSE}_1 = \mathrm{MSE}_2$, which corresponds to a complete fairness among substreams.

3.6 Optimum Linear Transmitter with Global Measure of Performance: Arbitrary Cost Functions

Section 3.4 dealt with the problem formulation with a global measure of performance based on the family of Schur-concave and Schur-convex functions. However, even though this family of functions embraces most interesting functions to design systems (cf. Sections 3.4.2–3.4.3 and Tables 3.1 and 3.2), it does not include all functions as illustrated in Figure 2.1. In particular, if the design is based on the minimization of average BER of the substreams, the corresponding cost function is neither Schur-concave nor Schur-convex (only if the constellations used on the substreams are equal, the function becomes Schur-convex as shown in Section 3.4.3.4). It is thus necessary to generalize the framework provided by Theorem 3.1 to arbitrary cost functions.

This section considers then problem formulations with an *arbitrary* global measure of performance (not necessarily Schur-concave or Schur-convex), either as the minimization of a global cost function subject to a power constraint as in (3.13) or the minimization of the power subject to a constraint on the global performance as in (3.14).

The optimal receiver $\mathbf{W}$ has already been obtained in Section 3.3 (see (3.29)) as the Wiener filter or MMSE receiver, and also as the ZF receiver under the ZF constraint. The MSE matrix is then given by (3.31) and the MSEs by (3.32): $\mathrm{MSE}_i = \left[\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}\right]_{ii}$, where $\nu = 0$ for the ZF receiver and $\nu = 1$ for the MMSE receiver. Therefore, the problem of minimizing a cost function of the MSEs as a function of the linear precoder $\mathbf{P}$ at the transmitter can be formulated as (recall that cost functions of the SINRs and BERs can always be reformulated as functions of the MSEs):

$$\begin{array}{ll} \underset{\mathbf{P}}{\text{minimize}} & f_0\left(\left\{\left[\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}\right]_{ii}\right\}\right) \\ \text{subject to} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^\dagger\right) \le P_0. \end{array} \tag{3.116}$$

The following result provides an equivalent simple reformulation of (3.116) based on majorization theory.[18]

Theorem 3.13. The optimal solution $\mathbf{P}$ to the complicated nonconvex matrix-valued problem in (3.116), where $f_0 : \mathbb{R}^L \to \mathbb{R}$ is a function increasing in each variable and minimized when the arguments are sorted in decreasing order (or, similarly, in some fixed ordering),[19] is given by

$$\mathbf{P} = \mathbf{V}_H \,\mathrm{diag}\,(\sqrt{\mathbf{p}})\,\mathbf{Q}, \tag{3.117}$$

where $\mathbf{V}_H$ is a (semi-)unitary matrix with columns equal to the right singular vectors of matrix $\mathbf{H}$ corresponding to the L largest singular values in increasing order, $\mathbf{Q}$ is a unitary matrix such that $\left[\left(\nu\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P}\right)^{-1}\right]_{ii} = \rho_i$ for all i (which can be computed with Algorithm 2.2), and the power allocation $\mathbf{p}$ is the solution to the following simple problem:

$$\begin{array}{ll} \underset{\mathbf{p},\boldsymbol{\rho}}{\text{minimize}} & f_0\left(\rho_1, \ldots, \rho_L\right) \\ \text{subject to} & \sum_{j=i}^{L} \frac{1}{\nu + p_j\, \lambda_{H,j}} \leq \sum_{j=i}^{L} \rho_j \quad 1 \leq i \leq L \\ & \rho_i \geq \rho_{i+1} \\ & \mathbf{1}^T\mathbf{p} \leq P_0 \\ & \mathbf{p} \geq \mathbf{0}, \end{array} \tag{3.118}$$

where the $\lambda_{H,i}$'s denote the L largest eigenvalues of matrix $\mathbf{H}^\dagger\mathbf{H}$ in increasing order.

Furthermore, if f_0 is a convex function, problem (3.118) is convex and the ordering constraint $\rho_i \geq \rho_{i+1}$ can be removed.

Proof. The proof follows closely that of Theorem 3.11 and hinges on majorization theory (cf. Chapter 2).

[18] Recall that for the sake of notation we assume $L \leq \min\left(n_R, n_T\right)$. Otherwise, Theorem 3.13 must be slightly modified similarly to Theorem 3.11 (as indicated in footnote 15).

[19] In fact, if the cost function f_0 is not minimized when the arguments are sorted in some specific ordering, the simplified problem in (3.118) can still be used removing the constraint $\rho_i \geq \rho_{i+1}$ and replacing the term $\sum_{j=i}^{L} \rho_j$ by $\sum_{j=i}^{L} \rho_{[j]}$ as described in (3.124).

Start by rewriting the original problem (3.116) as

$$\begin{array}{ll} \underset{\mathbf{P},\boldsymbol{\rho}}{\text{minimize}} & f_0(\boldsymbol{\rho}) \\ \text{subject to} & \mathbf{d}\left(\left(\nu\mathbf{I}+\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}\right)^{-1}\right) \leq \boldsymbol{\rho} \\ & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \leq P_0, \end{array} \tag{3.119}$$

where $\mathbf{d}(\cdot)$ denotes a vector with the diagonal elements of a matrix.

Now, for a given $\mathbf{P}$, we can choose a unitary matrix $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}\mathbf{Q}^{\dagger}$ is diagonal (with diagonal elements in increasing order). Then, matrix $\left(\nu\mathbf{I}+\tilde{\mathbf{P}}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}$, where $\tilde{\mathbf{P}}=\mathbf{P}\mathbf{Q}^{\dagger}$, is diagonal and the original MSE matrix is given by $\left(\nu\mathbf{I}+\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}\right)^{-1}=\mathbf{Q}^{\dagger}\left(\nu\mathbf{I}+\tilde{\mathbf{P}}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}\mathbf{Q}$. We can then rewrite the problem in terms of $\tilde{\mathbf{P}}$ and $\mathbf{Q}$ as

$$\begin{array}{ll} \underset{\tilde{\mathbf{P}},\mathbf{Q},\boldsymbol{\rho}}{\text{minimize}} & f_0(\boldsymbol{\rho}) \\ \text{subject to} & \tilde{\mathbf{P}}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\tilde{\mathbf{P}} \quad \text{diagonal (increasing diag. elements)} \\ & \mathbf{d}\left(\mathbf{Q}^{\dagger}\left(\nu\mathbf{I}+\tilde{\mathbf{P}}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}\mathbf{Q}\right) \leq \boldsymbol{\rho} \\ & \mathrm{Tr}\left(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^{\dagger}\right) \leq P_0. \end{array} \tag{3.120}$$

From majorization theory (see Lemma 2.4 and Corollary 2.3), we know that, for a given $\tilde{\mathbf{P}}$, we can always find a $\mathbf{Q}$ satisfying the QoS constraints if and only if

$$\boldsymbol{\lambda}\left(\left(\nu\mathbf{I}+\tilde{\mathbf{P}}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}\right) \succ^{w} \boldsymbol{\rho}, \tag{3.121}$$

where $\succ^{w}$ denotes the weakly majorization relation (see Definition 2.3)[20] and $\boldsymbol{\lambda}(\cdot)$ denotes a vector with the eigenvalues of a matrix. Therefore, we can use the weakly majorization relation in lieu of the QoS constraints:

$$\begin{array}{ll} \underset{\tilde{\mathbf{P}},\boldsymbol{\rho}}{\text{minimize}} & f_0(\boldsymbol{\rho}) \\ \text{subject to} & \tilde{\mathbf{P}}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\tilde{\mathbf{P}} \quad \text{diagonal (increasing diag. elements)} \\ & \mathbf{d}\left(\left(\nu\mathbf{I}+\tilde{\mathbf{P}}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\tilde{\mathbf{P}}\right)^{-1}\right) \succ^{w} \boldsymbol{\rho} \\ & \mathrm{Tr}\left(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^{\dagger}\right) \leq P_0. \end{array} \tag{3.122}$$

[20] The weakly majorization relation $\mathbf{y} \succ^{w} \mathbf{x}$ is defined as $\sum_{j=i}^{n} y_j \leq \sum_{j=i}^{n} x_j$ for $1 \leq i \leq n$, where the elements of $\mathbf{y}$ and $\mathbf{x}$ are assumed in decreasing order.

At this point, given that matrix $\tilde{\mathbf{P}}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\tilde{\mathbf{P}}$ is diagonal with diagonal elements in increasing order, we can invoke Lemma 3.16 in Appendix 3.B (as in the proof of Theorem 3.1) to conclude that the optimal $\mathbf{P}$ can be written as $\mathbf{P} = \mathbf{V}_H \mathbf{\Sigma}$, where $\mathbf{V}_H$ and $\mathbf{\Sigma} = \text{diag}\left(\sqrt{\mathbf{p}}\right)$ are defined in the theorem statement. Now, by using the structure $\mathbf{P} = \mathbf{V}_H \mathbf{\Sigma}$ and by rewriting the weakly majorization relation explicitly as in Definition 2.3, we get

$$\begin{array}{ll} \underset{\mathbf{p},\boldsymbol{\rho}}{\text{minimize}} & f_0\left(\boldsymbol{\rho}\right) \\ \text{subject to} & \sum_{j=i}^{L} \frac{1}{\nu + p_j \lambda_{H,j}} \leq \sum_{j=i}^{L} \rho_{[j]} \quad 1 \leq i \leq L \\ & \mathbf{1}^T \mathbf{p} \leq P_0 \\ & \mathbf{p} \geq \mathbf{0}, \end{array} \tag{3.123}$$

where $\rho_{[i]}$ denotes the elements of $\boldsymbol{\rho}$ in decreasing order and, as argued in the proof of Theorem 3.1, it is not necessary to explicitly include the constraints $p_i \lambda_{H,i} \leq p_{i+1} \lambda_{H,i+1}$ as the optimal solution already satisfies those constraints.

To deal with the $\rho_{[j]}$'s, we can use the relation [20]:

$$\sum_{j=i}^{L} \rho_{[j]} = \min\left\{\rho_{j_1} + \cdots + \rho_{j_{L-i+1}} \mid 1 \leq j_1 < \cdots < j_{L-i+1} \leq L\right\}, \tag{3.124}$$

which is clearly a concave function since it is the pointwise minimum of concave (affine) functions. This problem formulation is always valid. However, to simplify it without the need for (3.124), we can now use the assumption that the cost function $f_0\left(\boldsymbol{\rho}\right)$ is minimized when the arguments are in decreasing order to include the constraint $\rho_i \geq \rho_{i+1}$; the problem can then be further simplified by replacing the term $\sum_{j=i}^{L} \rho_{[j]}$ by $\sum_{j=i}^{L} \rho_j$, obtaining finally the simplified problem in (3.118).

To conclude, it turns out that if f_0 is convex, the constraint $\rho_i \geq \rho_{i+1}$ can be removed as any optimal solution satisfies it anyway as we now show. Suppose that $\rho_i < \rho_{i+1}$, then we can use instead $\tilde{\rho}_i = \tilde{\rho}_{i+1} = \left(\rho_i + \rho_{i+1}\right)/2$ with two consequences: (i) a lower (or at most equal) value of the cost function (to see this simply swap ρ_i and ρ_{i+1} to obtain a lower value because the function is minimized with arguments in decreasing order and then take any point on the line between $\left(\rho_i, \rho_{i+1}\right)$

and (ρ_{i+1}, ρ_i) using convexity of the function) and (ii) no effect on the other constraints as shown next.

Define

$$\alpha_i \triangleq \frac{1}{\nu + p_i \lambda_{H,i}}, \tag{3.125}$$

$$c_i \triangleq \sum_{j=i+2}^{L} \rho_j - \sum_{j=i+2}^{L} \frac{1}{\nu + p_j \lambda_{H,j}} \geq 0 \tag{3.126}$$

(recall that at an optimal point we must have $\alpha_i \geq \alpha_{i+1}$). Then, we can write the constraint $\sum_{j=i+1}^{L} \frac{1}{\nu + p_j \lambda_{H,j}} \leq \sum_{j=i+1}^{L} \rho_j$ as $\alpha_{i+1} \leq \rho_{i+1} + c_i$. We can always focus on the case with equality $\alpha_{i+1} = \rho_{i+1} + c_i$; otherwise, we could decrease ρ_{i+1} and increase ρ_i by the same amount. Then, we can rewrite the $(i+1)$th and ith constraints as

$$\alpha_{i+1} = \rho_{i+1} + c_i \geq \rho_{i+1}, \tag{3.127}$$

$$\alpha_i + \alpha_{i+1} \leq \rho_i + \rho_{i+1} + c_i = \rho_i + \alpha_{i+1}, \tag{3.128}$$

from which we can write

$$\alpha_i \leq \rho_i < \rho_{i+1} \leq \alpha_{i+1}, \tag{3.129}$$

which is a contradiction as we know that $\alpha_i \geq \alpha_{i+1}$ at an optimal point.

□

It is important to emphasize that we can consider instead the minimization of the power subject to a global constraint on the performance as in (3.14). This leads to the problem formulation

$$\begin{array}{ll} \underset{\mathbf{P}}{\text{minimize}} & \text{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \\ \text{subject to} & f_0\left(\left\{\left[\left(\nu\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}\right)^{-1}\right]_{ii}\right\}\right) \leq \alpha_0, \end{array} \tag{3.130}$$

whose solution is given by $\mathbf{P} = \mathbf{V}_H \, \text{diag}\left(\sqrt{\mathbf{p}}\right) \mathbf{Q}$, as in Theorem 3.13, and the problem can then be rewritten as

$$\begin{array}{ll} \underset{\mathbf{p},\boldsymbol{\rho}}{\text{minimize}} & \mathbf{1}^T\mathbf{p} \\ \text{subject to} & \sum_{j=i}^{L} \frac{1}{\nu + p_j \lambda_{H,j}} \leq \sum_{j=i}^{L} \rho_j \quad 1 \leq i \leq L \\ & \rho_i \geq \rho_{i+1} \\ & f_0\left(\rho_1, \ldots, \rho_L\right) \leq \alpha_0 \\ & \mathbf{p} \geq \mathbf{0}. \end{array} \tag{3.131}$$

Even more interestingly, we can easily restate the result in a more general framework by considering a cost function in terms of the alternative variables $b_i = g_i(\rho_i)$, where g_i is an invertible convex increasing function, e.g., the relation between the BER and the MSE as given in (3.34). For example, the problem formulation (3.130) in terms of b_i would be

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{b}}{\text{minimize}} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \\ \text{subject to} & g_i\left(\left[\left(\nu\mathbf{I}+\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}\right)^{-1}\right]_{ii}\right) \leq b_i \quad 1 \leq i \leq L \\ & f_0\left(b_1,\ldots,b_L\right) \leq \alpha_0. \end{array} \tag{3.132}$$

We can rewrite this problem in terms of the ρ_i's as

$$\begin{array}{ll} \underset{\mathbf{P},\boldsymbol{\rho}}{\text{minimize}} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \\ \text{subject to} & \left[\left(\nu\mathbf{I}+\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}\right)^{-1}\right]_{ii} \leq g_i^{-1}\left(b_i\right) \triangleq \rho_i \quad 1 \leq i \leq L \\ & f_0\left(g_1\left(\rho_1\right),\ldots,g_L\left(\rho_L\right)\right) \leq \alpha_0. \end{array} \tag{3.133}$$

Observe that if f_0 is increasing/convex then the new equivalent function $\tilde{f}_0\left(\rho_1,\ldots,\rho_L\right) = f_0\left(g_1\left(\rho_1\right),\ldots,g_L\left(\rho_L\right)\right)$ will be increasing/convex as well. At this point, if the function $\tilde{f}_0$ is minimized when the arguments are in decreasing order, then we can proceed as in Theorem 3.13 to finally rewrite the problem in a simplified way as

$$\begin{array}{ll} \underset{\mathbf{p},\boldsymbol{\rho}}{\text{minimize}} & \mathbf{1}^T\mathbf{p} \\ \text{subject to} & \sum_{j=i}^{L} \frac{1}{\nu+p_j\lambda_{H,j}} \leq \sum_{j=i}^{L} \rho_j \quad 1 \leq i \leq L \\ & \rho_i \geq \rho_{i+1} \\ & f_0\left(g_1\left(\rho_1\right),\ldots,g_L\left(\rho_L\right)\right) \leq \alpha_0 \\ & \mathbf{p} \geq \mathbf{0}. \end{array} \tag{3.134}$$

Finally, if f_0 is convex so will $\tilde{f}_0$, the constraint $\rho_i \geq \rho_{i+1}$ can be removed and the problem can be rewritten in terms of the ρ_i's as

$$\begin{array}{ll} \underset{\mathbf{p},\boldsymbol{\rho}}{\text{minimize}} & \mathbf{1}^T\mathbf{p} \\ \text{subject to} & \sum_{j=i}^{L} \frac{1}{\nu+p_j\lambda_{H,j}} \leq \sum_{j=i}^{L} g_i^{-1}\left(b_i\right) \quad 1 \leq i \leq L \\ & f_0\left(b_1,\ldots,b_L\right) \leq \alpha_0 \\ & \mathbf{p} \geq \mathbf{0}. \end{array} \tag{3.135}$$

3.6.1 Efficient Resolution via Primal Decomposition

Theorem 3.13 says that solving the complicated nonconvex problem (3.116) is equivalent to solving the simple problem (3.118) (similarly, problem (3.130) is equivalent to (3.131)). Problem (3.118) is indeed a simple problem to solve, especially when f_0 is convex, as the problem becomes then convex and then we can use any general-purpose method to solve it (cf. Appendix A). In the following, we give a very simple and efficient numerical method to solve problem (3.118) based on a primal decomposition technique (cf. Appendix A) as developed in [110].

The main idea is to decompose the joint minimization in (3.118) and (3.131) into two nested minimizations [20, Sec. 4.1.3] (see also [12, Sec. 6.4.2]):

$$\inf_{\mathbf{p},\boldsymbol{\rho}} \phi(\mathbf{p},\boldsymbol{\rho}) = \inf_{\boldsymbol{\rho}} \left(\inf_{\mathbf{p}} \phi(\mathbf{p},\boldsymbol{\rho}) \right). \tag{3.136}$$

The inner minimization over $\mathbf{p}$ for a fixed $\boldsymbol{\rho}$ applied to problem (3.131) is precisely problem (3.94) considered in Section 3.5 whose solution is the waterfilling in (3.111) for which we have a simple algorithm (Algorithm 3.4 in Appendix 3.E). In fact, the result of this inner minimization was characterized in Proposition 3.12, Section 3.5.3, and denoted by $P^\star(\boldsymbol{\rho})$. Thus, we can rewrite problems (3.130) and (3.131) as

$$\begin{array}{ll} \underset{\boldsymbol{\rho}}{\text{minimize}} & P^\star(\boldsymbol{\rho}) \\ \text{subject to} & \rho_i \geq \rho_{i+1} \\ & f_0(\rho_1,\ldots,\rho_L) \leq \alpha_0. \end{array} \tag{3.137}$$

Similarly, we can rewrite problems (3.116) and (3.118) as

$$\begin{array}{ll} \underset{\boldsymbol{\rho}}{\text{minimize}} & f_0(\rho_1,\ldots,\rho_L) \\ \text{subject to} & \rho_i \geq \rho_{i+1} \\ & P^\star(\boldsymbol{\rho}) \leq P_0. \end{array} \tag{3.138}$$

Problems (3.137) and (3.138) are known as *master problems* in the literature of decomposition techniques [12, 88, 142] (cf. Appendix A).

It is important to remark that if the cost function f_0 is defined instead as a function of the alternative variables $b_i = g_i(\rho_i)$, the same approach can be employed by using the function $P^\star(\mathbf{b})$ as characterized in Proposition 3.12, Section 3.5.3.

Now, solving the master problems (3.137) and (3.138) can be easily done in practice with a gradient/subgradient projection method (cf. Appendix A) since the subgradient of $P^\star(\boldsymbol{\rho})$ is well characterized in Proposition 3.12. This approach is illustrated in the next section for a very relevant example.

3.6.2 An Illustrative Example: Minimum BER Design

The design minimization of the BER averaged over the substreams is perhaps the most meaningful criterion. If the constellations used on the substreams are equal, then the cost function (average BER) turns out to be a Schur-convex function (cf. Section 3.4.3.4) and Theorem 3.1 can be invoked to easily find the optimal solution to the problem. However, if the constellations are different, then the function is not Schur-convex anymore and need to invoke the more general result for arbitrary cost functions in Theorem 3.13. This problem was studied in detail in [110].

Consider the following problem formulation:

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W}}{\text{minimize}} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^\dagger\right) \\ \text{subject to} & \frac{1}{L}\sum_{i=1}^{L} g_i(\mathrm{MSE}_i) \le \mathsf{ber}_0, \end{array} \tag{3.139}$$

where ber_0 is the desired averaged BER and each g_i relates the MSE to the BER on the ith substream (as defined in (3.34)):

$$\mathrm{BER}_i = g_i(\mathrm{MSE}_i) \quad 1 \le i \le L. \tag{3.140}$$

Recall that the function g_i depends on the particular constellation employed (cf. Section 3.3.2 and Appendix 3.A for a discussion and detailed characterization for QAM constellations).

In the sequel, we will use the following properties of the BER function g for a given constellation $\mathcal{C}$ (with cardinality denoted by $|\mathcal{C}|$):

P1 g is strictly increasing and $g(0) = 0$.

P2 g is strictly convex on the interval $[0, u]$ [21](for mathematical convenience g is defined $+\infty$ elsewhere[22]). In addition, $u_1 \geq u_2$ for $|\mathcal{C}_1| \leq |\mathcal{C}_2|$.

P3 $g_1(\rho) < g_2(\rho)$ and $g'_1(\rho) < g'_2(\rho)$ for $|\mathcal{C}_1| < |\mathcal{C}_2|$.

It is important to remark that the properties P1–P3 are indeed very natural for any reasonable family of constellations (see Figure 3.3). The increasingness of g is clear since a higher MSE must always be worse than a lower MSE. The convexity of g is a natural result for the range in which the MSE as a function of the SINR (which is a convex function) is approximately linear (this follows since we expect a system to have a BER at some SINR_0 smaller than (or at least equal to) the average BER that would be achieved by a time-division approach using two different SINRs satisfying $(\text{SINR}_1 + \text{SINR}_2)/2 = \text{SINR}_0$). Clearly, $g_1(\rho) < g_2(\rho)$ must be satisfied for $|\mathcal{C}_1| < |\mathcal{C}_2|$; otherwise, we could transmit more bits at a lower BER, which does not make any sense. Also, $g'_1(\rho) < g'_2(\rho)$ is a natural result since it simply means that larger constellations (normalized with unit energy) are expected to have a higher sensitivity with respect to changes in the MSE, which is an expected result because higher constellations have a smaller minimum distance.

The subsequent analytical characterization of the minimum BER problem is valid only for systems that work in the convex region, i.e., with a sufficiently small MSE at each established substream. This is a mild restriction because, if the gain of a substream is too low, it may be better not to use it at all and to decrease the total number of transmitted symbols. Nevertheless, it is worth mentioning that, in practice, the method proposed in this section also works well in the nonconvex region.

The following result will prove extremely useful in the sequel.

Lemma 3.14. [110, Lem. 1] Let g_1 and g_2 be two BER functions corresponding to the constellations $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively, with

[21] The value of u can be obtained by analyzing the convexity properties of the function g (see, for example, (3.37) and (3.38) for QAM constellations).

[22] Property P1 is then satisfied only on the interval $[0, u]$.

$|\mathcal{C}_1| \leq |\mathcal{C}_2|$ and satisfying properties P1–P3. Then, for $\rho_1 < \rho_2$,

$$g_1(\rho_2) + g_2(\rho_1) \leq g_1(\rho_1) + g_2(\rho_2) \tag{3.141}$$

$$g_1((\rho_1 + \rho_2)/2) + g_2((\rho_1 + \rho_2)/2) < g_1(\rho_1) + g_2(\rho_2). \tag{3.142}$$

In addition, if $|\mathcal{C}_1| < |\mathcal{C}_2|$, then (3.141) is satisfied with strict inequality (provided that $g_1(\rho_2) + g_2(\rho_1) < +\infty$).

As a consequence of Lemma 3.14, if the constellations employed on the L substreams are ordered in increasing cardinality, i.e., $|\mathcal{C}_1| \leq |\mathcal{C}_2| \leq \cdots \leq |\mathcal{C}_L|$, then the function $\sum_{i=1}^{L} g_i(\mathrm{MSE}_i)$ is minimized if the arguments are in decreasing order: $\mathrm{MSE}_i \geq \mathrm{MSE}_2 \geq \cdots \geq \mathrm{MSE}_L$. We can then invoke Theorem 3.13. In particular, since problem (3.132) simplifies to (3.135), it follows that the minimum BER problem in (3.139) simplifies to (denoting $b_i = g_i(\mathrm{MSE}_i)$)

$$\begin{array}{ll} \underset{\mathbf{p},\boldsymbol{\rho}}{\text{minimize}} & \mathbf{1}^T\mathbf{p} \\ \text{subject to} & \sum_{j=i}^{L} \frac{1}{\nu + p_j \lambda_{H,j}} \leq \sum_{j=i}^{L} g_i^{-1}(b_i) \quad 1 \leq i \leq L \\ & \mathbf{p} \geq \mathbf{0} \\ & \frac{1}{L}\sum_{i=1}^{L} b_i \leq \mathsf{ber}_0, \end{array} \tag{3.143}$$

which is a convex problem and can be optimally solved.

We can now go one step further and obtain a very efficient numerical algorithm by using the primal decomposition approach described in Section 3.6.1. In particular, we can rewrite problem (3.143) as

$$\begin{array}{ll} \underset{\mathbf{b}}{\text{minimize}} & P^{\star}(\mathbf{b}) \\ \text{subject to} & \mathbf{b} \geq \mathbf{0} \\ & \mathbf{1}^T\mathbf{b} \leq L\,\mathsf{ber}_0, \end{array} \tag{3.144}$$

where the function $P^{\star}(\mathbf{b})$ denotes the minimum power required to satisfy the BER constraints given by $\mathbf{b}$ and is fully characterized in Section 3.5.3. Recall that the evaluation of $P^{\star}(\mathbf{b})$ requires the water-filling solution in (3.111) which can be efficiently evaluated in practice with Algorithm 3.4 in Appendix 3.E (using $\rho_i = g_i^{-1}(b_i)$). (Observe that the additional constraint $\mathbf{b} \geq \mathbf{0}$ is actually redundant as $P^{\star}(\mathbf{b}) = +\infty$ if $\mathbf{b} \ngeq \mathbf{0}$.)

In Figure 3.9, several examples of the function $P^{\star}(\mathbf{b})$ are plotted between two feasible random points $\mathbf{b}_1$ and $\mathbf{b}_2$, from which the strict convexity and the differentiability can be easily observed.

Finally, to solve the master problem in (3.144) we will use a subgradient projection method (cf. Appendix A) since the subgradient of $P^{\star}(\mathbf{b})$ is well characterized in Proposition 3.12:

$$\mathbf{b}(t+1) = \left[\mathbf{b}(t) - \alpha(t)\,\mathbf{s}(t)\right]_{\mathcal{B}}, \tag{3.145}$$

where $s_i(t) = -\mu_i^2(\mathbf{b}(t))\left(g_i^{-1}\right)'(b_i(t))$ and $[\cdot]_{\mathcal{B}}$ denotes projection on the feasible set $\mathcal{B} \triangleq \left\{\mathbf{b} \mid \mathbf{b} \geq \mathbf{0}, \mathbf{1}^T\mathbf{b} \leq L\,\mathsf{ber}_0\right\}$ which is a simplex. Note that the projection is given by a waterfilling solution [109, Lemma 1]:

$$[\mathbf{b}]_{\mathcal{B}} = (\mathbf{b} - \mu)^+, \tag{3.146}$$

where μ as the minimum nonnegative value such that $\mathbf{1}^T\mathbf{b} \leq L\,\mathsf{ber}_0$.

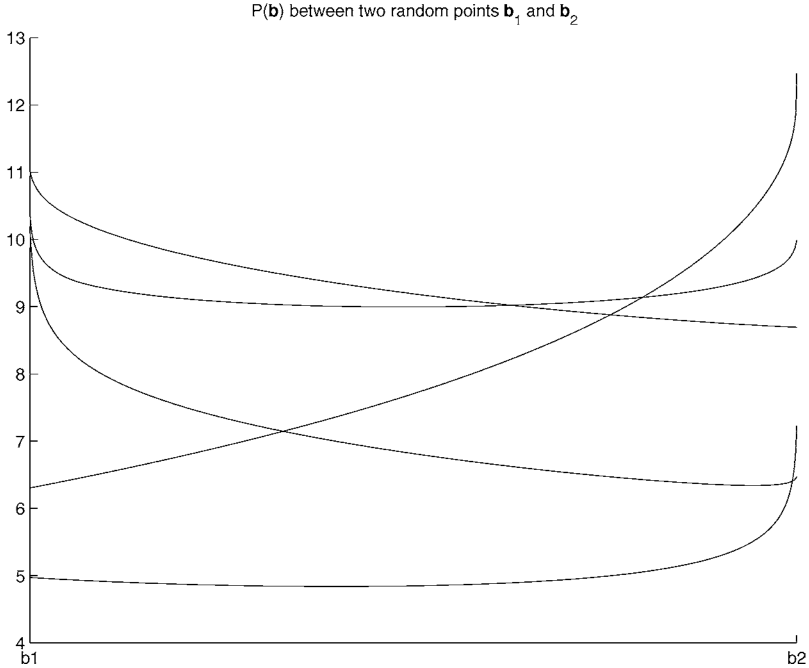


Fig. 3.9 Some examples of the function $P^{\star}(\mathbf{b})$ between two random points $\mathbf{b}_1$ and $\mathbf{b}_2$ (6×6 MIMO channel with $L = 5$ substreams using QPSK and 16-QAM constellations).

3.6.3 Numerical Results

Figure 3.10 gives the required power at the transmitter versus the global BER when transmitting $L = 12$ symbols drawn from different constellations (four QPSK, two 16-QAM, two 64-QAM, two 256-QAM, and two 512-QAM) over a 16×16 MIMO channel, for a random channel matrix $\mathbf{H}$ with i.i.d. elements Gaussian distributed with zero mean and unit variance. Three methods are compared: (i) a benchmark based on imposing a diagonal structure and the same BER on all substreams (called diag. structure + equal BER); (ii) a heuristic solution based on imposing the same BER on all substreams with the optimal design as in Section 3.5 (termed nondiag. structure + equal BER); and (iii) the optimal solution discussed in Section 3.6.2 (called nondiag. structure + nonequal BER). As can be observed, in this particular example the diagonal structure has a mild loss of performance (less than 1 dB).

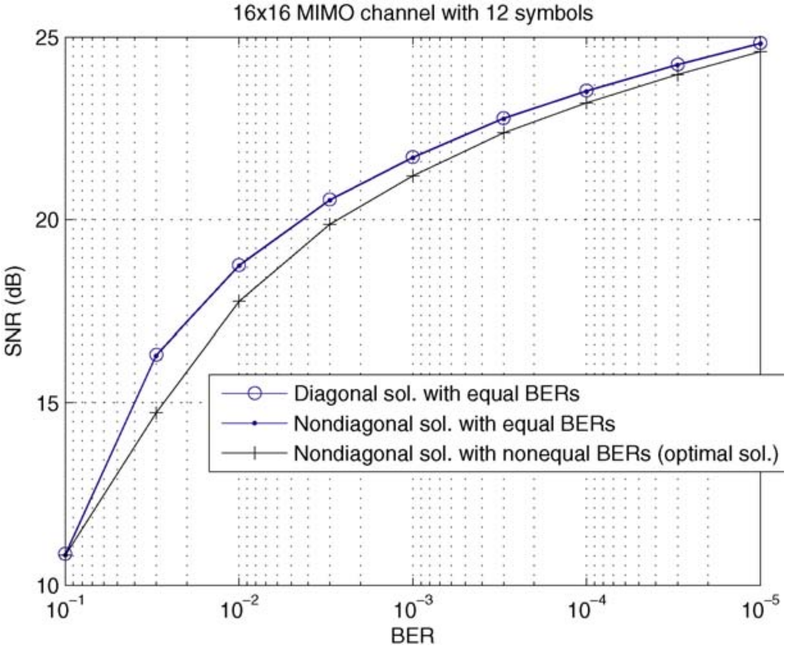


Fig. 3.10 Power versus global BER for a 16×16 MIMO channel with $L = 12$ transmitted symbols (four QPSK, two 16-QAM, two 64-QAM, two 256-QAM, and two 512-QAM).

Interestingly, the performance of the heuristic solution (nondiag. structure + equal BER) is indistinguishable from the optimal performance.

3.7 Extension to Multicarrier Systems

As mentioned in Section 3.1, multicarrier systems (and some other systems) may be more conveniently modeled as a communication through a set of parallel and non-interfering MIMO channels:

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{s}_k + \mathbf{n}_k \quad 1 \leq k \leq N, \tag{3.147}$$

where N is the number of carriers and k is the carrier index, rather than as a single MIMO channel $\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}$ with a block-diagonal equivalent channel matrix $\mathbf{H} = \operatorname{diag}(\{\mathbf{H}_k\})$.

At this point, we should distinguish between two very different scenarios:

(i) *Joint processing of all carriers*: in this case, the system can jointly process all carriers. It suffices to use the block-diagonal channel matrix $\mathbf{H} = \operatorname{diag}(\{\mathbf{H}_k\})$ and proceed as in the case of a single MIMO channel for which the results of this chapter apply.

(ii) *Independent processing of each carrier*: in this case, the system processes each carrier independently. We could still use the signal model with the block-diagonal channel matrix but the transmit and receive matrices would have an implicit block-diagonal structure as well, $\mathbf{P} = \operatorname{diag}(\{\mathbf{P}_k\})$ and $\mathbf{W} = \operatorname{diag}(\{\mathbf{W}_k\})$, which would complicate the design and, in particular, the results of this chapter would not be applicable anymore. In this case, however, it is actually more convenient to use the more explicit signal model in (3.147) with a set of transmit–receive matrices (one for each carrier), $(\mathbf{P}_k, \mathbf{W}_k)$, for which the results of this chapters can be readily applied as shown next.

As in Section 3.1, we can consider a problem formulation based on a global cost function of the MSEs, $f_0\big(\{\text{MSE}_{k,i}\}\big)$, as well as a design based on individual QoS constraints, $\text{MSE}_{k,i} \leq \rho_{k,i}$ for all k and i. If each carrier had an *individual power constraint*, $\operatorname{Tr}\left(\mathbf{P}_k \mathbf{P}_k^\dagger\right) \leq P_k$, then the problem would decouple into N problems, one for each carrier, that could be easily solved as we have seen in this chapter. We will consider instead the more challenging case of having a *global power constraint* $\sum_{k=1}^{N} \mathrm{Tr}\left(\mathbf{P}_k \mathbf{P}_k^\dagger\right) \leq P_0$ which couples all the carriers (individual power constraints can also be easily included).

The problem formulation with a global cost function can be conveniently formulated, similarly to (3.13), as

$$\begin{array}{lll} \underset{\{\mathbf{P}_k, \mathbf{W}_k, \alpha_k, P_k\}}{\text{minimize}} & f_0\left(\alpha_1, \ldots, \alpha_N\right) & \\ \text{subject to} & \alpha_k = f_k\left(\{\mathrm{MSE}_{k,i}\}_{i=1}^{L_k}\right) & 1 \leq k \leq N \\ & \mathrm{Tr}\left(\mathbf{P}_k \mathbf{P}_k^\dagger\right) \leq P_k & 1 \leq k \leq N \\ & \sum_{k=1}^{N} P_k \leq P_0, & \end{array} \tag{3.148}$$

where the global cost function has been separated for convenience into a set of functions f_k that evaluate the performance of each carrier and a global cost function f_0 that evaluates then the global performance (all functions are assumed to be increasing), and the variables $\{P_k\}$ have been introduced for convenience to denote the power distribution among the carriers.

The problem formulation with individual QoS constraints can be formulated, similarly to (3.15), as

$$\begin{array}{lll} \underset{\{\mathbf{P}_k, \mathbf{W}_k, P_k\}}{\text{minimize}} & \sum_{k=1}^{N} P_k & \\ \text{subject to} & \mathrm{Tr}\left(\mathbf{P}_k \mathbf{P}_k^\dagger\right) \leq P_k & 1 \leq k \leq N \\ & \mathrm{MSE}_{k,i} \leq \rho_{k,i} & 1 \leq k \leq N,\ 1 \leq i \leq L. \end{array} \tag{3.149}$$

It is not difficult to see that in the problem formulation with a global cost function in (3.148) all the carriers are coupled which makes the problem more complicated than the single MIMO channel counterpart in (3.13). Interestingly, it turns out that the problem formulation with individual QoS constraints in (3.149) actually decouples naturally into N individual problems each of them like (3.15) that can be solved as in Section 3.5. This observation is straightforward from (3.149) as minimizing the global power $\sum_{k=1}^{N} P_k$ is equivalent to minimizing each of the powers P_k subject to the individual MSE constraints.

To deal with problem (3.148), we can start as in Section 3.4 by obtaining the optimum transmitter, which is given by the MMSE receiver (or Wiener filter) and the ZF receiver (with the additional ZF constraint).

The problem reduces then to

$$\begin{array}{lll} \underset{\{\mathbf{P}_k,\alpha_k,\boldsymbol{\rho}_k,P_k\}}{\text{minimize}} & f_0\left(\alpha_1,\ldots,\alpha_N\right) & \\ \text{subject to} & \left[\left(\nu\mathbf{I}+\mathbf{P}_k^{\dagger}\mathbf{H}_k^{\dagger}\mathbf{H}_k\mathbf{P}_k\right)^{-1}\right]_{ii} \leq \rho_{k,i}, & 1\leq k\leq N,\ 1\leq i\leq L \\ & \alpha_k = f_k\left(\boldsymbol{\rho}_k\right) & 1\leq k\leq N \\ & \mathrm{Tr}\left(\mathbf{P}_k\mathbf{P}_k^{\dagger}\right)\leq P_k & 1\leq k\leq N \\ & \sum_{k=1}^{N} P_k \leq P_0. & \end{array} \tag{3.150}$$

This problem is very similar to the ones considered in Sections 3.4 and 3.6. In fact, the same simplifications based on majorization theory can be used, obtaining similar simplified problems with the additional complication of the power allocation among the carriers $\{P_k\}$. It is also possible to obtain closed-form solutions from the KKT conditions (as was done in Section 3.4 for a list of Schur-concave and Schur-convex functions) [108, 111]. For example, for the case of minimizing the average MSE over all carriers and substreams, it turns out that the solution is still given by the same waterfilling as in the single-carrier case. However, in general this problem is best approached via a primal decomposition approach (similar to the one considered in Section 3.6.1) as developed in detail in [109].

3.8 Summary

This chapter has considered the design of linear MIMO transceivers with perfect CSI under a very general framework based on majorization theory. Two different problem formulations have been considered: one based on a global cost function of the whole system and another based on imposing an individual QoS constraint on each transmitted data stream. Both formulations turn out to be very difficult nonconvex optimization problems with matrix-valued variables. With the aid of majorization theory, however, these problems can be reformulated as simple convex optimization problems with scalar-valued variables which can be optimally solved.

The optimal solution is characterized as follows: the optimum linear receiver is always the MMSE receiver (or Wiener filter) and the optimum linear precoder has the form $\mathbf{P} = \mathbf{V}_H \operatorname{diag}\left(\sqrt{\mathbf{p}}\right)\mathbf{Q}$, where $\mathbf{V}_H$ contains the right singular vectors of the channel matrix $\mathbf{H}$ (whose role is to diagonalize the channel), $\mathbf{p}$ denotes a power allocation over the channel eigenvalues, and $\mathbf{Q}$ is a unitary matrix that possibly rotates the symbols before being transmitted over the channel. The optimum power allocation $\mathbf{p}$ can be found in most cases in closed-form; otherwise, it can always be obtained by solving a simple problem numerically. The rotation $\mathbf{Q}$ can always be easily computed. In addition, two particular cases can be further particularized: if the cost function is Schur-concave, then $\mathbf{Q} = \mathbf{I}$, which means a fully diagonal transmission; and if the cost function is Schur-convex, then $\mathbf{Q}$ can be chosen as the DFT matrix or the unitary Hadamard matrix whose function is to spread the symbols over all used channel eigenvalues (in a CDMA fashion) such that they all have the same quality.

3.A Appendix: Characterization of BER Function for QAM Constellations

In this appendix, we prove that the BER function and also the corresponding Chernoff upper bound are convex decreasing functions of the SINR and convex increasing functions of the MSE (the latter for sufficiently small values of the MSE).

Before proceeding, recall that the symbol error probability can be analytically expressed as a function of the SINR as $P_e = \alpha\, \mathcal{Q}\left(\sqrt{\beta\,\mathrm{SINR}}\right)$, where the parameters $\alpha \geq 1$ and $\beta \leq 1$ depend on the constellation. As an example, for M-ary PAM, M-ary QAM, and M-ary PSK constellations this relation is specifically given by

$$P_e^{\mathrm{PAM}} \cong 2\left(1 - \frac{1}{M}\right)\mathcal{Q}\left(\sqrt{\frac{3}{M^2-1}\,\mathrm{SINR}}\right),$$

$$P_e^{\mathrm{QAM}} \cong 4\left(1 - \frac{1}{\sqrt{M}}\right)\mathcal{Q}\left(\sqrt{\frac{3}{M-1}\,\mathrm{SINR}}\right), \text{ and}$$

$$P_e^{\mathrm{PSK}} \cong 2\,\mathcal{Q}\left(\sqrt{2\sin^2\left(\frac{\pi}{M}\right)\mathrm{SINR}}\right) \quad \text{for } M \geq 4.$$

Table 3.3 Examples of parameters and convexity region of the BER for well-known constellations.

Constellation	M	k	α	β ($\simeq$)	Condition for convexity of BER(MSE)
BPSK	2	1	1	1	Always convex
4-PAM	4	2	1.5	0.2	BER $\leq 3.701 \times 10^{-2}$
16-PAM	16	4	1.875	0.0118	BER $\leq 1.971 \times 10^{-2}$
QPSK	4	2	2	1	Always convex
16-QAM	16	4	3	0.2	BER $\leq 3.701 \times 10^{-2}$
64-QAM	64	6	3.5	0.0476	BER $\leq 2.526 \times 10^{-2}$
8-PSK	8	3	2	0.2929	BER $\leq 3.576 \times 10^{-2}$
16-PSK	16	4	2	0.0761	BER $\leq 2.218 \times 10^{-2}$
32-PSK	32	5	2	0.0192	BER $\leq 1.692 \times 10^{-2}$

See Table 3.3 for specific values of the parameters (recall that M is the constellation size and $k = \log_2 M$ the number of bits per symbol).

3.A.1 BER as a Function of the SINR

To prove that the BER function is convex decreasing in the SINR, it suffices to show that the first and second derivatives of $\mathcal{Q}\left(\sqrt{\beta x}\right)$ are negative and positive, respectively (note that a positive scaling factor preserves monotonicity and convexity):

$$\begin{aligned} \frac{d}{dx}\mathcal{Q}\left(\sqrt{\beta x}\right) &= -\sqrt{\tfrac{\beta}{8\pi}}e^{-\beta x/2}x^{-1/2} < 0 & 0 < x < \infty \\ \frac{d^2}{dx^2}\mathcal{Q}\left(\sqrt{\beta x}\right) &= \tfrac{1}{2}\sqrt{\tfrac{\beta}{8\pi}}e^{-\beta x/2}x^{-1/2}\left(\tfrac{1}{x} + \beta\right) > 0 & 0 < x < \infty. \end{aligned} \tag{3.151}$$

The same can be done for the Chernoff upper bound $e^{-\beta x/2}$:

$$\begin{aligned} \frac{d}{dx}e^{-\beta x/2} &= -\tfrac{\beta}{2}e^{-\beta x/2} < 0 & 0 < x < \infty \\ \frac{d^2}{dx^2}e^{-\beta x/2} &= \left(\tfrac{\beta}{2}\right)^2 e^{-\beta x/s2} > 0 & 0 < x < \infty. \end{aligned} \tag{3.152}$$

3.A.2 BER as a Function of the MSE

To prove that the BER function is convex increasing in the MSE (assuming a ZF/MMSE receiver), it suffices to show that the first and second derivatives of $\mathcal{Q}\left(\sqrt{\beta\left(x^{-1}-\nu\right)}\right)$ are both positive:

$$\begin{aligned}
&\frac{d}{dx}\mathcal{Q}\left(\sqrt{\beta\left(x^{-1}-\nu\right)}\right)=\sqrt{\frac{\beta}{8\pi}}e^{-\beta\left(x^{-1}-\nu\right)/2}\left(x^{3}-\nu x^{4}\right)^{-1/2}\geq 0\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 0<x\leq 1\\
&\frac{d^{2}}{dx^{2}}\mathcal{Q}\left(\sqrt{\beta\left(x^{-1}-\nu\right)}\right)\\
&\quad=\frac{1}{2}\sqrt{\frac{\beta}{8\pi}}e^{-\beta\left(x^{-1}-\nu\right)/2}\left(x^{3}-\nu x^{4}\right)^{-1/2}\left(\frac{\beta}{x^{2}}-\frac{3-\nu 4x}{x-\nu x^{2}}\right)\geq 0\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 0<x\leq x_{z_1},\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad x_{z_2}\leq x\leq 1.
\end{aligned} \tag{3.153}$$

For the ZF receiver the condition for convexity is $x\leq\beta/3$. For the MMSE receiver, the zeros are $x_{z_1}=\left(\beta+3-\sqrt{\beta^{2}-10\beta+9}\right)/8$ and $x_{z_2}=\left(\beta+3+\sqrt{\beta^{2}-10\beta+9}\right)/8$ (it has been tacitly assumed that $\beta\leq 1$). It is remarkable that for $\beta=1$ both zeros coincide, which means that the BER function is convex for the whole range of MSE values. To be more specific, BPSK and QPSK constellations satisfy this condition and, consequently, their corresponding BER function is always convex in the MSE (see Table 3.3).

Consider now the Chernoff upper bound $e^{-\beta\left(x^{-1}-\nu\right)/2}$:

$$\begin{aligned}
&\frac{d}{dx}e^{-\beta\left(x^{-1}-\nu\right)/2}=\frac{\beta}{2}e^{-\beta\left(x^{-1}-\nu\right)/2}x^{-2}>0 && 0<x\leq 1\\
&\frac{d^{2}}{dx^{2}}e^{-\beta\left(x^{-1}-\nu\right)/2}=\frac{\beta}{2}e^{-\beta\left(x^{-1}-\nu\right)/2}x^{-4}\left(\frac{\beta}{2}-2x\right)\geq 0 && 0<x\leq\frac{\beta}{4}.
\end{aligned} \tag{3.154}$$

Thus, the convexity range given in (3.37) and (3.38) has been shown. As a rule-of-thumb, both the exact BER function and the Chernoff upper bound are convex increasing functions of the MSE for a BER $\leq 2\times 10^{-2}$ (see Table 3.3).

3.B Appendix: Optimum Left Singular Vectors of P

We first present a required basic result and then the desired result.

Lemma 3.15. [97, 9.H.1.h] If $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ positive semidefinite Hermitian matrices, then

$$\mathrm{Tr}\left(\mathbf{AB}\right) \geq \sum_{i=1}^{n} \lambda_{A,i}\, \lambda_{B,n-i+1},$$

where $\lambda_{A,i}$ and $\lambda_{B,i}$ are the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$, respectively, in decreasing order.

The following result was first proved in [111] and states that for a given matrix value of $\mathbf{P}^{\dagger}\mathbf{R}_H\mathbf{P}$ (assumed to be a diagonal matrix) we can always choose a matrix $\mathbf{P}$ of the form $\mathbf{P} = \mathbf{V}_H\boldsymbol{\Sigma}$, where $\mathbf{V}_H$ contains the eigenvectors of $\mathbf{R}_H$, in order to minimize the Frobenius norm of $\mathbf{P}$, i.e., the left singular vectors of $\mathbf{P}$ coincide with the eigenvectors of $\mathbf{R}_H$.

Lemma 3.16. Let $\mathbf{P}$ be an $n \times L$ matrix, $\mathbf{R}_H$ be an $n \times n$ positive semidefinite matrix, and $\mathbf{P}^{\dagger}\mathbf{R}_H\mathbf{P}$ be a diagonal matrix with diagonal elements in increasing order (possibly with some zero diagonal elements). Then, there is a matrix of the form $\tilde{\mathbf{P}} = \mathbf{V}_H\left[\mathbf{0}, \boldsymbol{\Sigma}\right]$ that satisfies

$$\tilde{\mathbf{P}}^{\dagger}\mathbf{R}_H\tilde{\mathbf{P}} = \mathbf{P}^{\dagger}\mathbf{R}_H\mathbf{P} \tag{3.155}$$

$$\mathrm{Tr}\left(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^{\dagger}\right) \leq \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right), \tag{3.156}$$

where $\mathbf{V}_H$ is a (semi-)unitary matrix with columns equal to the eigenvectors of matrix $\mathbf{R}_H$ corresponding to the $\min\left(n, L\right)$ largest eigenvalues in increasing order and $\boldsymbol{\Sigma}$ is square diagonal matrix of size $\min\left(n, L\right)$ (the left zero-block in $\tilde{\mathbf{P}}$ accounts for the case in which $\mathbf{P}$ is a fat matrix, i.e., $L > n$).

Proof. Since $\mathbf{P}^{\dagger}\mathbf{R}_H\mathbf{P}$ is diagonal with diagonal elements in increasing order, we can write

$$\mathbf{P}^{\dagger}\mathbf{R}_H\mathbf{P} = \tilde{\mathbf{D}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix},$$

where $\mathbf{D}$ is a diagonal matrix of size $\min(n, L)$. We can then write the singular value decomposition (SVD) $\mathbf{R}_H^{1/2}\mathbf{P} = \mathbf{Q}\,[\mathbf{0}, \mathbf{\Sigma}]$, where $\mathbf{Q}$ is an arbitrary (semi-)unitary matrix containing the left singular vectors, the right singular vector matrix (eigenvectors of $\tilde{\mathbf{D}}$) is the identity matrix, and matrix $\mathbf{\Sigma} = \mathbf{D}^{1/2}$ contains the singular values.

Assume for the moment and for the sake of notation that $\mathbf{R}_H$ is nonsingular with eigenvalue decomposition $\mathbf{R}_H = \tilde{\mathbf{V}}_H\tilde{\mathbf{D}}_H\tilde{\mathbf{V}}_H^\dagger$ (assume the diagonal elements of $\tilde{\mathbf{D}}_H$ in increasing order), then we can write

$$\mathbf{P} = \mathbf{R}_H^{-1/2}\mathbf{Q}\,[\mathbf{0}, \mathbf{\Sigma}] = \tilde{\mathbf{V}}_H\tilde{\mathbf{D}}_H^{-1/2}\tilde{\mathbf{V}}_H^\dagger\mathbf{Q}\,[\mathbf{0}, \mathbf{\Sigma}] \tag{3.157}$$

for some (semi-)unitary matrix $\mathbf{Q}$.

The idea now is to find another matrix $\tilde{\mathbf{P}}$ by properly choosing the (semi-)unitary matrix $\mathbf{Q}$ in (3.157) such that $\operatorname{Tr}\left(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^\dagger\right)$ has the smallest value (note that any matrix $\tilde{\mathbf{P}}$ obtained from (3.157) satisfies by definition the desired constraint $\tilde{\mathbf{P}}^\dagger\mathbf{R}_H\tilde{\mathbf{P}} = \tilde{\mathbf{D}}$).

In case that $\mathbf{R}_H$ is singular, we can clearly assume that $\tilde{\mathbf{P}}$ is orthogonal to the null space of $\mathbf{R}_H$, otherwise the nonorthogonal component could be made zero without changing the value of $\tilde{\mathbf{P}}^\dagger\mathbf{R}_H\tilde{\mathbf{P}}$ and decreasing $\operatorname{Tr}\left(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^\dagger\right)$. Knowing that $\tilde{\mathbf{P}}$ can be assumed to be orthogonal to the null space of $\mathbf{R}_H$ without loss of generality, expression (3.157) is still valid using the pseudo-inverse of $\mathbf{R}_H$ instead of the inverse.

Using now Lemma 3.15, $\operatorname{Tr}\left(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^\dagger\right)$ can be lower-bounded as follows:

$$\operatorname{Tr}\left(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^\dagger\right) = \operatorname{Tr}\left(\mathbf{\Sigma}\mathbf{\Sigma}^\dagger\tilde{\mathbf{V}}^\dagger\tilde{\mathbf{D}}_H^{-1}\tilde{\mathbf{V}}\right) \geq \sum_{i=1}^{\min(n,L)} d_i\,\lambda_{H,i}^{-1}, \tag{3.158}$$

where $\tilde{\mathbf{V}} \triangleq \tilde{\mathbf{V}}_H^\dagger\mathbf{Q}$, d_i is the ith diagonal element of $\mathbf{D}$ in increasing order and $\{\lambda_{H,i}\}_{i=1}^{\min(n,L)}$ are the $\min(n, L)$ largest eigenvalues of $\mathbf{R}_H$ in increasing order. The lower bound is achieved by matrix $\tilde{\mathbf{V}}$ choosing the $\min(n, L)$ largest diagonal elements of $\tilde{\mathbf{D}}_H$ in the same order as the d_i's, i.e., by $\tilde{\mathbf{V}} = [\mathbf{0}, \mathbf{I}]^\dagger$. From (3.157), we obtain that the optimal $\tilde{\mathbf{P}}$ (in the sense of minimizing the value of $\operatorname{Tr}\left(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^\dagger\right)$) is of the form $\tilde{\mathbf{P}} = \tilde{\mathbf{V}}_H\tilde{\mathbf{D}}_H^{-1/2}[\mathbf{0}, \mathbf{I}]^\dagger[\mathbf{0}, \mathbf{\Sigma}] = \mathbf{V}_H\mathbf{D}_H^{-1/2}[\mathbf{0}, \mathbf{\Sigma}] = \mathbf{V}_H\left[\mathbf{0}, \mathbf{D}_H^{-1/2}\mathbf{\Sigma}\right]$, which is the desired result by renaming the diagonal matrix $\mathbf{D}_H^{-1/2}\mathbf{\Sigma}$ as $\mathbf{\Sigma}$. □

In fact, a result stronger than Lemma 3.16 can be easily proved: for a given matrix value of $\mathbf{P}^{\dagger}\mathbf{R}_H\mathbf{P}$ (not necessarily diagonal) we can always choose matrix $\mathbf{P}$ of the form $\mathbf{P} = \mathbf{V}_H \boldsymbol{\Sigma} \mathbf{Q}^{\dagger}$ in order to minimize the Frobenius norm of $\mathbf{P}$, i.e., the left singular vectors of $\mathbf{P}$ coincide again with the eigenvectors of $\mathbf{R}_H$.

3.C Appendix: Optimum Ordering of Eigenvalues

Lemma 3.17. Let $p_1 \lambda_1 < p_2 \lambda_2$ and $\lambda_1 > \lambda_2 > 0$. Then, there exist $\tilde{p}_1$ and $\tilde{p}_2$ such that $\tilde{p}_1 \lambda_2 = p_1 \lambda_1$ and $\tilde{p}_2 \lambda_1 = p_2 \lambda_2$ with $\tilde{p}_1 + \tilde{p}_2 < p_1 + p_2$.

Proof. Clearly,

$$\begin{aligned}(\tilde{p}_1 + \tilde{p}_2) - (p_1 + p_2) &= p_1 \left(\frac{\lambda_1}{\lambda_2} - 1\right) + p_2 \left(\frac{\lambda_2}{\lambda_1} - 1\right)\\ &< p_2 \frac{\lambda_2}{\lambda_1}\left(\frac{\lambda_1 - \lambda_2}{\lambda_2}\right) + p_2 \left(\frac{\lambda_2 - \lambda_1}{\lambda_1}\right)\\ &= 0, \end{aligned} \tag{3.159}$$

where the inequality follows from $p_1 \lambda_1 < p_2 \lambda_2$. □

3.D Appendix: Proofs of Schur-Concavity/Convexity Lemmas

Proof of Lemma 3.2: $(f_0(\mathbf{x}) = \sum_i (\alpha_i x_i))$

Since the weights are in increasing order $\alpha_i \leq \alpha_{i+1}$, the function $f_0(\mathbf{x}) = \sum_i (\alpha_i x_i)$ is minimized with the x_i's in decreasing order $x_i \geq x_{i+1}$. To show this, suppose that for $i < j$ $(\alpha_i \leq \alpha_j)$ the arguments are such that $x_i < x_j$. It follows that the term $(\alpha_i x_i + \alpha_j x_j)$ can be minimized by simply swapping the arguments:

$$\begin{aligned} & x_i (\alpha_j - \alpha_i) \leq x_j (\alpha_j - \alpha_i) \\ \Longleftrightarrow \quad & \alpha_i x_i + \alpha_j x_j \geq \alpha_i x_j + \alpha_j x_i. \end{aligned}$$

To prove now that the function f_0 is Schur-concave (see Definition 2.4), define $\phi(\mathbf{x}) \triangleq -f_0(\mathbf{x}) = \sum_i g_i(x_i)$, where $g_i(x) = -\alpha_i x$. Function

ϕ is Schur-convex because $g_i'(a) \geq g_{i+1}'(b)$ (see Lemma 2.7) and, therefore, f_0 is Schur-concave (see Definition 2.4). □

Proof of Lemma 3.3: $\left(f_0(\mathbf{x}) = \prod_i x_i^{\alpha_i}\right)$

Since the weights are in increasing order $\alpha_i \leq \alpha_{i+1}$, the function $f_0(\mathbf{x}) = \prod_i x_i^{\alpha_i}$ is minimized with the x_i's in decreasing order $x_i \geq x_{i+1}$. To show this, suppose that for $i < j$ $(\alpha_i \leq \alpha_j)$ the arguments are such that $0 < x_i < x_j$. It follows that the term $x_i^{\alpha_i} x_j^{\alpha_j}$ can be minimized by simply swapping the arguments:

$$\begin{aligned} & x_i^{\alpha_j - \alpha_i} \leq x_j^{\alpha_j - \alpha_i} \\ \Longleftrightarrow \quad & x_i^{\alpha_i} x_j^{\alpha_j} \geq x_j^{\alpha_i} x_i^{\alpha_j}. \end{aligned}$$

If some argument is equal to zero, then the value of the function is also zero regardless of the ordering.

To prove now that the function f_0 is Schur-concave (see Definition 2.4), define $\phi(\mathbf{x}) \triangleq -\log f_0(\mathbf{x}) = \sum_i g_i(x_i)$, where $g_i(x) = -\alpha_i \log x$. Function ϕ is Schur-convex because $g_i'(a) \geq g_{i+1}'(b)$ whenever $a \geq b$ (see Lemma 2.7). Since $f_0(\mathbf{x}) = e^{-\phi(\mathbf{x})}$ and function e^{-x} is decreasing in x, f_0 is Schur-concave by Lemma 2.5. □

Proof of Lemma 3.4: $\left(f_0(\mathbf{x}) = -\sum_i \alpha_i \left(x_i^{-1} - \nu\right)\right)$

Since the weights are in increasing order $\alpha_i \leq \alpha_{i+1}$, the function $f_0(\mathbf{x}) = -\sum_i \alpha_i \left(x_i^{-1} - \nu\right)$ is minimized with the x_i's in decreasing order $x_i \geq x_{i+1} > 0$ (this can be shown similarly as in the proof of Lemma 3.2).

To prove that the function f_0 is Schur-concave (see Definition 2.4), define $\phi(\mathbf{x}) \triangleq -f_0(\mathbf{x}) = \sum_i g_i(x_i)$, where $g_i(x) = \alpha_i \left(x^{-1} - \nu\right)$. Function ϕ is Schur-convex because $g_i'(a) \geq g_{i+1}'(b)$ whenever $a \geq b$ (see Lemma 2.7) and, therefore, f_0 is Schur-concave by Definition 2.4. □

Proof of Lemma 3.5: $\left(f_0(\mathbf{x}) = -\prod_i \left(x_i^{-1} - \nu\right)^{\alpha_i}\right)$

Since the weights are in increasing order $\alpha_i \leq \alpha_{i+1}$, the function $f_0(\mathbf{x}) = -\prod_i \left(x_i^{-1} - \nu\right)^{\alpha_i}$ is minimized with the x_i's in decreasing order $1 \geq x_i \geq x_{i+1} > 0$ (this can be shown similarly as in the proof of Lemma 3.3).

To prove that the function f_0 is Schur-concave for $x_i \leq 0.5$ (see Definition 2.4), define $\phi(\mathbf{x}) \triangleq \log\left(-f_0(\mathbf{x})\right) = \sum_i g_i(x_i)$, where

$g_i(x) = \alpha_i \log\left(x^{-1} - 1\right)$. Function ϕ is Schur-convex because $g_i'(a) \geq g_{i+1}'(b)$ whenever $0.5 \geq a \geq b$[23] (see Lemma 2.7). Since $f_0(\mathbf{x}) = -e^{\phi(\mathbf{x})}$ and function $-e^x$ is decreasing in x, f_0 is Schur-concave by Lemma 2.5. □

Proof of Lemma 3.6: $\left(f_0(\mathbf{x}) = \prod_i g(x_i)\right)$

To prove that the function $f_0(\mathbf{x}) = \prod_i g(x_i)$ is Schur-concave for $\theta \geq x_i > 0$ (for sufficiently small θ such that $\left(\frac{\partial g(x)}{\partial x}\right)^2 \geq g(x)\frac{\partial^2 g(x)}{\partial x^2}$ for $\theta \geq x > 0$), define $\phi(\mathbf{x}) \triangleq -\log f_0(\mathbf{x}) = \sum_i h(x_i)$, where $h(x) = -\log g(x_i)$. Function h is convex for $0 < x \leq \theta$ because $\frac{\partial^2 h(x)}{\partial x^2} = \frac{1}{(g(x))^2}\left(\left(\frac{\partial g(x)}{\partial x}\right)^2 - g(x)\frac{\partial^2 g(x)}{\partial x^2}\right) \geq 0$ and ϕ is Schur-convex by Corollary 2.5. Since $f_0(\mathbf{x}) = e^{-\phi(\mathbf{x})}$ and function e^{-x} is decreasing in x, f_0 is Schur-concave by Lemma 2.5. □

Proof of Lemma 3.7: $(f_0(\mathbf{x}) = \max_i\{x_i\})$

From Definition 2.1, it follows that $f_0(\mathbf{x}) = \max_i\{x_i\} = x_{[1]}$. If $\mathbf{x} \prec \mathbf{y}$ it must be that $x_{[1]} \leq y_{[1]}$ (from Definition 2.2) and, therefore, $f_0(\mathbf{x}) \leq f_0(\mathbf{y})$. This means that f_0 is Schur-convex by Definition 2.4. □

Proof of Lemma 3.8: $\left(f_0(\mathbf{x}) = \sum_i \frac{x_i}{1-\nu x_i}\right)$

For $\nu = 0$ (ZF receiver), the function reduces to $f_0(\mathbf{x}) = \sum_i x_i$ which can be classified as Schur-convex (also as Schur-concave). For $\nu = 1$ (MMSE receiver), rewrite the function as $f_0(\mathbf{x}) = \sum_i g(x_i)$, where $g(x) = \frac{x}{1-x}$. Since function g is convex, it follows that f_0 is Schur-convex by Corollary 2.5. □

Proof of Lemma 3.9: $\left(f_0(\mathbf{x}) = \sum_i g(x_i)\right)$

To prove that the function $f_0(\mathbf{x}) = \sum_i g(x_i)$ is Schur-convex for $\theta \geq x_i > 0$ (for sufficiently small θ such that $\mathrm{BER}\left(x_i^{-1} - 1\right) \leq 10^{-2}$ $\forall i$), simply note that function g is convex within the range $(0, \theta]$ (see Section 3.3.2 and Appendix 3.A), it follows that f_0 is Schur-convex by Corollary 2.5. □

[23] Function $(1-x)x$ is increasing in x for $0 \leq x \leq 0.5$.

3.E Appendix: Waterfilling Algorithms

3.E.1 General Algorithm for Waterfillings with Single Waterlevel

In many cases, the optimal power allocation follows a waterfilling form: $p_i = (\mu a_i - b_i)^+$, where the a_i's and b_i's are some fixed numbers and μ is a waterlevel that has to be found to satisfy the power constraint with equality $\sum_i p_i = P_0$ or, more generally, some condition $g(\mu) = 0$, where g is an increasing function. The numerical evaluation of such waterfilling solutions can be done efficiently in practice either by bisection or by hypothesis testing as explored in detail in [112, 108]. For convenience, we reproduce in Algorithm 3.2 a method based on hypothesis testing with a worst-case complexity of L iterations.

Algorithm 3.2. Practical algorithm to evaluate waterfilling solutions of the form $p_i = (\mu a_i - b_i)^+$, for $1 \leq i \leq L$, subject to the constraint $g(\mu) = 0$.

Input: Set of pairs $\{(a_i, b_i)\}$ and constraint function g.

Output: Numerical solution $\{p_i\}$ and waterlevel μ.

0. Set $\tilde{L} = L$ and (if necessary) sort the set of pairs $\{(a_i, b_i)\}$ such that a_i/b_i are in decreasing order $a_i/b_i \geq a_{i+1}/b_{i+1}$ (define $a_{L+1}/b_{L+1} \triangleq 0$).
1. If $b_{\tilde{L}}/a_{\tilde{L}} < b_{\tilde{L}+1}/a_{\tilde{L}+1}$ and $g\left(b_{\tilde{L}}/a_{\tilde{L}}\right) < 0$, then accept the hypothesis and go to step 2.

   Otherwise reject the hypothesis, form a new one by setting $\tilde{L} = \tilde{L} - 1$, and go to step 1.
2. Find the waterlevel $\mu \in \left(b_{\tilde{L}}/a_{\tilde{L}}, b_{\tilde{L}+1}/a_{\tilde{L}+1}\right] \mid g(\mu) = 0$, obtain the numerical solution as

$$p_i = (\mu a_i - b_i)^+ \quad 1 \leq i \leq L,$$

   undo the sorting done at step 0 (if any), and finish.

For example, for the minimization of the unweighted sum of the MSEs subject to a power constraint in Section 3.4.2.1 and for the minimization of a Schur-convex function in Section 3.4.3, the optimal power allocation with an MMSE receiver is

$$p_i = \left(\mu \lambda_{H,i}^{-1/2} - \lambda_{H,i}^{-1}\right)^+ \quad 1 \le i \le L, \tag{3.160}$$

where μ is the waterlevel chosen such that $\sum_i p_i = P_0$. To obtain a practical algorithm, it suffices to use Algorithm 3.2 introducing the following particularizations:

$$\begin{aligned} a_i &= \lambda_{H,i}^{-1/2},\ b_i = \lambda_{H,i}^{-1}, \\ g(\mu) &= \mu \sum_{i=1}^{\tilde{L}} \lambda_{H,i}^{-1/2} - \sum_{i=1}^{\tilde{L}} \lambda_{H,i}^{-1} - P_T, \end{aligned}$$

where $\tilde{L}$ denotes the number of active substreams (with positive power). The comparison $g\left(b_{\tilde{L}}/a_{\tilde{L}}\right) < 0$ reduces to

$$\lambda_{H,\tilde{L}}^{-1/2} < \left(P_T + \sum_{i=1}^{\tilde{L}} \lambda_{H,i}^{-1}\right) \Big/ \sum_{i=1}^{\tilde{L}} \lambda_{H,i}^{-1/2} \tag{3.161}$$

and $\mu \mid g(\mu) = 0$ is obtained as

$$\mu = \left(P_T + \sum_{i=1}^{\tilde{L}} \lambda_{H,i}^{-1}\right) \Big/ \sum_{i=1}^{\tilde{L}} \lambda_{H,i}^{-1/2}. \tag{3.162}$$

3.E.2 Algorithm for Waterfilling with Single Waterlevel in the Equal-MSE Constrained Design

Another instance of Algorithm 3.2 is for the minimum power design with equal QoS constraints in Section 3.5. The optimal power allocation is

$$p_i = \left(\mu \lambda_{H,i}^{-1/2} - \nu \lambda_{H,i}^{-1}\right)^+ \quad 1 \le i \le L, \tag{3.163}$$

where μ is the waterlevel chosen such that $\sum_{i=1}^{L} \frac{1}{\nu + p_i \lambda_{H,i}} = \tilde{\rho} \triangleq L\rho$. After particularizing, the following algorithm is obtained for the case of an MMSE receiver ($\nu = 1$).

Algorithm 3.3. Practical algorithm to evaluate the waterfilling solution $p_i = \left(\mu \lambda_{H,i}^{-1/2} - \lambda_{H,i}^{-1}\right)^+$, subject to the constraint $\sum_{i=1}^{L} \frac{1}{1 + p_i \lambda_{H,i}} \le \tilde{\rho}$, corresponding to the design with equal MSE QoS requirements.

Input: Number of eigenvalues L, set of eigenvalues $\{\lambda_{H,i}\}$, and MSE constraint $\tilde{\rho}$.

Output: Numerical solution $\{p_i\}$ and waterlevel μ.

0. Set $\tilde{L} = L$ and (if necessary) sort the $\lambda_{H,i}$'s in decreasing order $\lambda_{H,i} \geq \lambda_{H,i+1}$ (define $\lambda_{H,L+1} \triangleq 0$).
1. Set $\mu = \lambda_{H,\tilde{L}}^{-1/2}$ (if $\lambda_{H,\tilde{L}} = \lambda_{H,\tilde{L}+1}$, then set $\tilde{L} = \tilde{L} - 1$ and go to step 1).
2. If $\mu \geq \left(\sum_{i=1}^{\tilde{L}} \lambda_{H,i}^{-1/2}\right) / \left(\tilde{\rho} - (L - \tilde{L})\right)$, then set $\tilde{L} = \tilde{L} - 1$ and go to step 1.

   Otherwise, obtain the definitive waterlevel μ and power allocation as

   $$\mu = \frac{\sum_{i=1}^{\tilde{L}} \lambda_{H,i}^{-1/2}}{\tilde{\rho} - \left(L - \tilde{L}\right)} \quad \text{and} \quad p_i = \left(\mu \lambda_{H,i}^{-1/2} - \lambda_{H,i}^{-1}\right)^+ \quad 1 \leq i \leq L,$$

   undo the sorting done at step 0, and finish.

Observe that for a ZF receiver ($\nu = 0$), Algorithm 3.3 is unnecessary as the optimal power allocation is directly given by

$$p_i = \lambda_{H,i}^{-1/2} \frac{\sum_{i=1}^{L} \lambda_{H,i}^{-1/2}}{\tilde{\rho}} \quad 1 \leq i \leq L. \tag{3.164}$$

3.E.3 Algorithm for Waterfilling with Multiple Waterlevels in the Different-MSE Constrained Design

Consider now the numerical evaluation of the waterfilling solution $p_i = \left(\mu_i \lambda_{H,i}^{-1/2} - \nu \lambda_{H,i}^{-1}\right)^+$ with multiple waterlevels constrained to satisfy the conditions in (3.112). This problem is much more involved than the previous cases of waterfillings with a single waterlevel. Nevertheless, it can still be done efficiently with the following algorithm, originally derived in [114], with a worst-case complexity of $L^2 (L+1)/2$ iterations.

Algorithm 3.4. Practical algorithm to solve the multilevel waterfilling solution $p_i = \left(\mu_i \lambda_{H,i}^{-1/2} - \nu \lambda_{H,i}^{-1}\right)^+$, subject to the conditions in (3.112), corresponding to the design with different MSE QoS requirements.

Input: Number of eigenvalues L, set of eigenvalues $\{\lambda_{H,i}\}$, and set of MSE constraints $\{\rho_i\}$ (note that the appropriate ordering of the $\lambda_{H,i}$'s and of the ρ_i's is independent of this algorithm).

Output: Numerical solution $\{p_i\}$ and set of waterlevels $\{\mu_i\}$.

0. Set $k_0 = 1$ and $\tilde{L} = L$.
1. Solve the equal MSE QoS constrained problem in $[k_0, \tilde{L}]$ using Algorithm 3.3 (or (3.164) for $\nu = 0$) with the set of $\tilde{L} - k_0 + 1$ eigenvalues $\{\lambda_{H,i}\}_{i=k_0}^{\tilde{L}}$ and with the MSE constraint given by $\tilde{\rho} = \sum_{i=k_0}^{\tilde{L}} \rho_i$.
2. If any intermediate constraint is not satisfied $\left(\sum_{i=k}^{\tilde{L}} \frac{1}{\nu + p_i \lambda_{H,i}} \not\leq \sum_{i=k}^{\tilde{L}} \rho_i,\ k_0 < k \leq \tilde{L}\right)$, then set k_0 equal to the smallest index whose constraint is not satisfied and go to step 1. Otherwise, if $k_0 = 1$ finish and if $k_0 > 1$ set $\tilde{L} = k_0 - 1$, $k_0 = 1$, and go to step 1.

4

Nonlinear Decision Feedback MIMO Transceivers

The preceding chapter focused on the design of *linear* MIMO transceivers, i.e., the combination of a linear precoder with a linear equalizer. In this chapter, we shall introduce another paradigm of MIMO transceiver designs where the linear equalizer is replaced by a (nonlinear) decision feedback equalizer (DFE). The distinction between the DFE and the linear equalizer is that the former exploits the finite alphabet property of communication signals and recovers signals successively. In doing so, the DFE enjoys a significant performance gain over the linear receiver. As a consequence, the DFE based MIMO transceiver designs have superior performance than the linear transceivers.

The research on joint design of a linear transmitter and a DFE can be traced back to Salz's work in the 1970s [128], where the design was done according to the MMSE criterion in the SISO intersymbol interference (ISI) channels. The paper [170] can somehow be regarded as an extension of [128] to the MIMO ISI channel, where the linear transmitter and DFE are optimized to minimize the determinant of the MSE matrix. Using majorization theory and the recent matrix decompositions [69, 70], one can design the decision feedback (DF) transceiver according to diverse design criteria. The recent development has been reported in [71, 72, 73]. The related work also appears in [169, 176]. We refer to this class of MIMO transceiver designs as *nonlinear DF MIMO transceiver designs*. In the context where no confusion arises, they are simply referred to as nonlinear designs.[1]

In this chapter, we will explore multiple aspects of the nonlinear transceiver designs. In the first five sections, we develop the nonlinear designs in a somewhat parallel manner to Chapter 3. After introducing the system model and problem formulation in Sections 4.1 and 4.2, we derive the optimum DFE in Section 4.3, which leads to the closed-form representation of MMSE–DFE. Sections 4.4 and 4.5 deal with the optimum transmitter design according to two different problem formulations: (i) the optimization of the global measure of system performance with input power constraint and (ii) the minimization of the overall input power subject to individual QoS constraints. The optimization problems are solved using majorization theory and recently developed matrix analysis tools [70]. Sections 4.4 and 4.5 reveal remarkable mathematical symmetry between the linear designs of Chapter 3 and the DF based nonlinear designs.

Besides the nonlinear design using the DFE which extracts the interfering substreams at the *receiver*, there is another implementation using dirty paper coding (DPC) for interference cancelation at the *transmitter*, as introduced in Section 4.6. Section 4.7 relates the design of MIMO transceivers to the problem of CDMA sequence optimization by realizing that the latter is actually a special case of the former. Therefore, the preceding MIMO transceiver designs can also be applied to the optimization CDMA sequences with little modification. Section 4.8 summarizes the chapter.

4.1 System Model

In this chapter, we consider the same communication system as introduced in Section 3.1. The received sampled baseband signal is

$$\mathbf{y} = \mathbf{HPx} + \mathbf{n}, \tag{4.1}$$

[1] Other nonlinear designs are possible, such as the ML-based transceiver introduced in Section 5.3.

where $\mathbf{H} \in \mathbb{C}^{n_R \times n_T}$ is the channel matrix with rank K, $\mathbf{P} \in \mathbb{C}^{n_T \times L}$ is the linear precoder matrix, and the additive noise $\mathbf{n}$ is zero-mean circularly symmetric Gaussian with covariance matrix $\mathbb{E}[\mathbf{n}\mathbf{n}^\dagger] = \mathbf{I}$.[2] We also assume that $\mathbb{E}[\mathbf{x}\mathbf{x}^\dagger] = \mathbf{I}_L$. Hence the input power of the system is

$$P_T = \mathbb{E}[\|\mathbf{P}\mathbf{x}\|^2] = \mathrm{Tr}(\mathbf{P}\mathbf{P}^\dagger). \tag{4.2}$$

We consider a MIMO transceiver structure illustrated in Figure 4.1. Comparing it to Figure 3.1, we see that the linear receiver in Figure 3.1 is replaced by a (nonlinear) DFE. Figure 4.2 shows the details of the receiver. Distinct from the linear receiver which applies linear filters to estimate the substreams *simultaneously*, the DFE detects the substreams *successively* with the Lth substream (x_L) detected first and the first substream detected last. At each step, a linear feed-forward filter is applied to the received signal vector and the previously detected substreams are fed back to facilitate the detection. The block $Q[\cdot]$ stands for mapping the "analog" estimate $\hat{x}_i$ to the closest point in the signal constellation. If the detection is erroneous, it may cause more errors in the subsequent detections, which is the so-called error propagation effect. To simplify the system design and performance analysis, we invoke the usual assumption that the system is free of error propagation. (In fact, this assumption can be justified by Shannon's theory; if long decoding latency is allowed, one can apply powerful coding to achieve arbitrarily small error probability given that the information rate of the data sub-

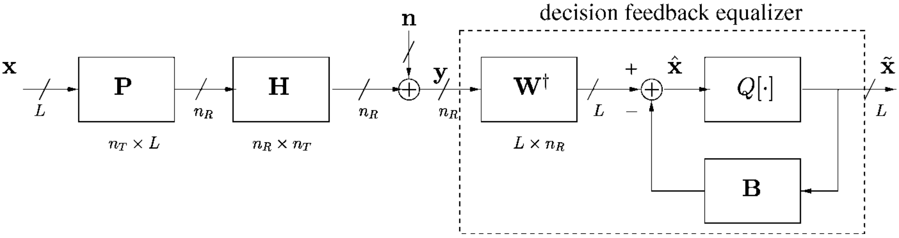


Fig. 4.1 Scheme of a MIMO Communication system with DFE receiver.
©2007 IEEE. Reprinted with permission from IEEE.

[2] If the noise is spatially correlated with known covariance matrix $\mathbf{R}_n$, we can always prewhiten the received data vector $\mathbf{y}$ to $\tilde{\mathbf{y}} = \mathbf{R}_n^{-1/2}\mathbf{y} = \mathbf{R}_n^{-1/2}\mathbf{H}\mathbf{P}\mathbf{x} + \mathbf{R}_n^{-1/2}\mathbf{n}$ so that the covariance of the transformed noise vector $\tilde{\mathbf{n}} = \mathbf{R}_n^{-1/2}\mathbf{n} \sim N(0, \mathbf{I})$.

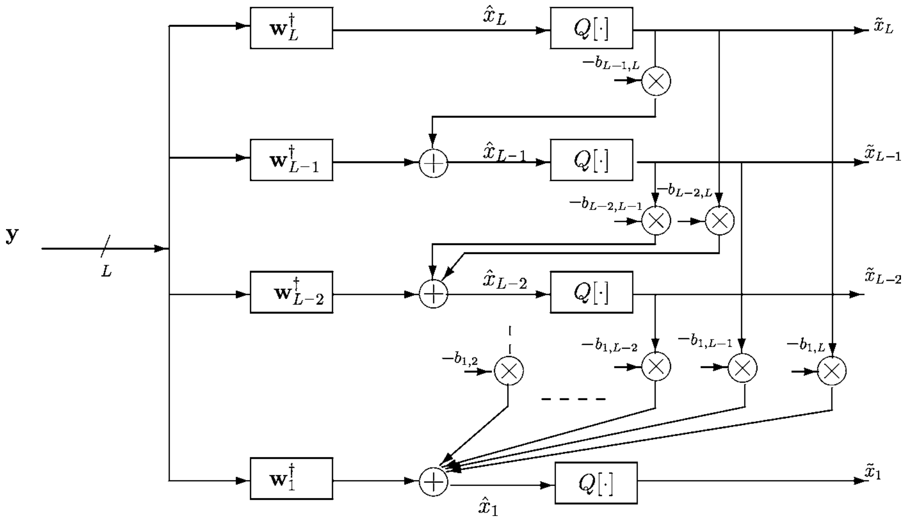


Fig. 4.2 DFE, the substreams are detected successively from x_L to x_1.

stream is less than the subchannel capacity.[3]) Hence the influence of error propagation can also be made negligible. If the DFE is combined with error control coding, the symbol-by-symbol detection block $Q[\cdot]$ should be replaced by a decoder.

We note the fundamental difference between a linear receiver and the DFE: the linear receiver used in Chapter 3 does not assume "digital" nature of the signal $\mathbf{x}$. In contrast, the nonlinear DFE exploits the fact that the digital communication signals are drawn from a finite alphabet. As Forney put it in [42]: "the ideal decision feedback assumption is the decisive break between the classical analog estimation theory of Wiener *et al.* and the digital Shannon theory." Somewhat related to this observation, using "digital" DFE rather than "analog" linear Wiener receiver leads to a different paradigm of MIMO transceiver design from those established in Chapter 3. In Appendix 4.A, we provide the insight that a linear receiver usually incurs information loss.

[3] Since L substreams are simultaneously transmitted through the MIMO channel, we may regard the system as if each substream is transmitted through a scalar subchannel.

With the error-propagation-free assumption, the "analog" estimate $\hat{x}_i$ can be written as

$$\hat{x}_i = \mathbf{w}_i^\dagger \mathbf{y} - \sum_{j=i+1}^{L} b_{ij} x_j, \quad 1 \leq i \leq L. \tag{4.3}$$

Denoting $\hat{\mathbf{x}} = [\hat{x}_1, \ldots, \hat{x}_L]^T$, $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_L] \in \mathbb{C}^{n_R \times L}$, and $\mathbf{B} \in \mathbb{C}^{L \times L}$ as a strictly upper triangular matrix with entries b_{ij}, (4.3) may be represented in the matrix form:

$$\hat{\mathbf{x}} = \mathbf{W}^\dagger \mathbf{y} - \mathbf{B}\mathbf{x} = (\mathbf{W}^\dagger \mathbf{H}\mathbf{P} - \mathbf{B})\mathbf{x} + \mathbf{W}^\dagger \mathbf{n}. \tag{4.4}$$

4.1.1 Measures of performance: MSE, SINR, and Channel Capacity

Denote $\mathbf{p}_i$ the ith column of $\mathbf{P}$. The performance of the substreams given in (4.3), measured in terms of the MSE and SINR, is

$$\begin{aligned} \mathrm{MSE}_i &\triangleq \mathbb{E}[|\hat{x}_i - x_i|^2] = |\mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_i - 1|^2 \\ &\quad + \sum_{j=i+1}^{L} |\mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_j - b_{ij}|^2 + \sum_{j=1}^{i-1} |\mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_j|^2 + \|\mathbf{w}_i\|^2 \end{aligned} \tag{4.5}$$

$$\begin{aligned} \mathrm{SINR}_i &\triangleq \frac{\text{desired component}}{\text{undesired component}} \\ &= \frac{|\mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_i|^2}{\sum_{j=i+1}^{L} |\mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_j - b_{ij}|^2 + \sum_{j=1}^{i-1} |\mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_j|^2 + \|\mathbf{w}_i\|^2}. \end{aligned} \tag{4.6}$$

It is trivial to see from (4.5) and (4.6) that to minimize the MSEs and maximize the SINRs the DF coefficients should be

$$b_{ij} = \mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_j, \quad 1 \leq i < j \leq L, \tag{4.7}$$

in which case the (improved) MSEs and SINRs become

$$\mathrm{MSE}_i = |\mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_i - 1|^2 + \sum_{j=1}^{i-1} |\mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_j|^2 + \|\mathbf{w}_i\|^2, \tag{4.8}$$

$$\mathrm{SINR}_i = \frac{|\mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_i|^2}{\sum_{j=1}^{i-1} |\mathbf{w}_i^\dagger \mathbf{H}\mathbf{p}_j|^2 + \|\mathbf{w}_i\|^2}. \tag{4.9}$$

Now we see that due to the feedback component the ith substream is only subject to $i-1$ interference terms. In contrast, for the linear equalizers every substream has $L-1$ interference terms.

Conditioned on a specific channel realization, the mutual information between $\mathbf{x}$ and $\mathbf{y}$ is [148]

$$I(\mathbf{x};\mathbf{y}) = \log|\mathbf{I} + \mathbf{H}\mathbf{P}\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}|. \tag{4.10}$$

With CSI at transmitter (CSIT), the transmitter may choose $\mathbf{P}$ to maximize $I(\mathbf{x};\mathbf{y})$. Denote the singular value decomposition (SVD) $\mathbf{H} = \mathbf{U}_H\boldsymbol{\Sigma}_H\mathbf{V}_H^{\dagger}$, where $\boldsymbol{\Sigma}_H$ is a K by K diagonal matrix with entries $\sigma_{H,i}$, $1 \le i \le K$. We have seen in Chapter 3 (see (3.67)) that the precoder maximizing the mutual information has the following structure:

$$\mathbf{P}\mathbf{P}^{\dagger} = \mathbf{V}_H\mathrm{diag}(\mathbf{p})\mathbf{V}_H^{\dagger}, \tag{4.11}$$

where $\mathbf{p}$ is the power allocation chosen according to the waterfilling algorithm [33]

$$p_i = \left(\mu - \frac{1}{\sigma_{H,i}^2}\right)^{+}, \quad 1 \le i \le K. \tag{4.12}$$

The mutual information corresponding to this precoder is channel capacity.[4] The channel capacity is one of the benchmarks for our MIMO transceiver design. It is important to note that for any semi-unitary matrix $\boldsymbol{\Omega} \in \mathbb{C}^{L\times K}$ with $L \ge K$, the precoder $\mathbf{P} = \mathbf{V}_H\mathrm{diag}(\sqrt{\mathbf{p}})\boldsymbol{\Omega}^{\dagger}$ satisfies (4.11). Hence the semi-unitary matrix $\boldsymbol{\Omega}$ does not influence the channel capacity. However, by designing $\boldsymbol{\Omega}$, we can control MSE_i.

The close approximation of BER as a function of SINR is given in Section 3.1.1.

4.2 Problem Formulation

We now introduce the general problem of optimizing the nonlinear DF MIMO transceiver, i.e., the precoder $\mathbf{P}$, the DFE feed-forward filter

[4] By using the terminology "capacity," we have implicitly assumed that the channel coherent time is so long that a capacity-achieving coding scheme can be applied for almost arbitrarily reliable communication as long as the data rate is less than the maximum mutual information.

matrix $\mathbf{W}$, and the feedback filter matrix $\mathbf{B}$ at the receiver. We will consider two problem formulations: the optimization of a global performance measure subject to the constraint of overall input power and the minimization of the overall input power subject to individual QoS constraints.

4.2.1 Global Measure of Performance

Consider that the performance of the system is measured by a cost function f_0 chosen based on some criterion of practical significance. Then the problem can be formulated as

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W},\mathbf{B}}{\text{minimize}} & f_0(\{\text{MSE}_i\}) \\ \text{subject to} & \text{Tr}\ (\mathbf{P}\mathbf{P}^\dagger) \leq P_0, \end{array} \tag{4.13}$$

where the cost function $f_0(\{\text{MSE}_i\})$ is increasing in each argument. The variables to be optimized are the precoder matrix $\mathbf{P}$, the feed-forward filter matrix $\mathbf{W}$, and the feedback matrix $\mathbf{B}$ (see Figure 4.1).

In Section 4.4, we shall first consider the solution to (4.13) for a general f_0. Then we specialize the cost function to the case where the composite function $f_0 \circ \exp$ is either Schur-concave or Schur-convex, which leads to exceedingly simple solutions.

4.2.2 Individual QoS Constraints

Consider now a formulation in terms of individual QoS constraints. The problem can be formulated as the minimization of the transmitted power subject to the constraints:

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W},\mathbf{B}}{\text{minimize}} & \text{Tr}\ (\mathbf{P}\mathbf{P}^\dagger) \\ \text{subject to} & \text{SINR}_i \geq \gamma_i, \quad 1 \leq i \leq L, \end{array} \tag{4.14}$$

where the constraints ensure that output SINR of the ith detected substream is no less than γ_i.

Both problems (4.13) and (4.14) require joint optimization of the precoder $\mathbf{P}$ and the DFE ($\mathbf{W}$, $\mathbf{B}$). In a somewhat similar vein to the derivations in Chapter 3, it will be shown that the transmitter and receiver parts can be optimized in two decoupled steps. First, the optimum DFE parameters are expressed as functions of channel and precoder, which concentrates out $\mathbf{W}$ and $\mathbf{B}$. In the second step, we only need to optimize the precoder $\mathbf{P}$.

We note that for any f_0, the minimum cost function achieved in (3.13) is lower bounded by the optimum nonlinear design. This claim is quite straightforward to see since the cost function in (4.13) is minimized over three matrices $(\mathbf{P},\mathbf{W},\mathbf{B})$ while (3.13) is minimized over only two $(\mathbf{P},\mathbf{W})$. Similarly, with the same QoS constraint, the nonlinear design requires no more power than the linear one.

4.3 Optimum Decision Feedback Receiver

It has been observed in (4.7) that the optimal feedback matrix must be

$$\mathbf{B} = \mathcal{U}(\mathbf{W}^\dagger \mathbf{H}\mathbf{P}), \tag{4.15}$$

where $\mathcal{U}(\cdot)$ stands for keeping the strictly upper triangular entries of the matrix while setting the others to zero. Now we only need to find the optimum feed-forward filter matrix $\mathbf{W}$ whose ith column minimizes MSE_i given in (4.8). Let $\mathbf{G} \triangleq \mathbf{H}\mathbf{P}$ denote the effective channel matrix. Denoting $\mathbf{G}_i \in \mathbb{C}^{n_R \times i}$ as the submatrix consisting of the first i columns of $\mathbf{G}$ and $\mathbf{g}_i \in \mathbb{C}^{n_R \times 1}$ as the ith column of $\mathbf{G}$, we rewrite (4.8) as

$$\mathrm{MSE}_i = \mathbf{w}_i^\dagger(\mathbf{G}_i\mathbf{G}_i^\dagger + \mathbf{I})\mathbf{w}_i - \mathbf{w}_i^\dagger \mathbf{g}_i - \mathbf{g}_i^\dagger \mathbf{w}_i + 1. \tag{4.16}$$

To minimize MSE_i, let us equate the gradient of MSE_i with respect to $\mathbf{w}_i$ to zero, which yields

$$(\mathbf{G}_i\mathbf{G}_i^\dagger + \mathbf{I})\mathbf{w}_i - \mathbf{g}_i = \mathbf{0}. \tag{4.17}$$

Hence the optimum receiver for the ith substream is

$$\mathbf{w}_i = (\mathbf{G}_i\mathbf{G}_i^\dagger + \mathbf{I})^{-1}\mathbf{g}_i, \quad 1 \le i \le L. \tag{4.18}$$

Inserting $\mathbf{W} = [\mathbf{w}_1,\ldots,\mathbf{w}_L]$ into (4.15) yields $\mathbf{B}$, which constitutes the optimum DFE.

The following result provides an alternative expression for the optimum DFE, which is clearly computationally more efficient [52, 61].

Lemma 4.1. Let the QR decomposition of the augmented matrix be[5]

$$\mathbf{G}_a \triangleq \begin{bmatrix} \mathbf{G} \\ \mathbf{I}_L \end{bmatrix}_{(n_R+L)\times L} = \mathbf{QR}. \tag{4.19}$$

Partition $\mathbf{Q}$ into

$$\mathbf{Q} = \begin{bmatrix} \bar{\mathbf{Q}} \\ \underline{\mathbf{Q}} \end{bmatrix}, \tag{4.20}$$

where $\bar{\mathbf{Q}} \in \mathbb{C}^{n_R \times L}$ and $\underline{\mathbf{Q}} \in \mathbb{C}^{L\times L}$. The optimum feed-forward and feedback matrices that minimize the MSEs are

$$\mathbf{W} = \bar{\mathbf{Q}}\mathbf{D}_R^{-1}, \quad \text{and} \quad \mathbf{B} = \mathbf{D}_R^{-1}\mathbf{R} - \mathbf{I}, \tag{4.21}$$

where $\mathbf{D}_R$ is a diagonal matrix with the same diagonal as $\mathbf{R}$. The resulting MSE matrix is diagonal:

$$\mathbf{E} \triangleq \mathbb{E}[(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^\dagger] = \mathbf{D}_R^{-2}. \tag{4.22}$$

Proof. See Appendix 4.B. □

Since both $\mathbf{W}$ and $\mathbf{B}$ are designed to minimize the MSEs, the optimum DFE is also called MMSE–DFE. It can be seen from Appendix 4.B that given an effective channel $\mathbf{G}$ the MMSE–DFE involves no tradeoff among the MSEs. Indeed, as we see from (4.8), MSE_i does not depend on $\mathbf{w}_j$ for $j \neq i$.

Observe that the MSE matrix is diagonal for any channel realization. In contrast, the MSE matrix of the linear Wiener receiver is $(\mathbf{I} + \mathbf{G}^\dagger\mathbf{G})^{-1}$ which is diagonal only if $\mathbf{G}$ has orthogonal columns (see Appendix 4.A). In fact, if the columns of $\mathbf{G}$ are orthogonal, $\mathbf{R}$ becomes diagonal and it follows from (4.21) that $\mathbf{B} = \mathbf{0}$. In this case the DFE degenerates into a linear receiver.

Similar to the linear receivers, which include both the MMSE (Wiener) receiver andthe ZF receiver, the DFE also has two forms:

[5] To make the QR decomposition unique, the diagonal of $\mathbf{R}$ is set to be positive real.

the MMSE–DFE and the ZF–DFE. By (4.117), the lower triangular entries of $\mathbf{W}^\dagger\mathbf{G}$ are nonzero, which means that the ith substream is subject to the interference leakage from the $j(< i)$th substream. The ZF–DFE imposes the constraint that the interference leakage be zero. Let $\mathbf{G} = \mathbf{Q}_G\mathbf{R}_G$ be the QR decomposition. The feed-forward and feedback matrices of ZF–DFE are [72]

$$\mathbf{W}^{\text{zf}-\text{dfe}} = \mathbf{Q}_G\mathbf{D}_{R_G}^{-1} \tag{4.23}$$

and

$$\mathbf{B}^{\text{zf}-\text{dfe}} = \mathbf{D}_{R_G}^{-1}\mathbf{R}_G - \mathbf{I}, \tag{4.24}$$

where $\mathbf{D}_{R_G}$ is the diagonal of $\mathbf{R}_G$. The MSE matrix of ZF–DFE is

$$\mathbf{E}^{\text{zf}-\text{dfe}} = \mathbf{D}_{R_G}^{-2}. \tag{4.25}$$

It is worthwhile emphasizing that $\mathbf{R}_G$ can only have up to K (the rank of $\mathbf{H}$) nonzero diagonal entries, since the rank of the effective channel $\mathbf{HP}$ is not greater than K. Consequently, the ZF–DFE can only recover $L \leq K$ substreams. The MMSE–DFE, however, can handle any number of substreams since the rank of $\mathbf{G}_a$ defined in (4.19) is not limited by the dimensionality of $\mathbf{H}$.

Some comments on the DFE are in order. In fact, the optimum feedback matrix $\mathbf{B}$ is chosen such that the substreams detected in the previous steps are eliminated completely from the received signal vector. Hence the DFE shown in Figure 4.2 can actually be equivalently represented by the more straightforward scheme shown in Figure 4.3. At each step, a linear receiver (MMSE or ZF) is applied to detect *one* substream, and then the detected substream is removed from the data vector so that the number of interferences in the next step detection is reduced by one (the detection is assumed correct, i.e., $\tilde{x}_i = x_i$). In contrast, for the linear receiver given in Section 3.3, each substream is estimated by regarding *all* the other substreams as interference. Given the assumption of perfect interference cancelation, the DFE should yield lower MSEs than a linear receiver for all except the first detected substream. (At the first step of successive detection, the DFE, just like the linear receiver, has $L - 1$ interference terms to suppress.)

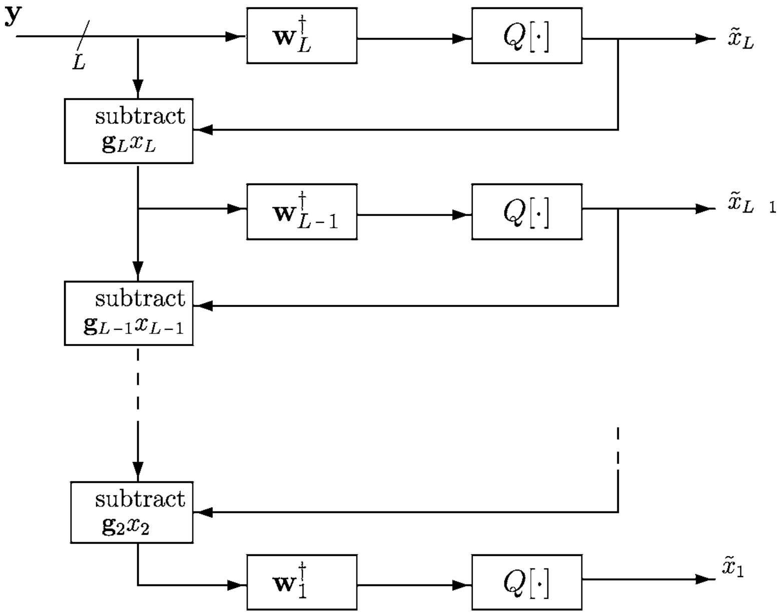


Fig. 4.3 Another representation of DFE.

4.3.1 Relationship Among Different Measures of Performance

Since the DFE applies linear MMSE or ZF successively, the SINR and MSE of the substreams at the output of DFE are related in the same way as those of the linear receiver (see (3.33)):

$$\mathrm{SINR}_i = \begin{cases} \frac{1}{\mathrm{MSE}_i} - 1 & \text{for MMSE–DFE} \\ \frac{1}{\mathrm{MSE}_i} & \text{for ZF–DFE.} \end{cases} \tag{4.26}$$

For MMSE–DFE, it follows from (4.22) and (4.26) that

$$\mathrm{SINR}_i = [\mathbf{R}]_{ii}^2 - 1, \tag{4.27}$$

where $[\mathbf{R}]_{ii}$ is the ith diagonal element of $\mathbf{R}$. With a sufficiently long capacity-achieving code, the data rate of the ith substream obtained by MMSE–DFE can be up to

$$R_i = \log(1 + \mathrm{SINR}_i) = \log[\mathbf{R}]_{ii}^2. \tag{4.28}$$

The sum rate is

$$\begin{aligned}\sum_{i=1}^{L}\log[\mathbf{R}]_{ii}^2 &= \log|\mathbf{R}^\dagger\mathbf{R}| = \log|\mathbf{G}_a^\dagger\mathbf{G}_a|, \\ \text{(see (4.19))} &= \log|\mathbf{I}+\mathbf{HPP}^\dagger\mathbf{H}^\dagger|, \end{aligned} \tag{4.29}$$

which is the mutual information given in (4.10). Hence we may regard the MMSE–DFE as a channel partitioner, which decomposes, in an information lossless manner, a MIMO channel into L scalar subchannels with capacity $\log[\mathbf{R}]_{ii}^2$ for $1 \le i \le L$ [156].

It follows from (4.26) and (4.25) that the ZF–DFE decomposes the MIMO channel into L $(L \le K)$ subchannels with output SNRs[6]

$$\mathrm{SNR}_i^{\mathrm{zf-dfe}} = [\mathbf{R}_G]_{ii}^2. \tag{4.30}$$

It is well-known that the MMSE–DFE has superior performance to the ZF–DFE. Hence we will focus on the MIMO transceiver design based on the MMSE–DFE.

We have seen that the system performance measures, including the MSEs, the output SINRs, and the mutual information of the subchannels, are all directly determined by the diagonal of $\mathbf{R}$ (or $\mathbf{R}_G$ in the ZF–DFE case) which in turn depends on the linear precoder $\mathbf{P}$. Given the CSIT, the transmitter can design $\mathbf{P}$ to transform the channel $\mathbf{H}$ to the effective channel $\mathbf{G} = \mathbf{HP}$ and hence control the performance of each substream. Since the DFE parameters $\mathbf{W}$ and $\mathbf{B}$ have been concentrated out of the optimization problem, we now focus on designing the transmitter $\mathbf{P}$.

4.4 Optimum Transmitter with Global Measure of Performance

This section deals with the design of the optimum transmitter $\mathbf{P}$ according to problem (4.13). While the linear MIMO transceiver design in Chapter 3 relies on the *additive* majorization theory, the nonlinear DF transceiver design relies on the concept of *multiplicative* majorization

[6] The ZF–DFE yields substreams free of interference. Thus, we use the terminology SNR rather than SINR.

(see Definition 2.6) and the recent matrix decomposition called the generalized triangular decomposition (GTD).To facilitate our discussion, we reproduce the main idea here. The GTD is detailed in Appendix B.1.

Theorem 4.2. (GTD Theorem) Let $\mathbf{H} \in \mathbb{C}^{m \times n}$ be a matrix with rank K and singular values $\sigma_{H,1} \geq \sigma_{H,2} \geq \cdots \geq \sigma_{H,K} > 0$.[7] There exists an upper triangular matrix $\mathbf{R} \in \mathbb{C}^{K \times K}$ and semi-unitary matrices $\mathbf{Q}$ and $\mathbf{P}$ such that

$$\mathbf{H} = \mathbf{QRP}^{\dagger}$$

if and only if the diagonal elements of $\mathbf{R}$ satisfy $|\mathbf{r}| \prec_{\times} \boldsymbol{\sigma}_H$,[8] where $|\mathbf{r}|$ is a vector with the absolute values of $\mathbf{r}$ element-wise.

It follows from this theorem that there exists the decomposition

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

where the upper triangular matrix $\mathbf{R}$ has equal diagonal elements [69] of value

$$[\mathbf{R}]_{ii} = \bar{\sigma}_H \triangleq \left(\prod_{i=1}^{K} \sigma_{H,i} \right)^{\frac{1}{K}}, \quad 1 \leq i \leq K. \tag{4.31}$$

The existence of such a decomposition is clear by noting that $\bar{\sigma}_H \mathbf{1} \prec_{\times} \boldsymbol{\sigma}_H$.[9]

We start with the problem with a general cost function f_0 and then particularize for two classes of functions.

4.4.1 General Solution for Arbitrary Cost Functions

Recall from (4.22) that the MSEs with the MMSE–DFE are

$$\mathrm{MSE}_i = [\mathbf{R}]_{ii}^{-2} \quad \text{with } \mathbf{R} \text{ from } \quad \begin{pmatrix} \mathbf{HP} \\ \mathbf{I} \end{pmatrix} = \mathbf{QR}. \tag{4.32}$$

[7] Throughout this chapter, the singular values are in non-increasing order.

[8] We have adopted the notation used in the original paper [71]. The notations $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{P}$ are certainly *not* the ones given in (4.19) or the precoder matrix.

[9] In the [69], we call such a decomposition geometric mean decomposition as the diagonal elements of $\mathbf{R}$ equal to the geometric mean of the singular values of $\mathbf{H}$.

Therefore the problem (4.13) can be reformulated as

$$\begin{array}{ll} \underset{\mathbf{P}}{\text{minimize}} & f_0\left(\left\{[\mathbf{R}]_{ii}^{-2}\right\}\right) \\ \text{subject to} & \begin{pmatrix} \mathbf{HP} \\ \mathbf{I} \end{pmatrix} = \mathbf{QR} \\ & \text{Tr}(\mathbf{PP}^\dagger) \le P_0. \end{array} \tag{4.33}$$

The problem (4.33) is a complicated matrix optimization problem. However, it can be transformed into a much simplified optimization problem with scalar-valued variables, as summarized below.

Theorem 4.3. The solution to (4.33) has the form:

$$\mathbf{P} = \mathbf{V}_H \text{diag}(\sqrt{\mathbf{p}})\mathbf{\Omega}^\dagger, \tag{4.34}$$

where $\mathbf{V}_H \in \mathbb{C}^{n_T \times K}$ follows from the SVD $\mathbf{H} = \mathbf{U}_H \mathbf{\Sigma}_H \mathbf{V}_H^\dagger$, and the power allocation $\mathbf{p} \in \mathbb{R}_+^K$ is the solution to the optimization problem:

$$\begin{array}{ll} \underset{\mathbf{p},\, \{[\mathbf{R}]_{ii}\}}{\text{minimize}} & f_0\left(\left\{[\mathbf{R}]_{ii}^{-2}\right\}\right) \\ \text{subject to} & \left([\mathbf{R}]_{11}^2, \ldots, [\mathbf{R}]_{LL}^2\right) \prec_\times \left(\left\{1 + \sigma_{H,i}^2 p_i\right\}_{i=1}^K, \underbrace{1, \ldots, 1}_{L-K}\right) \\ & \mathbf{1}^T\mathbf{p} \le P_0 \\ & \mathbf{p} \ge \mathbf{0}, \end{array} \tag{4.35}$$

and the semi-unitary matrix $\mathbf{\Omega} \in \mathbb{C}^{L \times K}$ is obtained such that the matrix $\mathbf{R}$ in the QR decomposition

$$\begin{pmatrix} \mathbf{HP} \\ \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{U}_H \mathbf{\Sigma}_H \text{diag}(\sqrt{\mathbf{p}})\mathbf{\Omega}^\dagger \\ \mathbf{I} \end{pmatrix} = \mathbf{QR} \tag{4.36}$$

has diagonal elements $\{[\mathbf{R}]_{ii}\}$ being the solution to (4.35) (which can be computed with the procedure described later in Table 4.1).

Proof. The proof is rather technical and we relegate it to Appendix 4.C. However, a fair understanding of the proof is needed to follow the procedure of the precoder design given in Table 4.1 □

Table 4.1 Procedure of calculating DF transceiver with global performance measure.

Step	Operation
1	Compute the SVD $\mathbf{H} = \mathbf{U}_H \boldsymbol{\Sigma}_H \mathbf{V}_H^\dagger$.
2	Obtain $\mathbf{p}$ and $\{[\mathbf{R}]_{ii}\}$ by solving (4.35).
3	Calculate the generalized triangular decomposition $\begin{bmatrix} \mathbf{U}_H \boldsymbol{\Sigma}_H [\text{diag}(\sqrt{\mathbf{p}}) : \mathbf{0}_{K \times (L-K)}] \\ \mathbf{I}_L \end{bmatrix} = \mathbf{Q}_J \mathbf{R} \mathbf{P}_J^\dagger,$ where $\mathbf{R}$ has diagonal elements $\{[\mathbf{R}]_{ii}\}$ from step 2.
4	Obtain the precoder as $\mathbf{P} = \mathbf{V}_H [\text{diag}(\sqrt{\mathbf{p}}) \vdots \mathbf{0}_{K \times (L-K)}] \mathbf{P}_J$.
5	Compute the DFE $\mathbf{B} = \mathbf{D}_R^{-1} \mathbf{R}$ and $\mathbf{W} = \bar{\mathbf{Q}} \mathbf{D}_R^{-1}$, where $\bar{\mathbf{Q}}$ consists of the first n_R rows of $\mathbf{Q}_J$.

The derivations given in Appendices 4.C and 4.D present the "roadmap" for designing the optimum transceiver with global performance measure, which is summarized in Table 4.1.

Observe that Theorem 4.4 does not restrict L in any way and we can have $L \geq K$. In the case that $L < K$, the majorization constraint in (4.35) reduces to

$$\left([\mathbf{R}]_{11}^2, \ldots, [\mathbf{R}]_{LL}^2\right) \prec_\times \left(1 + \sigma_{H,1}^2 p_1, \ldots, 1 + \sigma_{H,L}^2 p_L\right).$$

Distribution of rates among substreams: Recall from Section 4.3.1 that the MMSE–DFE decomposes a MIMO channel into multiple subchannels with rates

$$R_i = \log [\mathbf{R}]_{ii}^2, \quad 1 \leq i \leq L. \tag{4.37}$$

Now, from Lemma 4.9 in Appendix 4.C, we know that as $\boldsymbol{\Omega}$ goes over the Stiefel manifold [68]

$$\mathcal{S}(L;K) \triangleq \left\{ \mathbf{Q} \in \mathbb{C}^{L \times K} : \mathbf{Q}^\dagger \mathbf{Q} = \mathbf{I} \right\}, \tag{4.38}$$

the achievable set of the diagonal of $\mathbf{R}$ is $\{\mathbf{r} \in \mathbb{C}^L : |\mathbf{r}| \prec_\times \boldsymbol{\sigma}_{G_a}\}$. Hence, we have

$$(R_1, \ldots, R_L) \prec_+ (C_1, \ldots, C_L), \tag{4.39}$$

where

$$C_i \triangleq \log \sigma_{G_a,i}^2 = \begin{cases} \log(1 + \sigma_{H,i}^2 p_i) & 1 \leq i \leq K \\ 0 & K+1 \leq i \leq L, \end{cases} \tag{4.40}$$

and $C_1, \ldots, C_K$ is the capacities of the K eigen-subchannels for some given power allocation. Therefore by changing $\boldsymbol{\Omega}$ within the Stiefel manifold, we may achieve a flexible rate distribution (always with the same total rate). We illustrate this point by the following toy example. Recall that the set $\{\mathbf{x} \in \mathbb{R}_+^n \,|\, \mathbf{x} \prec_+ \mathbf{y}\}$ is the convex hull spanned by the $n!$ points which are formed from the permutations of the elements of $\mathbf{y}$ [97, pp. 9]. Consider a MIMO channel $\mathbf{H}$ with rank $K = 3$. We assume that the capacities of the 3 eigen-subchannels obtained via SVD and power allocation are $C_1 \geq C_2 \geq C_3$. For $L = 3$, the nonlinear DFE based transceiver can decompose the MIMO channel into 3 subchannels with capacities R_1, R_2, and R_3 if and only if the vector $[R_1, R_2, R_3]^T$ lies on the convex hull

$$\text{Co}\left\{\begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}, \begin{pmatrix} C_1 \\ C_3 \\ C_2 \end{pmatrix}, \begin{pmatrix} C_2 \\ C_1 \\ C_3 \end{pmatrix}, \begin{pmatrix} C_2 \\ C_3 \\ C_1 \end{pmatrix}, \begin{pmatrix} C_3 \\ C_2 \\ C_1 \end{pmatrix}, \begin{pmatrix} C_3 \\ C_1 \\ C_2 \end{pmatrix}\right\}, \tag{4.41}$$

where "Co" stands for the convex hull defined as

$$\text{Co}\{S\} = \{\theta_1 x_1 + \cdots + \theta_K x_K | x_i \in S, \theta_i \geq 0, \theta_1 + \cdots + \theta_K = 1\}. \tag{4.42}$$

The gray area in Figure 4.4 shows the convex hull of (4.41) with $C_1 = 3$, $C_2 = 2$, and $C_3 = 1$. In this case, the 6 vertices lie in the 2D plane $\{\mathbf{x} : \sum_{i=1}^{3} x_i = 6\}$. From this example, we see that the nonlinear DF transceiver is much more flexible than the linear design which can only achieve the corner points. The rate tuple at the center of the convex hull is of special interests as we shall see soon.

4.4.2 Schur-Convex and Schur-Concave Cost Functions

As we have seen in the previous section, the solution to the general problem with an arbitrary cost function has a nice structure. In this section, we further specialize the cost function to the case where the composite function $f_0 \circ \exp : \mathbb{R}^L \to \mathbb{R}$ is either Schur-convex or Schur-concave. The composite function $f_0 \circ \exp$ is defined as

$$f_0 \circ \exp(\mathbf{x}) \triangleq f_0(e^{x_1}, e^{x_2}, \ldots, e^{x_L}). \tag{4.43}$$

For such special cost functions, the nonlinear MIMO transceiver turns out to be exceedingly simple.

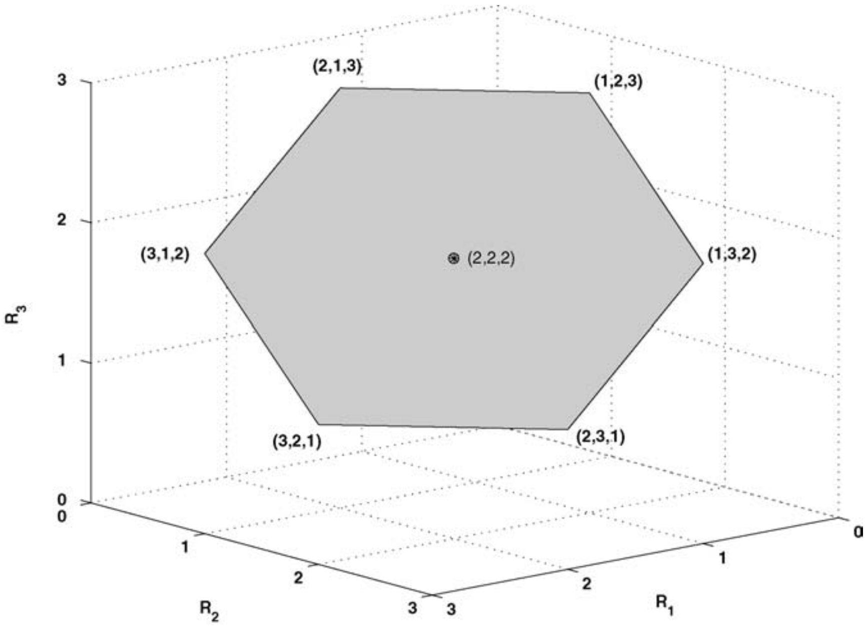


Fig. 4.4 Illustration of the achievable rate region by choosing different $\mathbf{\Omega}$ (for a given power allocation). We assume $L = 3$, $C_1 = 3$, $C_2 = 2$, and $C_3 = 1$.
©2006 IEEE. Reprinted with permission from IEEE.

Theorem 4.4. An optimal solution $\mathbf{P}$ of the problem (4.33), where $f_0 : \mathbb{R}^L \to \mathbb{R}$ is a function increasing in each argument, can be characterized as follows:

- If $f_0 \circ \exp$ is Schur-concave on $\mathcal{D}_L \triangleq \{\mathbf{x} \in \mathbb{R}^L : x_1 \geq \cdots \geq x_L\}$ (assuming without loss of generality that it is minimized when the arguments are in decreasing order), then

$$\mathbf{P} = \mathbf{V}_H \text{diag}(\sqrt{\mathbf{p}}), \tag{4.44}$$

where $\mathbf{V}_H$ is the right singular vector matrix of $\mathbf{H}$ and $\mathbf{p} \in \mathbb{R}_+^K$ is the solution to

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & f_0\left(\left\{\frac{1}{1+\sigma_{H,i}^2 p_i}\right\}\right) \\ \text{subject to} & \mathbf{1}^T\mathbf{p} \leq P_0 \\ & \mathbf{p} \geq \mathbf{0}, \end{array} \tag{4.45}$$

where $\sigma_{H,i}$ is the ith largest singular values of $\mathbf{H}$. In this case $L \leq K$, otherwise some argument of $f_0 \circ \exp$ should be set to zero.

- If $f_0 \circ$ expis Schur-convex on $\mathbb{R}^L$, then

$$\mathbf{P} = \mathbf{V}_H \text{diag}(\sqrt{\mathbf{p}})\boldsymbol{\Omega}^\dagger, \tag{4.46}$$

where $\mathbf{V}_H$ is as before and $\mathbf{p} \in \mathbb{R}_+^K$ is given by the standard waterfilling power allocation

$$p_i = \left(\mu - \frac{1}{\sigma_{H,i}^2}\right)^+, \quad 1 \leq i \leq K, \tag{4.47}$$

where μ is chosen such that $\sum_{i=1}^K p_i = P_0$. The semi-unitary matrix $\boldsymbol{\Omega}$ is chosen so that the QR decomposition $\begin{bmatrix} \mathbf{HP} \\ \mathbf{I} \end{bmatrix} = \mathbf{QR}$ yields $\mathbf{R}$ with equal diagonal elements. In this case, L is not constrained by the dimensionality of $\mathbf{H}$.

Proof. Based on Theorem 4.3, we start with the optimization problem (4.35)

$$\begin{array}{ll} \underset{\mathbf{p},\, \{[\mathbf{R}]_{ii}\}}{\text{minimize}} & f_0\left(\left\{[\mathbf{R}]_{ii}^{-2}\right\}\right) \\ \text{subject to} & \left([\mathbf{R}]_1^2,\ldots,[\mathbf{R}]_L^2\right) \prec_\times \left(\left\{1+\sigma_{H,i}^2 p_i\right\}_{i=1}^K, \underbrace{1,\ldots,1}_{L-K}\right) \\ & \mathbf{1}^T\mathbf{p} \leq P_0 \\ & \mathbf{p} \geq \mathbf{0}. \end{array} \tag{4.48}$$

Let $R_i \triangleq \log[\mathbf{R}]_{ii}^2$ (R_i stands for the achievable rate of the ith subchannel as shown in (4.28)). Problem (4.48) can be rewritten as

$$\begin{array}{ll} \underset{\mathbf{p},\, \{R_i\}}{\text{minimize}} & f_0\left(\left\{e^{-R_i}\right\}\right) \\ \text{subject to} & (R_1,\ldots,R_L) \prec_+ \left(\left\{\log(1+\sigma_{H,i}^2 p_i)\right\}_{i=1}^K, 0,\ldots,0\right) \\ & \mathbf{1}^T\mathbf{p} \leq P_0 \\ & \mathbf{p} \geq \mathbf{0}. \end{array} \tag{4.49}$$

Note that for any function f, $f(\mathbf{x})$, and $f(-\mathbf{x})$ have the same Schur-convexity/concavity. Thus, if $f_0 \circ \exp$ is Schur-concave, then $f_0\left(\left\{e^{-R_i}\right\}\right)$ is also a Schur-concave function of $(R_1, \ldots, R_L)$. Hence, the solution to (4.49) occurs when

$$R_i = \begin{cases} \log(1 + \sigma_{H,i}^2 p_i) & 1 \leq i \leq K \\ 0 & K+1 \leq i \leq L, \end{cases} \tag{4.50}$$

which corresponds to $\mathbf{\Omega} = [\mathbf{I}_K : \mathbf{0}_{K\times(L-K)}]^T$. For $L > K$, there are subchannels with capacity $R_i = 0$. Assuming for simplicity of notation that $L \leq K$, (4.49) can be simplified to be

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & f_0\left(\left\{\frac{1}{1+\sigma_{H,i}^2 p_i}\right\}\right) \\ \text{subject to} & \mathbf{1}^T\mathbf{p} \leq P_0 \\ & \mathbf{p} \geq \mathbf{0}. \end{array} \tag{4.51}$$

If $f_0 \circ \exp$ is Schur-convex, $f_0\left(\left\{e^{-R_i}\right\}\right)$ is a Schur-convex function of $[R_1, \ldots, R_L]$. Hence the solution to (4.49) occurs when

$$R_i = \frac{1}{L}\sum_{i=1}^{K} \log(1 + \sigma_{H,i}^2 p_i) \quad \text{for}\, 1 \leq i \leq L. \tag{4.52}$$

Using the property that f_0 is an increasing function of each argument, we further simplify (4.49) to be

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{maximize}} & \sum_{i=1}^{K} \log(1 + \sigma_{H,i}^2 p_i) \\ \text{subject to} & \mathbf{1}^T\mathbf{p} \leq P_0 \\ & \mathbf{p} \geq \mathbf{0}. \end{array} \tag{4.53}$$

The solution is the standard waterfilling power allocation given in (4.47). The semi-unitary matrix $\mathbf{\Omega}$ is chosen such that the diagonal elements of $\mathbf{R}$ in

$$\begin{pmatrix} \mathbf{HP} \\ \mathbf{I} \end{pmatrix} = \mathbf{QR} \tag{4.54}$$

are equal, i.e.,

$$[\mathbf{R}]_{ii} = \exp(R_i) = \left(\prod_{k=1}^{K} (1 + \sigma_{H,i}^2 p_i)\right)^{\frac{1}{L}}, \quad \text{for } 1 \leq i \leq L. \tag{4.55}$$

According to Lemma 4.9, such an $\mathbf{\Omega}$ exists. □

4.4.2.1 Linear vs. Decision Feedback MIMO Transceivers

For the first case in Theorem 4.4, the precoder $\mathbf{P} = \mathbf{V}_H \mathrm{diag}(\sqrt{\mathbf{p}})$ orthogonalizes the MIMO channel into multiple eigen-subchannels and the power allocation applied to the subchannels is given by the solution to (4.51). In this case, the columns of $\mathbf{HP}$ are orthogonal and the QR decomposition $\begin{pmatrix} \mathbf{HP} \\ \mathbf{I} \end{pmatrix} = \mathbf{QR}$ yields diagonal $\mathbf{R}$. Hence, the feedback matrix $\mathbf{B} = \mathbf{D}_R^{-1}\mathbf{R} - \mathbf{I} = \mathbf{0}$. The optimal DFE therefore degenerates to linear diagonal transmission. In fact, this transceiver design is the same as the linear design corresponding to the first case in Theorem 3.1. However, the condition here is that the composite function $f_0 \circ \exp$ be Schur-concave, while the condition in Theorem 3.1 is that f_0 be Schur-concave. The relationship between the two conditions is given in Lemma 2.12: $f_0 \circ \exp$ being Schur-concave implies that f_0 is Schur-concave, but not vice versa. Define $\mathcal{S}_{\text{nonlin,diag}}$ ($\mathcal{S}_{\text{lin,diag}}$) as the set of cost functions such that the optimum nonlinear (linear) transceiver degenerates to diagonal transmission. We have

$$\mathcal{S}_{\text{nonlin,diag}} \subset \mathcal{S}_{\text{lin,diag}}.$$

In the second case of Theorem 4.4, for *any* cost function such that the composite $f_0 \circ \exp$ is Schur-convex, the optimum *nonlinear* MIMO transceiver design is the same. Lemma 2.11 sheds light on the relationship between Theorems 4.4 and 3.1: if f_0 is Schur-convex, which implies that the optimum linear transceiver uses non-diagonal transmission and all the substreams have the same MSE, then $f_0 \circ \exp$ is also Schur-convex, which means that the optimum nonlinear design also uses non-diagonal transmission and all the substreams have the same MSE, but not vice versa. Define $\mathcal{S}_{\text{nonlin,nondiag}}$ ($\mathcal{S}_{\text{lin,nondiag}}$) as the set of cost functions such that the optimum nonlinear (linear) transceiver applies the non-diagonal transmission. We have

$$\mathcal{S}_{\text{lin,nondiag}} \subset \mathcal{S}_{\text{nonlin,nondiag}}.$$

The linear and nonlinear design may have quite different performance when the cost function is not in $\mathcal{S}_{\text{lin,nondiag}}$. For instance, in the second case of Theorem 4.4 ($f_0 \in \mathcal{S}_{\text{nonlin,nondiag}}$), the nonlinear design applies the standard capacity-achieving waterfilling power allocation,

$$p_i = \left(\mu - \sigma_{H,i}^{-2}\right)^+ . \tag{4.56}$$

For the nondiagonal linear design ($f_0 \in \mathcal{S}_{\text{lin,nondiag}}$), however, the power allocation is the MMSE waterfilling solution (see (3.84) where $\lambda_{H,i} = \sigma_{H,i}^2$)

$$p_i = \left(\mu \sigma_{H,i}^{-1} - \sigma_{H,i}^{-2}\right)^+ , \tag{4.57}$$

which yields smaller mutual information.

From (4.55) and (4.27), we see that the optimum nonlinear design corresponding to the second case of Theorem 4.4 yields L subchannels with the same output SINR:

$$\text{SINR}_i = \left[\prod_{i=1}^{K}(1 + \sigma_{H,i}^2 p_i)\right]^{\frac{1}{L}} - 1 \tag{4.58}$$

and with rate

$$R_i = \frac{1}{L}\sum_{i=1}^{K}\log(1 + p_i \sigma_{H,i}^2) = \frac{C}{L}, \tag{4.59}$$

where C is the capacity of the MIMO channel. We refer to this MIMO transceiver design as the uniform channel decomposition (UCD) scheme, as it uniformly decomposes, in an information lossless manner, a MIMO channel into L *identical* subchannels. As we will see soon, the UCD scheme is the optimum solution to a wide class of cost functions, including minimum average BER.

Because of its special importance, we summarize the procedure of UCD in Table 4.2, which are different from Table 4.1 in Steps 2 and 3; in step 2 the standard waterfilling power allocation is applied and in step 3 we constrain the diagonal elements of $\mathbf{R}$ to be the same. The paper [73] also presents fast implementation of Steps 3–5.

4.4.3 Examples of Cost Functions with Schur-Convex Composite $f_0 \circ \exp$

In this section, we study some examples of cost function for which $f_0 \circ \exp$ is Schur-convex and hence the optimum design is the UCD

Table 4.2 Procedure for designing UCD: The solution for Schur-convex composite $f_0 \circ \exp$.

Step	Operation
1	Calculate SVD $\mathbf{H} = \mathbf{U}_H \mathbf{\Sigma}_H \mathbf{V}_H^\dagger$.
2	Obtain the water filling power allocation $p_i = \left(\mu - \frac{1}{\sigma_{H,i}^2}\right)^+$.
3	Calculate a special GTD $\mathbf{J} = \begin{bmatrix} \mathbf{U}_H \mathbf{\Sigma}_H [\mathrm{diag}(\sqrt{\mathbf{p}}) \vdots \mathbf{0}_{K\times(L-K)}] \\ \mathbf{I}_L \end{bmatrix} = \mathbf{Q}_J \mathbf{R} \mathbf{P}_J^\dagger$, where the diagonal elements of $\mathbf{R}$ are equal.
4	Obtain precoder $\mathbf{P} = \mathbf{V}_H [\mathrm{diag}(\sqrt{\mathbf{p}}) : \mathbf{0}_{K\times(L-K)}] \mathbf{P}_J$.
5	Compute DFE $\mathbf{W} = \bar{\mathbf{Q}} \mathbf{D}_R^{-1}$ and $\mathbf{B} = \mathbf{D}_R^{-1} \mathbf{R} - \mathbf{I}$.

which decomposes the MIMO channel into L identical subchannels. As noted earlier, $f_0 \circ \exp$ is Schur-convex if f_0 is Schur-convex and increasing in each argument. Therefore, all the Schur-convex functions listed in Table 3.2 lead to the same solution: UCD. We list them in the following for the ease of reference:

- Minimization of the maximum of the MSEs.
- Maximization of the harmonic mean of the SINRs.
- Maximization of the minimum of the SINRs.
- Minimization of the average BER.
- Minimization of the maximum of the BERs.

Interestingly, even for some Schur-*concave* functions f_0, the composite function $f_0 \circ \exp$ may be Schur-*convex*. This is illustrated with the following examples from the list of Schur-concave function listed in Table 3.1.

4.4.3.1 Minimization of the Sum of MSEs

The cost function is

$$f_0(\{\mathrm{MSE}_i\}) = \sum_{i=1}^{L} \mathrm{MSE}_i, \tag{4.60}$$

which is both Schur-concave and Schur-convex. The composite function $f_0 \circ \exp(\mathbf{x}) = \sum_i \exp(x_i)$ is Schur-convex (but not Schur-concave). By

Theorem 4.4, the optimum solution is the UCD scheme. Indeed, by the relationship $\mathrm{MSE}_i = [\mathbf{R}]_{ii}^{-2} = \exp(-R_i)$, the cost function

$$f_0(\{\exp(-R_i)\}) = \sum_{i=1}^{L} \exp(-R_i), \tag{4.61}$$

is minimized when $R_i = C/L$, where C is the MIMO channel capacity.

4.4.3.2 Minimization of the Determinant of the MSE Matrix

As noted in (4.22), the MSE matrix of the DFE is always diagonal. Hence the cost function is

$$f_0(\{\mathrm{MSE}_i\}) = \prod_{i=1}^{L} \mathrm{MSE}_i. \tag{4.62}$$

By Example 2.4, although $f_0(x)$ is a Schur-concave function, the composite $f_0 \circ \exp(\mathbf{x}) = \exp(\sum_i x_i)$ is both Schur-convex and Schur-concave. Hence the solution is not unique. Both the UCD and diagonal transmission are optimal. In fact, for any semi-unitary matrix $\mathbf{\Omega}$, the precoder of the form $\mathbf{P} = \mathbf{V}_H \mathrm{diag}(\sqrt{\mathbf{p}})\mathbf{\Omega}^\dagger$ with $\mathbf{p}$ obtained via the standard waterfilling power allocation is the optimum solution to this problem. (This fact is reflected in cf. Figure 4.4 where any rate tuple within the convex hull corresponds to the same determinant MSE.)

4.4.3.3 Maximization of the Mutual Information

By (4.29) and (4.22), the channel mutual information is $I = \log|\mathbf{E}|$. Hence to maximize the mutual information is equivalent to minimize the determinant of the MSE matrix.

4.4.3.4 Maximization of the Product of SINRs

Since $\mathrm{SINR}_i = 1/\mathrm{MSE}_i - 1$, the objective function is to maximize $\prod_{i=1}^{L}(1/\mathrm{MSE}_i - 1)$. Hence the cost function to minimize is

$$f_0(\{\mathrm{MSE}_i\}) = -\prod_{i=1}^{L}\left(\frac{1}{\mathrm{MSE}_i} - 1\right), \tag{4.63}$$

which is shown in Lemma 3.5 to be Schur-concave. The composite function $f_0 \circ \exp(\mathbf{x}) = -\prod_{i=1}^{L}(\exp(-x_i) - 1)$ is, however, Schur-convex. To prove this point, we note that $\sum_{i=1}^{L}\log(\exp(-x_i) - 1)$ is Schur-concave, as the second order derivative:

$$\frac{\partial^2 \log(\exp(x) - 1)}{\partial x^2} = -\frac{\exp(x)}{(\exp(x) - 1)^2} \leq 0. \tag{4.64}$$

According to Lemma 2.5, $\prod_{i=1}^{L}(\exp(-x_i) - 1) = \exp\left(\sum_{i=1}^{L}\log(\exp(-x_i) - 1)\right)$ is Schur-concave. Therefore $f_0 \circ \exp(\mathbf{x}) = -\prod_{i=1}^{L}(\exp(-x_i) - 1)$ is Schur-convex.

All the previous examples are summarized in Table 4.3.

4.4.4 Examples of Cost Functions with Schur-Concave Composite $f_0 \circ \exp$

We present two examples from Table 3.1 for which $f_0 \circ \exp$ is Schur-concave and the nonlinear design degenerates to linear diagonal transmission.

4.4.4.1 Minimization of the Exponentially Weighted Product of MSEs

The cost function is

$$f_0(\{\mathrm{MSE}_i\}) = \prod_{i=1}^{L}\mathrm{MSE}_i^{\alpha_i}. \tag{4.65}$$

Table 4.3 Examples of problem formulations that can be rewritten as minimization of a Schur-convex cost function of $(R_1, \ldots, R_L)$.

Problem	Optimal solution
Minim. maximum of MSEs	Uniform channel decomposition
Maxim. minimum of SINRs	
Maxim. harmonic mean of SINRs	
Minim. average BER (equal constellations)	
Minim. maximum of BERs	
Minim. sum of MSEs	(See Table 4.2)
Minim. det$(\mathbf{E})$	
Maxim. mutual information	
Maxim. product of SINRs	

Without loss of generality, it is assumed that $0 < \alpha_1 \le \cdots \le \alpha_L$. The composite function is

$$f_0 \circ \exp(\mathbf{x}) = \exp\left(\sum_{i=1}^{L} \alpha_i x_i\right), \tag{4.66}$$

which is minimized in the region $\mathbf{x} \in \mathcal{D}_L$. According to Lemma 3.2, $\sum_{i=1}^{L} \alpha_i x_i$ (assuming $\alpha_i \le \alpha_{i+1}$) is a Schur-concave function on $\mathcal{D}_L$, so is $\exp\left(\sum_{i=1}^{L} \alpha_i x_i\right)$ by Lemma 2.5. In this case, the optimum nonlinear DFE based transceiver design degenerates to (linear) diagonal transmission.

4.4.4.2 Maximization of the Weighted Sum of SINRs

The objective function to maximize is $\sum_{i=1}^{L} \alpha_i \mathrm{SINR}_i$, where $\alpha_1 \le \cdots \le \alpha_L$. Then the cost function to minimize is

$$f_0(\{\mathrm{MSE}_i\}) = -\sum_{i=1}^{L} \alpha_i \left(\frac{1}{\mathrm{MSE}_i} - 1\right), \tag{4.67}$$

and the composite function is

$$f_0 \circ \exp(\mathbf{x}) = -\sum_{i=1}^{L} \alpha_i \left(\exp(-x_i) - 1\right). \tag{4.68}$$

Similar to the proof of Lemma 3.2, it can be shown that the function $f_0 \circ \exp(\mathbf{x})$ (assuming $\alpha_i \le \alpha_{i+1}$) is minimized in the region $\mathbf{x} \in \mathcal{D}_L$. Note that for $\mathbf{x} \in \mathcal{D}_L$ and $\alpha_i \le \alpha_{i+1}$, the derivative

$$\frac{\partial f_0 \circ \exp(\mathbf{x})}{\partial x_i} = \alpha_i \exp(-x_i)$$

is increasing in $i = 1, \ldots, L$. By Lemma 2.8, $f_0 \circ \exp(\mathbf{x})$ is a Schur-concave function over $\mathcal{D}_L$.

4.4.5 Other Cost Functions

There exist cost functions for which the composite function $f_0 \circ \exp$ is neither Schur-convex nor Schur-concave. The following two examples are from Table 3.1.

4.4.5.1 Minimization of the Weighted Sum of MSEs

The cost function is $f_0(\{\mathrm{MSE}_i\}) = \sum_{i=1}^{L} \alpha_i \mathrm{MSE}_i$, where $\alpha_i \leq \alpha_{i+1}$. The composite function is

$$f_0 \circ \exp(\mathbf{x}) = \sum_{i=1}^{L} \alpha_i \exp(x_i). \tag{4.69}$$

By Example 2.5, although $f_0(\mathbf{x})$ is a Schur-concave function, $f_0 \circ \exp(\mathbf{x})$ is neither Schur-concave nor Schur-convex on $\mathbf{x} \in \mathcal{D}_L$.

4.4.5.2 Maximization of the Exponentially Weighted Product of SINRs

The objectivefunction to maximize is $\prod_{i=1}^{L} \mathrm{SINR}_i^{\alpha}$. The cost function to minimize is

$$f_0(\{\mathrm{MSE}_i\}) = \sum_{i=1}^{L} \log((\mathrm{MSE}_i^{-1} - 1)^{-\alpha_i}). \tag{4.70}$$

The composite function $f_0 \circ \exp$ is

$$f_0 \circ \exp(\mathbf{x}) = -\sum_{i=1}^{L} \alpha_i \log\left(\exp(-x_i) - 1\right). \tag{4.71}$$

For $\alpha_1 \leq \cdots \leq \alpha_L$, the optimum transceiver minimizing the cost function occurs in $\mathbf{x} \in \mathcal{D}_L$. But for $\alpha_i \leq \alpha_{i+1}$ and, the derivative

$$\frac{\partial f_0 \circ \exp(\mathbf{x})}{\partial x_i} = \frac{\alpha_i \exp(-x_i)}{\exp(-x_i) - 1}. \tag{4.72}$$

is *not* monotonous in $i = 1, \ldots, L$ for some $\mathbf{x} \in \mathcal{D}_L$. Thus $f_0 \circ \exp(\mathbf{x})$ is neither Schur-concave nor Schur-convex.

Theorem 4.3 gives the solution for such cost functions. In this cases, the power allocation has to be solved according to (4.35), which is more complicated than the power allocation in Theorem 4.4.

4.4.6 Numerical Results

We present two numerical examples to compare the optimum nonlinear design minimizing the BER, i.e., the UCD scheme, against the

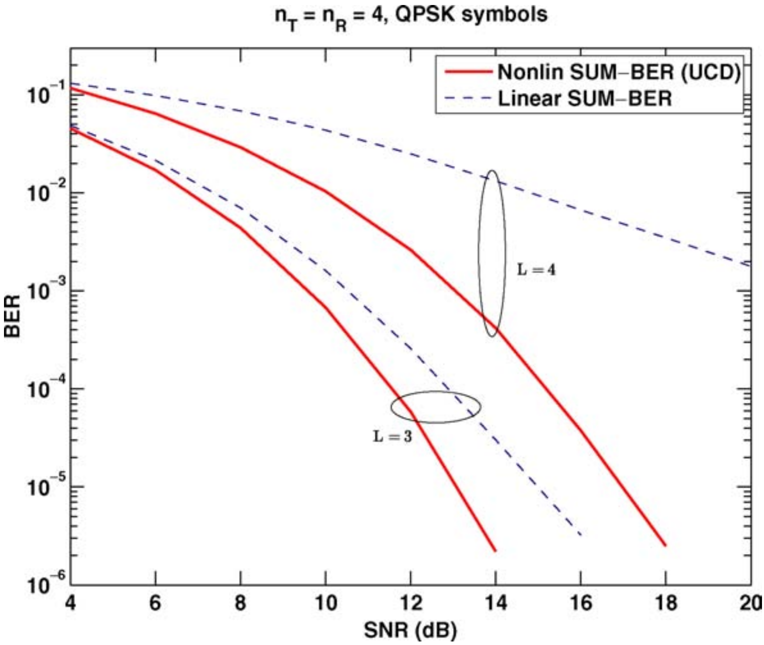


Fig. 4.5 BER performances of the linear and nonlinear transceiver designs minimizing BER. L is the number of substreams.

optimal linear SUM-BER design obtained in Section 3.4.3.4 and the maximum likelihood (ML) detector without precoder [164] which is a scheme assuming no CSIT.

In the first example, we compare the BER performance of the linear and nonlinear transceiver designs. The simulation result in Figure 4.5 is based on 10,000 Monte Carlo trials of 4×4 i.i.d. Rayleigh flat fading channels with QPSK symbols as input. Although the linear SUM-BER method is optimum among all the *linear* designs (see Figure 3.6), the *nonlinear* design (UCD) can still outperform it significantly. Moreover, the advantage of the nonlinear UCD design is larger in a fully loaded system ($L = 4$). This is because the linear design is not robust against rank-deficient channel realizations.

In the second example, we compare the BER performance of the UCD scheme against the ML receiver without precoder. The channel is 3×3 i.i.d. Rayleigh fading channel with 16-QAM symbols as input.

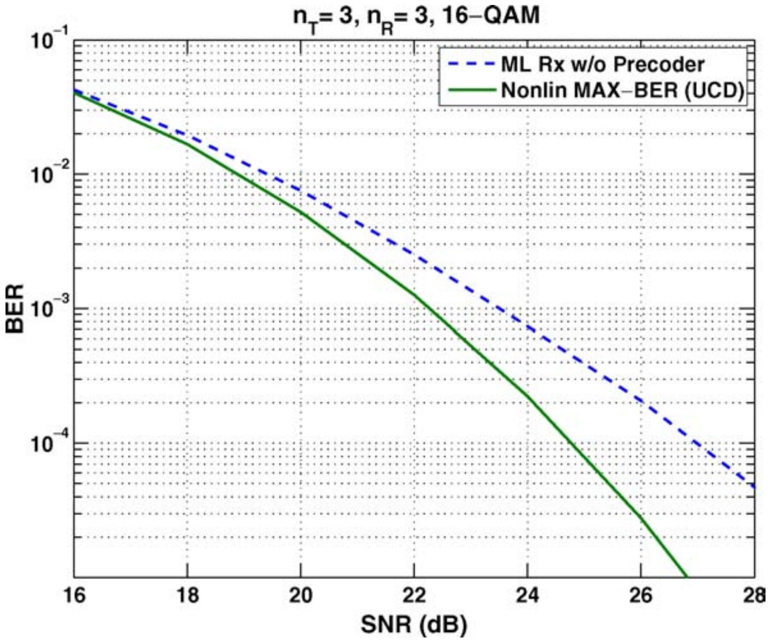


Fig. 4.6 BER performances of the UCD–DFE and the ML receiver without precoder in i.i.d. Rayleigh flat fading channel with $n_T = 3$ and $n_R = 3$ and 16-QAM input. The result is based on 5000 channel realizations.

Three substreams ($L = 3$) are transmitted simultaneously. We see from Figure 4.6 that the nonlinear transceiver design outperforms the unprecoded ML receiver and has higher diversity gain. Indeed, the UCD has maximal diversity gain $n_T n_R = 9$ while the ML only has diversity gain equal to $n_R = 3$ [73].

4.5 Optimum Transmitter with Individual QoS Constraints

This section deals with the problem formulation in (4.14) reproduced next for convenience:

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W},\mathbf{B}}{\text{minimize}} & \text{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \\ \text{subject to} & \text{SINR}_i \geq \gamma_i, \quad 1 \leq i \leq L. \end{array} \tag{4.73}$$

The optimal DFE matrices $\mathbf{W}$ and $\mathbf{B}$ have already been derived in (4.21), with resulting SINRs given by $\text{SINR}_i = [\mathbf{R}]_{ii}^2 - 1$ for $1 \leq i \leq L$, where $[\mathbf{R}]_{ii}$ is the ith diagonal element of $\mathbf{R}$ given in the QR decomposition $\begin{pmatrix} \mathbf{HP} \\ \mathbf{I}_L \end{pmatrix} = \mathbf{QR}$. Hence, (4.73) may be reformulated as

$$\begin{array}{ll} \underset{\mathbf{P}}{\text{minimize}} & \text{Tr}\ (\mathbf{PP}^\dagger) \\ \text{subject to} & \begin{pmatrix} \mathbf{HP} \\ \mathbf{I}_L \end{pmatrix} = \mathbf{QR} \\ & [\mathbf{R}]_{ii} \geq \sqrt{1+\gamma_i}, \quad 1 \leq i \leq L. \end{array} \tag{4.74}$$

Theorem 4.5. The solution to (4.74) must be of form:

$$\mathbf{P} = \mathbf{V}_H \text{diag}(\sqrt{\mathbf{p}}) \boldsymbol{\Omega}^\dagger,$$

where $\mathbf{p}$ is solved according to

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & \sum_{k=1}^K p_k \\ \text{subject to} & (\{1+\sigma_{H,i}^2 p_i\}_{i=1}^K, 1, \ldots, 1) \prec_\times (1+\gamma_1, \ldots, 1+\gamma_L), \end{array} \tag{4.75}$$

and $\boldsymbol{\Omega}$ is chosen such that the diagonal element of $\mathbf{R}$ in $\begin{pmatrix} \mathbf{HP} \\ \mathbf{I}_L \end{pmatrix} = \mathbf{QR}$ are $[\mathbf{R}]_{ii} = \sqrt{1+\gamma_i}$, $1 \leq i \leq L$.

Proof. The proof is not difficult given the proof of Theorem 4.3. We only give a sketch here. Readers are referred to [71] for the details.

First, observe that an optimal solution to (4.74) must satisfy $[\mathbf{R}]_{ii} = \sqrt{1+\gamma_i}$ for $\forall i$. The reason is the following. Suppose $[\mathbf{R}]_{ii} > \sqrt{1+\gamma_i}$. We can scale down the ith column of $\mathbf{P}$ until $[\mathbf{R}]_{ii} = \sqrt{1+\gamma_i}$, and the overall input power $\text{Tr}(\mathbf{PP}^\dagger)$ is reduced. Meanwhile, the other diagonal entries of $\mathbf{R}$ do not decrease. Indeed, $[\mathbf{R}]_{jj}$, for $j < i$, remain the same and $[\mathbf{R}]_{jj}$, for $j > i$, may even increase as the jth substream is now subject to weaker interference from the ith substream.

Second, denote $\mathbf{P} = \mathbf{U}_P \text{diag}(\sqrt{\mathbf{p}}) \boldsymbol{\Omega}^\dagger$ as its SVD. Then it can be proven that the constraint

$$\begin{pmatrix} \mathbf{HP} \\ \mathbf{I}_L \end{pmatrix} = \mathbf{QR} \ \text{ with } [\mathbf{R}]_{ii} = \sqrt{1+\gamma_i}, \quad 1 \leq i \leq L$$

is relaxed to the maximal extent when $\mathbf{U}_P = \mathbf{V}_H$.

Finally, according to the proof of Lemma 4.9, there exists a semi-unitary matrix $\mathbf{\Omega}$ such that $[\mathbf{R}]_{ii} = \sqrt{1+\gamma_i}$, for $1 \leq i \leq L$, if and only if the constraint of (4.75) is satisfied. □

Denoting $\gamma_{[i]}$ the ith largest element of $\{\gamma_i\}_{i=1}^{L}$, the problem (4.75) may be rewritten more explicitly as

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & \sum_{k=1}^{K} p_k \\ \text{subject to} & \prod_{i=1}^{k}(1+\sigma_{H,i}^2 p_i) \geq \prod_{i=1}^{k}(1+\gamma_{[i]}), \;\; 1 \leq k \leq K-1, \\ & \prod_{i=1}^{K}(1+\sigma_{H,i}^2 p_i) = \prod_{i=1}^{L}(1+\gamma_{[i]}). \end{array} \tag{4.76}$$

Note that the constraints in (4.76) are non-convex. However, after a closer look at the structure of this problem, we can still find an efficient algorithm to produce the optimal solution to this problem as considered in the next section.

4.5.1 Optimal Power Allocation

We begin with a change of variables to further simplify the formulation in (4.76). Define

$$\begin{array}{ll} \psi_i = p_i + \frac{1}{\sigma_{H,i}^2}, & 1 \leq i \leq K, \\ \beta_i = \frac{1+\gamma_{[i]}}{\sigma_{H,i}^2}, & 1 \leq i < K, \\ \beta_K = \frac{1}{\sigma_{H,K}^2} \prod_{i=K}^{L}(1+\gamma_{[i]}). & \end{array} \tag{4.77}$$

With these definitions, problem (4.76) reduces to

$$\begin{array}{lll} \underset{\boldsymbol{\psi}}{\text{minimize}} & \sum_{i=1}^{K} \psi_i & \\ \text{subject to} & \prod_{i=1}^{k} \psi_i \geq \prod_{i=1}^{k} \beta_i & 1 \leq k \leq K, \\ & \psi_k \geq \frac{1}{\sigma_{H,k}^2} & 1 \leq k \leq K. \end{array} \tag{4.78}$$

To find an efficient solution, we shall need to establish three lemmas whose proofs are relegated to Appendix 4.E.

The first result refers to a relaxed version of problem (4.78).

Lemma 4.6. Any solution $\boldsymbol{\psi}$ to the problem

$$\min_{\boldsymbol{\psi}} \sum_{i=1}^{K} \psi_i \text{ subject to } \prod_{i=1}^{k} \psi_i \geq \prod_{i=1}^{k} \beta_i, \ 1 \leq k \leq K, \tag{4.79}$$

has the property that $\psi_{i+1} \leq \psi_i$ for each i.

Lemma 4.6 provides some insight into the structure of a solution to (4.78), which leads to the next result.

Lemma 4.7. There exists a solution $\boldsymbol{\psi}$ to (4.78) with the property that for some integer $j \in [1, K]$,

$$\begin{aligned} \psi_{i+1} \leq \psi_i \text{ for all } i < j, \quad & \psi_{i+1} \geq \psi_i \text{ for all } i \geq j, \\ & \psi_i = \frac{1}{\sigma_{H,i}^2} \text{ for all } i > j. \end{aligned} \tag{4.80}$$

In particular, $\psi_j \leq \psi_i$ for all i.

By Lemma 4.7, ψ_i is a decreasing function of i for $i \in [1, j]$ while $\psi_i = 1/\sigma_{H,i}^2$ for $i > j$. Since $\sigma_{H,i}$ is a decreasing function of i, it follows that $p_i = \psi_i - 1/\sigma_{H,i}^2$ is a decreasing function of i for $i \in [1, j]$ with $p_i \geq 0$, while $p_i = 0$ for $i > j$. Hence, p_i is a decreasing function of $i \in [1, K]$.

We refer to the index j in Lemma 4.7 as the "break point." At the break point, the lower bound constraint $\psi_i \geq 1/\sigma_{H,i}^2$ changes from inactive to active. We now use Lemma 4.7 to obtain an algorithm for (4.78).

Lemma 4.8. Let γ_k denote the kth geometric mean of $\{\beta_i\}_{i=1}^{k}$:

$$\gamma_k = \left(\prod_{i=1}^{k} \beta_i \right)^{1/k},$$

and let l denote an index for which γ_k is the largest:

$$l = \arg\max\{\gamma_k : 1 \leq k \leq K\}. \tag{4.81}$$

If $\gamma_l \geq 1/\sigma_{H,l}^2$, then setting $\psi_i = \gamma_l$ for all $i \leq l$ is optimal in (4.78). If $\gamma_l < 1/\sigma_{H,l}^2$, then $\psi_i = 1/\sigma_{H,i}^2$ for all $i \geq l$ is an optimal solution to (4.78).

Based on Lemma 4.8, we can use the following strategy to solve (4.78). We form the geometric mean described in Lemma 4.8 and we evaluate l. If $\gamma_l \geq 1/\sigma_{H,l}^2$, then we set $p_i = \gamma_l$ for $i \leq l$, and we simplify (4.78) by removing ψ_i, $1 \leq i \leq l$, from the problem. If $\gamma_l < 1/\sigma_{H,l}^2$, then we set $\psi_i = 1/\sigma_{H,i}^2$ for $i \geq l$, and we simplify (4.78) by removing ψ_i, $l \leq i \leq K$, from the problem. The Matlab code `PowerLoadQoS` implementing this algorithm appears in Table 4.4. From the output $\boldsymbol{\psi}$ of the Matlab function, we have the power allocation $p_i = \psi_i - \frac{1}{\sigma_{H,i}^2}$ for $1 \leq i \leq K$ according to (4.77).

Table 4.5 summarizes the procedure to design the DF MIMO transceiver as the solution to (4.14), which is quite similar to the procedure given in Table 4.1 except steps 2 and 3.

Table 4.4 A Matlab function to solve (4.76).

```
function $\mathbf{p}$ = PowerLoadQoS ($\boldsymbol{\beta}$,$\boldsymbol{\sigma}$)
L = 1 ; R = length ($\boldsymbol{\beta}$) ; $\psi$ = zeros (1, R) ;
$\boldsymbol{\zeta}$ = cumsum (log ($\boldsymbol{\beta}$)) ;
while R $\geq$ L
     [t, l] = max ($\boldsymbol{\zeta}$(L:R)./[1:R-L+1]) ;
     $\gamma_l$ = exp (t) ; L1 = L + l - 1 ;
     if $\gamma_l$ $\geq$ 1/$\boldsymbol{\sigma}$(L1)^2
          $\boldsymbol{\psi}$(L:L1) = $\gamma_l$ ;
          L = L + l ;
          $\boldsymbol{\zeta}$(L:R) = $\boldsymbol{\zeta}$(L:R) - $\boldsymbol{\zeta}$(L-1) ;
     else
          $\boldsymbol{\psi}$(L1:R) = 1./($\boldsymbol{\sigma}$(L1:R).^2) ;
          $\boldsymbol{\zeta}$(L1-1) = $\boldsymbol{\zeta}$(R) - sum (log ($\boldsymbol{\psi}$(L1:R))) ;
          R = L1 - 1 ;
     end
end
$\mathbf{p} = \boldsymbol{\psi}$-(1./$\boldsymbol{\sigma}$).^2;
```

Table 4.5 Procedure of designing DF transceiver with individual QoS constraint.

Step	Operation
1	Calculate SVD $\mathbf{H} = \mathbf{U}_H \mathbf{\Sigma}_H \mathbf{V}_H^\dagger$.
2	Obtain power allocation $\mathbf{p}$ from Table 4.4.
3	Calculate GTD $\mathbf{J} = \left[\; \mathbf{U}_H \mathbf{\Sigma}_H [\mathrm{diag}(\sqrt{\mathbf{p}}) \vdots \mathbf{0}_{K \times (L-K)}] \mathbf{I}_L \; \right] = \mathbf{Q}_J \mathbf{R} \mathbf{P}_J^\dagger$, where $\mathbf{R}$ has diagonal elements $[\mathbf{R}]_{ii} = \sqrt{1+\gamma_i}$.
4	Obtain linear precoder $\mathbf{P} = \mathbf{V}_H [\mathrm{diag}(\sqrt{\mathbf{p}}) \vdots \mathbf{0}_{K \times (L-K)}] \mathbf{P}_J$.
5	Compute the DFE $\mathbf{W} = \bar{\mathbf{Q}} \mathbf{D}_R^{-1}$ and $\mathbf{B} = \mathbf{D}_R^{-1} \mathbf{R} - \mathbf{I}$, where $\bar{\mathbf{Q}}$ consists of the first n_R rows of $\mathbf{Q}_J$.

The PowerLoadQoS routine given in Table 4.4 can be regarded as a multi-level waterfilling algorithm. Note that if the "if" part of the matlab code is run only once, then the routine makes the leading elements of $\boldsymbol{\psi}$ all equal while setting the trailing elements to be $\psi_i = 1/\sigma_{H,i}^2$. In this case, the power allocation turns out to be the standard waterfilling algorithm. In this case, the nonlinear transceiver scheme is optimal in terms of maximizing the channel mutual information given the input power constraint. On the other hand, if some substream has a very high prescribed SINR such that the l given in (4.81) is less than the "break point" j, then $\boldsymbol{\psi}$ leads to a multi-level waterfilling power allocation, which suffers from overall capacity loss. (In this case the "if" part of the routine is run more than once.) This happens when the target rate vector $[R_1, \ldots, R_L]$ is *not* majorized by $[C_1, \ldots, C_K, 0, \ldots, 0]$, where $C_k, k = 1, \ldots, K$, are the capacities of the eigen subchannels with *standard* waterfilling power allocation. In other words, the target rate tuple is outside of the convex hull spanned by the $L!$ permutations of $[C_1, \ldots, C_K, 0, \ldots, 0]$ (cf. Figure 4.4). When it happens, the power allocation algorithm applies multi-level waterfilling to "expand the convex hull" such that the target rate vector lies in it.

4.5.2 Numerical Examples

Figure 4.7 gives the required power at the transmitter versus the different QoS constraints in terms of BER. The simulation setting is the same as Figure 3.7.

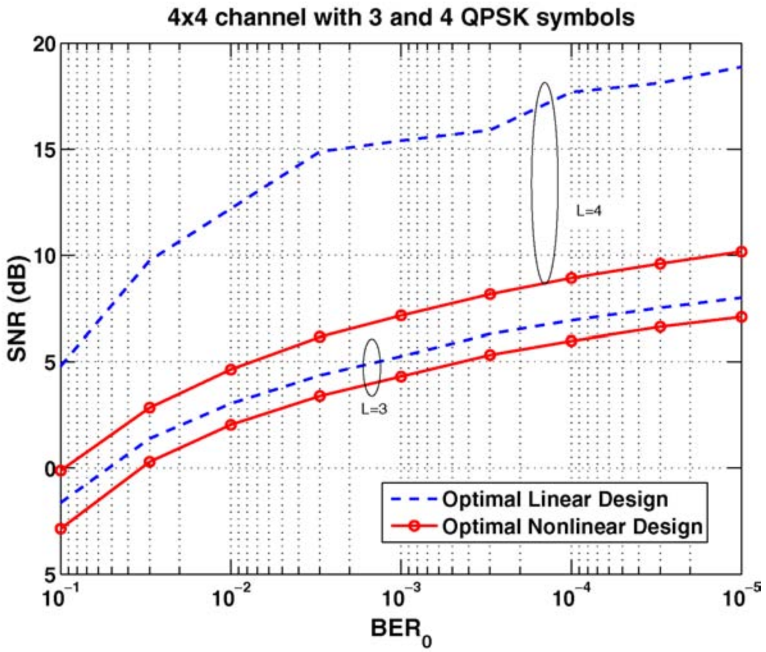


Fig. 4.7 Power versus the different QoS constraints in terms of BER for a 4 × 4 MIMO channel with $L = 3$ and $L = 4$ transmitted QPSK symbols. (The BER of the first substream is along the x-axis, and the BER of the second, third, and possibly fourth substreams are given by scaling with the factors 0.5, 0.5, and 0.1).

In particular, the optimum linear design and nonlinear design are compared with a difference of about 1.5 dB for $L = 3$ and over 5 dB for $L = 4$. This result agrees with Figure 4.5 where we have seen that the advantage of the nonlinear design is more prominent in a fully loaded system. It is worthwhile noting that the nonlinear DF transceiver design has error propagation. Therefore the DFE should detect the substream with the smallest error probability first to minimize the error propagation toward the substreams detected later. Hence compared to the linear scheme, the nonlinear scheme has the extra constraint of detection ordering. But such a constraint is usually harmless.

Figure 4.8 compares the achievable regions of linear and nonlinear transceivers in terms of MSE for a 2 × 2 channel with singular values $\boldsymbol{\sigma} = [5;\ 1.2]$. The input SNR is 15 dB. The boundary of achievable region is called Pareto-optimal boundary as illustrated in Figure 4.8.

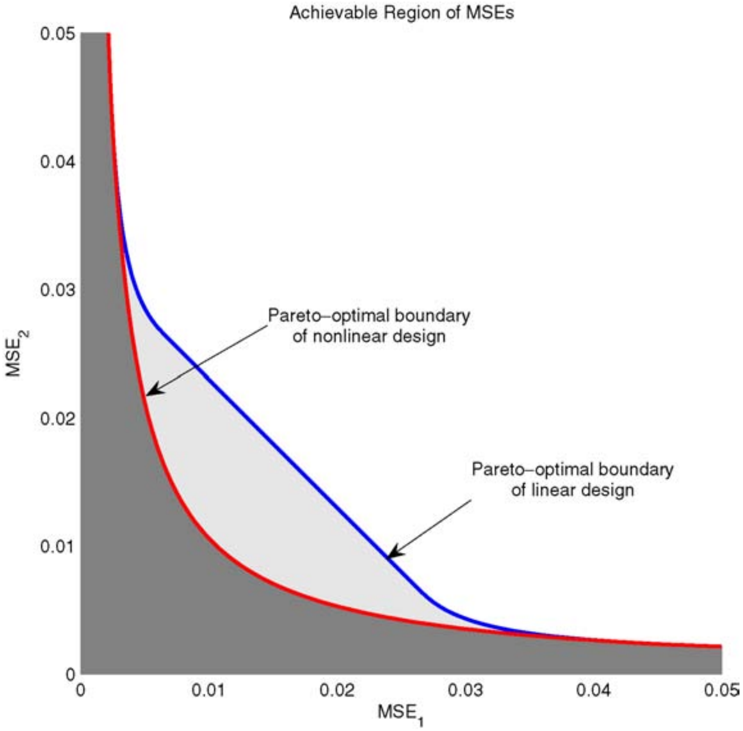


Fig. 4.8 The Pareto-optimal boundaries of linear and nonlinear designs in terms of MSE. The achievable region is above the boundary. The 2 × 2 channel has singular values [5; 1.2].

The achievable region of the linear transceiver design is a proper subset of that of the nonlinear design. Interestingly, a section of the linear design's Pareto-optimal boundary overlaps with the straight line $\text{MSE}_1 + \text{MSE}_2 = c_1$ while part of nonlinear design's Pareto-optimal boundary is a hyperbola $\text{MSE}_1 \cdot \text{MSE}_2 = c_2$, where c_1 and c_2 are some constants. When MSE_1 and MSE_2 are disparate enough, the two boundaries coincide. This corresponds to the scenario where both transceiver designs degenerate to (linear) diagonal transmission.

Figure 4.9 is similar to Figure 4.8 but with the only difference that the 2 × 2 channel matrix here has singular values $\boldsymbol{\sigma} = [4;\ 1.5]$. Comparing Figures 4.8 and 4.9, one can readily see that the advantage of the nonlinear design is smaller in the channel with less disparate singular values.

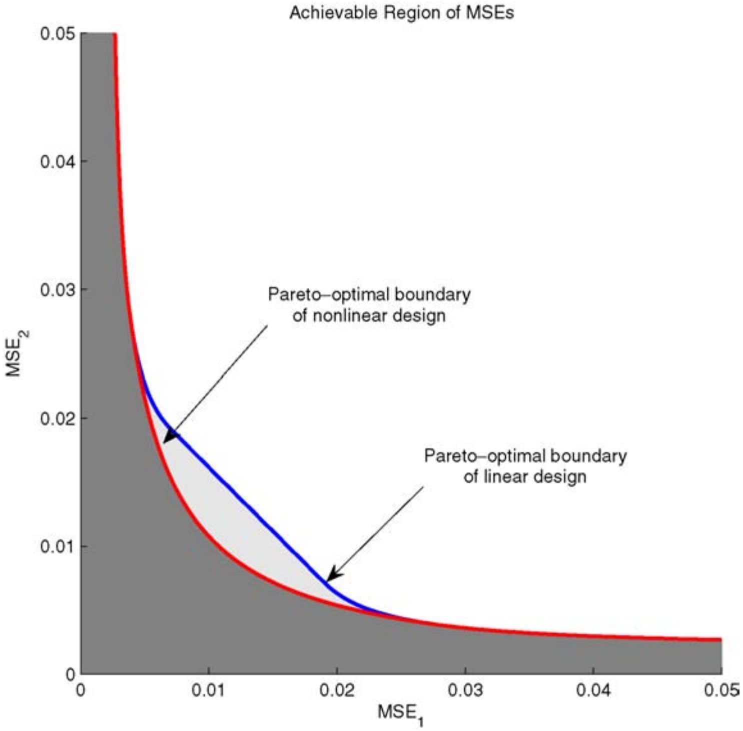


Fig. 4.9 The Pareto-optimal boundaries of linear and nonlinear designs in terms of MSE. The 2×2 channel has singular values [4; 1.5].

4.5.3 Transceiver Designs with ZF–DFE Receiver

Now we have solved the transceiver design problems (4.13) and (4.14). The previous discussion concentrated on the designs based on the MMSE–DFE. However, as introduced in Section 4.3, there is another type of DFE, i.e., ZF–DFE. Correspondingly, we may design transceivers with the ZF–DFE receiver.

The feed-forward and feedback matrices of ZF–DFE was introduced in Section 4.3 (see (4.23) and (4.24)). The ZF–DFE yields K subchannels with output SNRs (see (4.30))

$$\mathrm{SNR}_i^{\mathrm{zf-dfe}} = [\mathbf{R}_G]_{ii}^2, \quad 1 \leq i \leq K, \tag{4.82}$$

where $[\mathbf{R}_G]_{ii}$ are the diagonal entries of the QR decomposition $\mathbf{HP} = \mathbf{Q}_G\mathbf{R}_G$. Hence similar to (4.33) and (4.74), we may optimize the

ZF–DFE-based MIMO transceivers by solving one of the following problems:

$$\begin{array}{ll} \underset{\mathbf{P}}{\text{minimize}} & f_0\left(\{[\mathbf{R}_G]_{ii}^{-2}\}\right) \\ \text{subject to} & \mathbf{HP} = \mathbf{Q}_G\mathbf{R}_G \\ & \mathrm{Tr}(\mathbf{PP}^\dagger) \leq P_0, \end{array} \tag{4.83}$$

and

$$\begin{array}{ll} \underset{\mathbf{P}}{\text{minimize}} & \mathrm{Tr}\ (\mathbf{PP}^\dagger) \\ \text{subject to} & \mathbf{HP} = \mathbf{Q}_G\mathbf{R}_G \\ & [\mathbf{R}_G]_{ii} \geq \sqrt{1+\gamma_i}, \quad 1 \leq i \leq K. \end{array} \tag{4.84}$$

These problems present no fundamental difficulty given our previous study on the transceiver design based on MMSE–DFE. The readers are referred to [72] for an interesting special case of (4.83).

4.6 A Dual Form Based on Dirty Paper Coding

In the preceding sections, we have studied the nonlinear MIMO transceiver design as a combination of a linear precoder with an MMSE–DFE. In this section, we introduce an alternative implementation form of the MIMO transceiver designs, which is inspired by the uplink–downlink duality revealed in [158, 160, 172] as well as the DPC [32, 39, 40, 173].

4.6.1 Dirty Paper Coding

Consider a scalar Gaussian channel

$$y = x + s + z, \tag{4.85}$$

where s and z are independent interference and Gaussian noise with s known to the receiver. Clearly, the capacity of the channel (4.85) is exactly the same as the additive white Gaussian noise (AWGN) channel

$$y = x + z, \tag{4.86}$$

since the receiver may remove the known interference s prior to signal detection, which is what the DFE does.

Now consider the signal model in (4.85) where the interference s is *unknown* to the receiver but known to the transmitter (non-causally). It is natural to raise the following question: can we exploit this information for better communication? Costa studied this problem in [32] and gave a rather surprising answer. He showed that the capacity of the channel (4.85) is equal to that of the classic AWGN channel $y = x + z$ for the case of Gaussian interference s, i.e., the communication is as good as if the interference did not exist at all! It is further shown in [30, 39] that the same capacity can be achieved even if the interference is of arbitrary distribution. Costa's result suggests the existence of a code that can eliminate the influence from the interference s without consuming additional input power. Such a coding technique is called "dirty paper coding" (DPC).

Now we have peeked into the duality between the DPC and DFE in the context of MIMO communications: DPC cancels the known interference from the other substreams at transmitter, while the DFE cancels the known interference, i.e., the previously detected substreams, at receiver.

To have a basic idea of how the seemingly magic DPC works, we give a brief introduction to the Tomlinson–Harashima code [58, 149], which is a simple but suboptimal implementation of the DPC. As illustrated in Figure 4.10, the Tomlinson–Harashima code partitions the real line into infinitely many intervals, each of which has length Δ. All the points having distance $k\Delta$ for $k \in \mathbb{Z}$ among each other are regarded as an equivalent class. For instance, the information u can be represented by

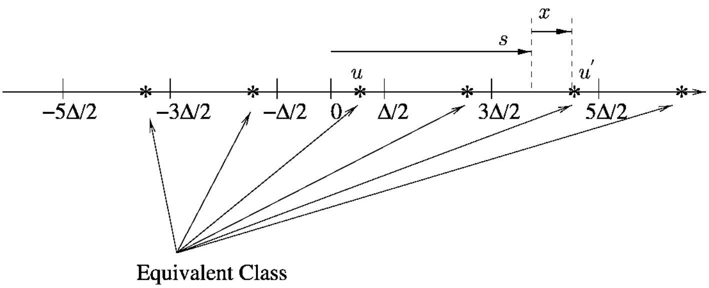


Fig. 4.10 An illustration of the Tomlinson–Harashima code.

an equivalent class consisting of infinitely many points $\{u + k\Delta\}_{k=-\infty}^{\infty}$, as illustrated in Figure 4.10. Suppose the transmitter wants to send the information u and the channel is subject to the interference s. Then the transmitter finds the representative point $u' = u + k\Delta$ which is closest to s among the equivalent class and sends signal $x = u' - s$. Note that the amplitude of the transmitted signal is constrained to be $|x| \leq \Delta/2$ and, hence, the average input power is constrained disregarding the distribution of s. At the receiver side $y = x + s + z = u + k\Delta + z$. Applying modulo operation so that $y \bmod \Delta \in (-\Delta/2, \Delta/2]$, we can recover u which is free of the interference from s but is only subject to the modulo additive noise. So the output of DPC decoder $\tilde{\mathbf{u}} = \mathbf{u} + \breve{\mathbf{n}}$, where $\breve{\mathbf{n}}$ is a vector of amplitude-constrained modulo noise.

In the high SNR regime the Tomlinson–Harashima code has 1.53 dB performance loss compared to the theoretic limit. It is shown in [39] that this performance loss is due to the fact that the Tomlinson–Harashima code only makes interference cancelation in a symbol-by-symbol manner and does not exploit the non-causal information on s. The improved DPCs are proposed in [40] and [173].

4.6.2 Nonlinear MIMO Transceiver Design Using DPC

We now introduce the nonlinear MIMO transceiver design using DPC.

The layout of the DPC-based transceiver design is given in Figure 4.11, where the DPC encoder and DPC decoder are illustrated in Figures 4.12 and 4.13, respectively, and $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{W}}$ are the linear precoder and equalizer matrices to determine. In Figure 4.12, α is an MMSE scaling constant[10] [39, 41] and η_{ij} is the (i, j)th entry of $\tilde{\mathbf{W}}^{\dagger}\mathbf{H}\tilde{\mathbf{P}}$.

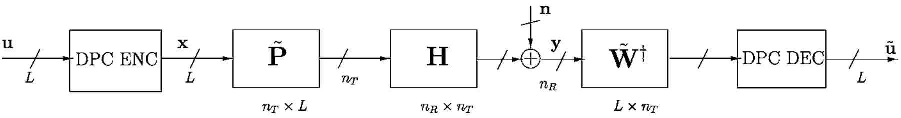


Fig. 4.11 MIMO transceiver design based on DPC.

[10] We may set $\alpha = 1$ for simplicity, incurring minor degradation at high SNR.

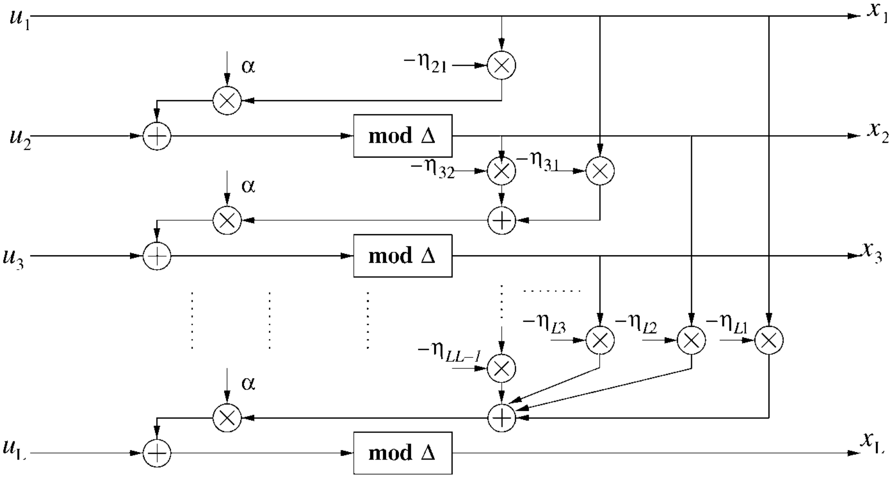


Fig. 4.12 Dirty paper encoder.

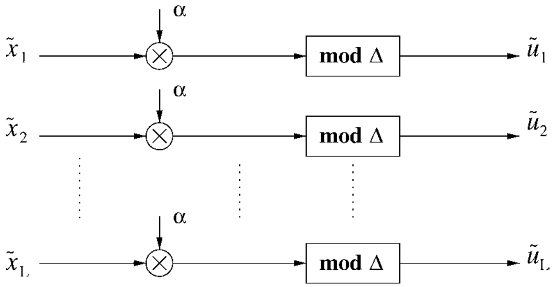


Fig. 4.13 Dirty paper decoder.

A direct construction of $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{W}}$ is difficult. Instead, we resort to the uplink–downlink duality [160] to obtain $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{W}}$.

According to the uplink–downlink duality, we first calculate the precoder and equalizer of the DFE based transceiver design in the reverse channel

$$\mathbf{y} = \mathbf{H}^{\dagger}\mathbf{P}\mathbf{x} + \mathbf{n}, \tag{4.87}$$

where the roles of the transmitter and receiver are exchanged and the channel matrix $\mathbf{H}$ is replaced by $\mathbf{H}^{\dagger}$. Then, we obtain the precoder $\mathbf{P}$

and the equalizer $\mathbf{W} \triangleq [\mathbf{w}_1, \ldots, \mathbf{w}_L]$ according to Table 4.1 or Table 4.5. (Note that since $\mathbf{H}$ is replaced by $\mathbf{H}^\dagger$, the roles of $\mathbf{U}_H$ and $\mathbf{V}_H$ should also be exchanged.)

From the solution to the reverse channel, the precoder of the DPC based transceiver design should be

$$\tilde{\mathbf{P}} = [\sqrt{q_1}\bar{\mathbf{w}}_1, \ldots, \sqrt{q_L}\bar{\mathbf{w}}_L], \tag{4.88}$$

where $\bar{\mathbf{w}}_l$ are obtained by scaling $\mathbf{w}_l$ such that $\|\bar{\mathbf{w}}_l\| = 1$, and $\{q_i\}_{i=1}^L$ will be determined later in (4.92). The receiving matrix $\tilde{\mathbf{W}}$ is obtained by scaling the columns of $\mathbf{P}$ such that the diagonal $[\tilde{\mathbf{W}}^\dagger \mathbf{H} \tilde{\mathbf{P}}]_{ii} = 1$. To obtain q_i, $1 \le i \le L$, we can assume for the moment that the columns of $\tilde{\mathbf{W}}$, $\bar{\mathbf{p}}_1, \ldots, \bar{\mathbf{p}}_L$, have unit Euclidean norm since scaling the receiving vector does not change the output SINR. The received signal at the ith substream is

$$y_i = \bar{\mathbf{p}}_i^\dagger \mathbf{H} \bar{\mathbf{w}}_i \sqrt{q_i} x_i + \sum_{j=i+1}^{L} \bar{\mathbf{p}}_i^\dagger \mathbf{H} \bar{\mathbf{w}}_j \sqrt{q_j} x_j + \sum_{j=1}^{i-1} \bar{\mathbf{p}}_i^\dagger \mathbf{H} \bar{\mathbf{w}}_j \sqrt{q_j} x_j + \bar{\mathbf{p}}_i^\dagger \mathbf{n}. \tag{4.89}$$

We apply DPC substream-by-substream in the order of increasing index. For the ith substream, the information u_i is encoded into x_i by treating $\sum_{j=1}^{i-1} \bar{\mathbf{p}}_i^\dagger \mathbf{H} \bar{\mathbf{w}}_j \sqrt{q_j} x_j$ as the interference known at the transmitter. It should be emphasized that the interference term $\sum_{j=i+1}^{L} \bar{\mathbf{p}}_i^\dagger \mathbf{H} \bar{\mathbf{w}}_j \sqrt{q_j} x_j$ is unknown at this step. Also note that here we precode the first substream first while for the DFE based design, we detect the Lth substream first. At the receiver side, after applying the linear filter $\tilde{\mathbf{W}}$ and the DPC decoder, we obtain L equivalent subchannels given by

$$\tilde{u}_i = \bar{\mathbf{p}}_i^\dagger \mathbf{H} \bar{\mathbf{w}}_i \sqrt{q_i} u_i + \sum_{j=i+1}^{L} \bar{\mathbf{p}}_i^\dagger \mathbf{H} \bar{\mathbf{w}}_j \sqrt{q_j} x_j + \bar{\mathbf{p}}_i^\dagger \mathbf{n} \tag{4.90}$$

with output SINR

$$\gamma_i = \frac{q_i |\bar{\mathbf{p}}_i^\dagger \mathbf{H} \bar{\mathbf{w}}_i|^2}{1 + \sum_{j=i+1}^{L} q_j |\bar{\mathbf{p}}_i^\dagger \mathbf{H} \bar{\mathbf{w}}_j|^2}, \quad \text{for } i = 1, 2, \ldots, L, \tag{4.91}$$

which may be written in a matrix form:

$$\begin{bmatrix} a_{11} & -\gamma_1 a_{12} & \cdots & -\gamma_1 a_{1L} \\ 0 & a_{22} & \cdots & -\gamma_2 a_{2L} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{LL} \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_L \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_L \end{bmatrix}, \tag{4.92}$$

where $a_{ij} \triangleq |\bar{\mathbf{p}}_i^\dagger \mathbf{H} \bar{\mathbf{w}}_j|^2$. It is easy to see that $q_i > 0$, $1 \leq i \leq L$. It is proven in [160] that $\sum_{i=1}^{L} q_i = \mathrm{Tr}(\tilde{\mathbf{P}}\tilde{\mathbf{P}}^\dagger) = \mathrm{Tr}(\mathbf{P}\mathbf{P}^\dagger) = \sum_{i=1}^{L} p_i$. That is, to obtain L subchannels with SINRs $\{\gamma_i\}_{i=1}^{L}$, the DPC based transceiver design needs exactly the same power as the DFE based design. Here we have adopted output SINR as the measure of QoS, which does not lose generality according to the relationship given in (4.26).

Because the DPC makes the interference cancelation at transmitter and is free of error propagation, the DPC based design may be a better choice than the DFE based design for a system with large dimensionality.

4.6.3 Numerical Example

Figure 4.14 compares the BER performance of the DFE based UCD (UCD–DFE) and the DPC based UCD (UCD–DPC) in a 4 × 4 Rayleigh flat fading channel. Four substreams of uncoded 64-QAM symbols are transmitted simultaneously. Here we use the simple Tomlinson–Harashima precoder as a suboptimal DPC for interference cancelation. To present a benchmark, we also include UCD-genie as the fictitious scheme for which a genie would eliminate the influence of erroneous detections from the previous substreams when using UCD–DFE. Figure 4.14 shows that the small BER degradation of UCD–DFE (about 0.5 dB for $\mathrm{BER} = 10^{-4}$) compared with UCD-genie, which is due to error propagation, can be effectively eliminated by the UCD–DPC. The slight SNR loss of UCD–DPC compared to the UCD-genie is mainly due to the inherent power loss and modulo loss of the Tomlinson–Harashima precoder [139, 173].

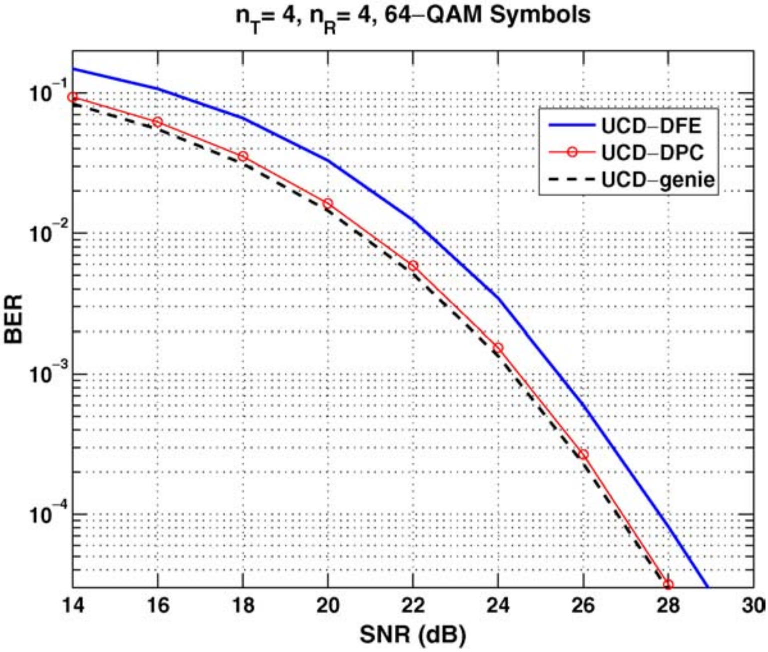


Fig. 4.14 BER performances of the UCD–DPC, UCD–DFE schemes, and the imaginary UCD-genie scheme in an i.i.d. Rayleigh flat fading channel with $n_T = 4$ and $n_R = 4$. The result is based on 5000 channel realizations.

4.7 A Particular Case: CDMA Sequence Design

Similar to MIMO transceiver design, the symbol synchronous CDMA sequence design problem also has been studied intensively over the past decade (e.g., [55, 127, 161, 163]). However, these two topics were investigated in the signal processing and information theory communities independently. In this section, we provide a perspective which identifies the problem of CDMA sequence optimization as a special case of the MIMO transceiver design. Consequently, the MIMO transceiver schemes introduced before can be applied to design optimal CDMA sequences with little modifications.

The signal model is as follows. Consider an idealized CDMA system where the channel does not experience any fading or near-far effect. In the uplink channel, L mobile users modulate their information symbols via spreading sequences $\{\mathbf{s}_i\}_{i=1}^{L}$ of length N which is the processing gain of the CDMA system. The discrete-time baseband CDMA signal received at the (single-antenna) base-station can be represented as [127]

$$\mathbf{y} = \mathbf{S}\mathbf{x} + \mathbf{n}, \tag{4.93}$$

where $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_L] \in \mathbb{R}^{N \times L}$ and the lth $(1 \leq l \leq L)$ entry of $\mathbf{x}$, x_l, stands for the information symbol from the lth user. In the downlink channel, the base station multiplexes the information dedicated to the L mobile users through the spreading sequences, which are the columns of $\mathbf{S}$, and all the mobiles receive a signal that follows the signal model in (4.93). It is important to remark that (4.93) can also be written as the general MIMO channel model in (4.1) with channel $\mathbf{H} = \mathbf{I}_N$ and linear precoder $\mathbf{P} = \mathbf{S}$. Here $n_R = n_T = N$ is the processing gain and the input power is $P_T = \mathrm{Tr}(\mathbf{S}\mathbf{S}^\dagger)$. Hence, optimizing the spreading sequences contained in $\mathbf{S}$ amounts to optimizing the linear precoder $\mathbf{P}$ in a MIMO system!

The goal here is to design the spreading sequences $\mathbf{S}$ such that the overall input power is minimized while the prescribed QoS constraints (in terms of output SINR $\{\gamma_i\}_{i=1}^{L}$) of each mobile user are satisfied. This problem is a particular case of the MIMO transceiver design problem with individual QoS constraints for the channel matrix $\mathbf{H} = \mathbf{I}$. The solution to this problem for the linear model is a particular case of the method given in Section 3.5, whereas for the DF model the problem can be solved via the scheme proposed in Section 4.5. But due to the simple channel matrix, the design procedure can be significantly simplified and has some interesting structure. In the following, we elaborate on the solution for the DF formulation.

4.7.1 Uplink Scenario

For the uplink scenario, i.e., the mobiles to base station case, the base station calculates the optimal CDMA sequence for each mobile user and the corresponding DFE matrices $\mathbf{W}$ and $\mathbf{B}$ needed at the base station. Then the base station informs the mobile users their designated CDMA sequences.

Using the scheme developed in Section 4.5, we first need to calculate the power loading levels. Similar to (4.76), we need to solve the problem (note that $\sigma_{H,i} = 1$ since here $\mathbf{H} = \mathbf{I}$)

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & \sum_{i=1}^{N} p_i \\ \text{subject to} & (\{1+p_i\}_{i=1}^{N}, \underbrace{1,\ldots,1}_{L-N}) \succ_{\times} \{1+\gamma_i\}_{i=1}^{L}. \end{array} \tag{4.94}$$

Similar to (4.77), (4.94) can be further simplified using the variables

$$\psi_i = p_i + 1, \quad \beta_i = 1 + \gamma_{[i]} \quad \text{for } i < N, \quad \text{and} \quad \beta_N = \prod_{i=N}^{L} (1+\gamma_i).$$

The simplified problem is

$$\begin{array}{ll} \underset{\boldsymbol{\psi}}{\text{minimize}} & \sum_{i=1}^{N} \psi_i \\ \text{subject to} & \prod_{i=1}^{k} \psi_i \geq \prod_{i=1}^{k} \beta_i \\ & \psi_k \geq 1, \ \ 1 \leq k \leq N. \end{array} \tag{4.95}$$

The algorithm `PowerLoadQoS` in Table 4.4 is simplified immensely when we apply it to (4.95). Since $\beta_i \geq 1 = 1/\sigma_{H,i}$ for all i, the constraints $\psi_i \geq 1$ are inactive. Also, because $\beta_i \leq \beta_{i-1}$ for all $i < N$, the geometric means satisfy $\gamma_i \leq \gamma_{i-1}$ for all $i < N$. Hence, in Lemma 4.8, the value of l is either 1 or N. If $l = 1$, then we set $\psi_1 = \beta_1$ and we remove ψ_1 from the problem. If $l = N$, then $\psi_i = \gamma_N$ for all i. It follows that there exists an index j with the property that

$$\psi_i = \beta_i \quad \text{for all } i \leq j \quad \text{and} \quad \psi_i = \left(\prod_{i=j+1}^{N} \beta_i \right)^{1/(N-j)} \quad \text{for all } i > j.$$

Thanks to the simple channel matrix ($\mathbf{H} = \mathbf{I}$), matrix $\mathbf{J}$ defined in Step 3 of Table 4.5 becomes

$$\mathbf{J} = \left[\begin{array}{c} [\mathrm{diag}(\sqrt{\mathbf{p}}) \vdots \mathbf{0}_{N \times (L-N)}] \\ \mathbf{I} \end{array} \right], \tag{4.96}$$

where $\mathbf{p}$ has entries $p_i = \psi_i - 1$. The SVD of $\mathbf{J}$ is

$$\mathbf{J} = \left[\begin{array}{c} [\mathrm{diag}(\sqrt{\mathbf{p}}) \vdots \mathbf{0}_{N \times (L-N)}] \boldsymbol{\Psi}^{-\frac{1}{2}} \\ \boldsymbol{\Psi}^{-\frac{1}{2}} \end{array} \right] \boldsymbol{\Psi}^{\frac{1}{2}} \mathbf{I}, \tag{4.97}$$

where $\boldsymbol{\Psi}$ denote an $L \times L$ identity matrix with its first N diagonal elements replaced by ψ_i for $1 \leq i \leq N$. We then apply the GTD algorithm to $\boldsymbol{\Psi}^{\frac{1}{2}}$ to obtain

$$\boldsymbol{\Psi}^{\frac{1}{2}} = \mathbf{Q}_{\Psi}\mathbf{R}\mathbf{P}_{\Psi}^{T}. \tag{4.98}$$

so that $\mathbf{R}$ has diagonal elements $\{\sqrt{\gamma_i + 1}\}_{i=1}^{L}$. According to Steps 4 and 5 of Table 4.5, we have the CDMA sequence

$$\mathbf{S} = \left[\operatorname{diag}(\sqrt{\mathbf{p}}) \vdots \mathbf{0}_{N\times(L-N)}\right]\mathbf{P}_{\Psi}, \tag{4.99}$$

and the feed-forward and feedback matrices

$$\mathbf{W} = [\operatorname{diag}(\sqrt{\mathbf{p}}) \vdots \mathbf{0}_{N\times(L-N)}]\boldsymbol{\Psi}^{-\frac{1}{2}}\mathbf{Q}_{\Psi}\mathbf{D}_{R}^{-1}, \quad \mathbf{B} = \mathbf{D}_{R}^{-1}\mathbf{R} - \mathbf{I}, \tag{4.100}$$

where $\mathbf{D}_R$ is the diagonal matrix consisting of the diagonal elements of $\mathbf{R}$. In summary, the base station needs to run the following three steps:

1. Solve the optimization problem (4.95).
2. Apply the GTD algorithm to $\boldsymbol{\Psi}^{\frac{1}{2}}$ in (4.98).
3. Obtain the spreading sequences for all mobile users, $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_L]$ by (4.99), and the feed-forward and feedback matrices for the base station from (4.100).

4.7.2 Downlink Scenario

In the downlink case, the mobiles cannot cooperate with each other for decision feedback. Hence the MMSE–DFE is impractical at the receivers. However, we can apply the DPC approach as introduced in Section 4.6 to cancel known interferences at the transmitter, i.e., the base station. We can convert the downlink problem as an uplink one and exploit the downlink–uplink duality as we have done in Section 4.6.Note that $\mathbf{H} = \mathbf{H}^{\dagger} = \mathbf{I}$, i.e., the downlink and uplink channels are the same.

Consider the case where the uplink and downlink communications are symmetric, i.e., for each mobile user, the QoS of the communications from the user to the base station and the base station to the user are the same. After obtaining the spreading sequences $[\mathbf{s}_1, \ldots, \mathbf{s}_L]$ for the mobile users, and the nulling vectors $[\mathbf{w}_1, \ldots, \mathbf{w}_L]$ used at the base station for the uplink case, we immediately know that in the downlink case the spreading sequences transmitted from the base station are exactly $[\mathbf{w}_1, \ldots, \mathbf{w}_L]$ and the nulling vectors used at the mobiles are the spreading sequences, $[\mathbf{s}_1, \ldots, \mathbf{s}_L]$, used in the uplink case. The only parameters that remain to be calculated are $q_1, \ldots, q_N$ (cf. (4.92)). Hence in this symmetric case, the base station only needs to inform the mobiles their designated spreading sequences once in the two-way communication. Each mobile uses the same sequence for both data spreading in the uplink channel and interference nulling in the downlink channel. More detailed discussion on this problem can be found in [71].

4.8 Summary

This chapter has introduced the design of nonlinear DF MIMO transceivers with full CSI based on majorization theory and the generalized triangular decomposition (GTD) algorithm. Two different problem formulations have been considered: one based on a global performance measure subject to the overall input power constraint and another one based on minimizing the input power subject to individual QoS constraints.

The optimal solution is characterized as follows. The optimum DFE is always the MMSE–DFE, which can be easily computed via the QR decomposition of the augmented matrix $\binom{\mathbf{HP}}{\mathbf{I}} = \mathbf{QR}$. In a remarkably similar vein to the *linear* transceiver designs, the precoder $\mathbf{P}$ for the nonlinear DF transceiver design also has the form $\mathbf{P} = \mathbf{V}_H \,\mathrm{diag}\left(\sqrt{\mathbf{p}}\right) \boldsymbol{\Omega}^\dagger$, where $\mathbf{V}_H$ contains the right singular vectors of the channel matrix $\mathbf{H}$ (whose role is to diagonalize the channel), $\mathbf{p}$ denotes a power allocation over the eigen-subchannels, and $\boldsymbol{\Omega}$ is a unitary matrix that can be obtained using the GTD algorithm.

For the design based on some global performance measure f_0, all the cost functions for which the composite function $f_0 \circ \exp$ is Schur-convex lead to the same solution, i.e., the so-called UCD. The precoder of the UCD scheme applies the waterfilling power allocation $\mathbf{p}$ to maximize the channel mutual information and the rotation $\boldsymbol{\Omega}$ such that the DFE yields substreams with identical MSE performance. On the other hand, if $f_0 \circ \exp$ is Schur-concave, then the optimum precoder $\mathbf{P} = \mathbf{V}_H \operatorname{diag}\left(\sqrt{\mathbf{p}}\right)$ and the nonlinear DF transceiver design degenerates to linear diagonal transmission.

For the design with individual QoS constraints, we simplify the transceiver optimization problem to a power allocation problem for which we have a simple multi-level waterfilling algorithm. We have shown through several numerical examples that the DF based nonlinear designs have remarkable performance gain over the linear designs.

Interestingly, the nonlinear MIMO transceiver designs can be implemented in two forms by exploiting the duality between the DFE and DPC as well as the uplink–downlink duality. Finally, the problem of designing CDMA sequences has been shown to be a special case of the MIMO transceiver design problem.

4.A Appendix: Mutual Information and Wiener Filter

Consider Gaussian input $\mathbf{x} \sim N(0, \mathbf{I})$. For the MIMO channel given in (4.1), the instantaneous mutual information between the input and output of the channel is [147]

$$I(\mathbf{x};\mathbf{y}) = \log|\mathbf{I} + \mathbf{H}\mathbf{P}\mathbf{P}^\dagger\mathbf{H}^\dagger| \quad \text{bps/Hz}. \tag{4.101}$$

There is an interesting link between the mutual information and the MSE matrix when using the Wiener filter (or, equivalently, the linear MMSE receiver). As shown in (3.18), the Wiener filter yields an estimate of $\mathbf{x}$ with MSE matrix (see (3.21))

$$\mathbf{E} = (\mathbf{I} + \mathbf{P}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{P})^{-1}. \tag{4.102}$$

It follows from (4.101) and the equality $|\mathbf{I} + \mathbf{A}\mathbf{B}| = |\mathbf{I} + \mathbf{B}\mathbf{A}|$ that

$$I(\mathbf{x};\mathbf{y}) = -\log|\mathbf{E}|. \tag{4.103}$$

With the Wiener filter, the original MIMO channel is converted into L scalar subchannels

$$x_i = \hat{x}_i + e_i, \quad 1 \le i \le L, \tag{4.104}$$

where x_i ($\hat{x}_i$) is the ith entry of $\mathbf{x}$ ($\hat{\mathbf{x}}$) and e_i is the estimation error with zero mean and variance $[\mathbf{E}]_{ii}$, the ith diagonal element of $\mathbf{E}$.

By the orthogonality principle of the MMSE estimation theory [157], Gaussian random variables $\hat{x}_i$ and e_i are statistically independent. Hence, $\mathbb{E}[|\hat{x}_i|^2] + \mathbb{E}[|e_i|^2] = \mathbb{E}[|x_i|^2] = 1$. Therefore, the mutual information between x_i and $\hat{x}_i$ is

$$I(x_i;\hat{x}_i) = \log\left(1 + \frac{\mathbb{E}[|\hat{x}_i|^2]}{[\mathbf{E}]_{ii}}\right) = \log\left(\frac{1}{[\mathbf{E}]_{ii}}\right), \tag{4.105}$$

and the sum of the mutual information of the L subchannels is

$$\sum_{l=1}^{L} I(x_i;\hat{x}_i) = -\log\left(\prod_{l=1}^{L}[\mathbf{E}]_{ii}\right). \tag{4.106}$$

From Hadamard's inequality,

$$|\mathbf{E}| \leq \prod_{i=1}^{L}[\mathbf{E}]_{ii} \tag{4.107}$$

and the equality holds if and only if $\mathbf{E}$ is diagonal. Therefore, we see from (4.103) that

$$\sum_{l=1}^{L} I(x_i;\hat{x}_i) = -\log\left(\prod_{l=1}^{L}[\mathbf{E}]_{ii}\right) \leq -\log|\mathbf{E}| = I(\mathbf{x};\mathbf{y}). \tag{4.108}$$

In general, the MSE matrix $\mathbf{E}$ is non-diagonal. Since the Wiener filter does not exploit the spatial correlation between the estimation errors $\{e_i\}_{l=1}^{L}$, it incurs in an information loss quantified by

$$I_{\text{loss}} = \sum_{l=1}^{L} \log[\mathbf{E}]_{ii} - \log|\mathbf{E}| \quad \text{bps/Hz}. \tag{4.109}$$

Hence, the linear MMSE receiver is information lossless only if $\mathbf{E}$ is diagonal or, equivalently, if the columns of the effective channel matrix $\mathbf{G} \triangleq \mathbf{HP}$ are orthogonal.

The Wiener filter is inherently an "analog" equalizer which does not take advantage of the "digital" feature of the signal. However, in practice $\mathbf{x}$ belongs to a finite alphabet. To exploit the "digital" feature of communication signals, one may adopt DFE.

4.B Appendix: Proof of Lemma 4.1

It can be seen from (4.18) that $\mathbf{w}_i$ is the ith column of $(\mathbf{G}_i^\dagger \mathbf{G}_i + \mathbf{I})^{-1}\mathbf{G}_i$. Using the matrix inversion lemma (see Appendix B.2)

$$\begin{aligned}(\mathbf{G}_i^\dagger \mathbf{G}_i + \mathbf{I})^{-1}\mathbf{G}_i &= (\mathbf{I} - \mathbf{G}_i(\mathbf{G}_i^\dagger \mathbf{G}_i + \mathbf{I})^{-1}\mathbf{G}_i^\dagger)\mathbf{G}_i \\ &= \mathbf{G}_i(\mathbf{I} - (\mathbf{G}_i^\dagger \mathbf{G}_i + \mathbf{I})^{-1}\mathbf{G}_i^\dagger \mathbf{G}_i) \\ &= \mathbf{G}_i(\mathbf{G}_i^\dagger \mathbf{G}_i + \mathbf{I})^{-1}. \end{aligned} \tag{4.110}$$

Hence

$$\mathbf{w}_i = [\mathbf{G}_i(\mathbf{G}_i^\dagger \mathbf{G}_i + \mathbf{I})^{-1}]_{:,i}, \tag{4.111}$$

where $[\cdot]_{:,i}$ stands for the ith column of the matrix. Let $\mathbf{G}_{a,i} \in \mathbb{C}^{(n_R+L)\times i}$ be the submatrix consisting of the first i columns of $\mathbf{G}_a$. Then

$$\mathbf{G}_i(\mathbf{G}_i^\dagger \mathbf{G}_i + \mathbf{I})^{-1} = \mathbf{G}_i(\mathbf{G}_{a,i}^\dagger \mathbf{G}_{a,i})^{-1}, \tag{4.112}$$

which is a submatrix of $\mathbf{G}_{a,i}(\mathbf{G}_{a,i}^\dagger \mathbf{G}_{a,i})^{-1}$. From the QR decomposition given in (4.19), $\mathbf{G}_{a,i} = \mathbf{Q}_i\mathbf{R}_i$, where $\mathbf{Q}_i \in \mathbb{C}^{(n_R+L)\times i}$ is the submatrix of $\mathbf{Q}$, and $\mathbf{R}_i \in \mathbb{C}^{i\times i}$ is the leading principal submatrix of $\mathbf{R}$. Partitioning $\mathbf{Q}_i$ into $\mathbf{Q}_i = \begin{bmatrix}\bar{\mathbf{Q}}_i \\ \underline{\mathbf{Q}}_i\end{bmatrix}$, where $\bar{\mathbf{Q}}_i \in \mathbb{C}^{n_R\times i}$, we obtain $\mathbf{G}_i = \bar{\mathbf{Q}}_i\mathbf{R}_i$ and

$$\mathbf{G}_i(\mathbf{G}_{a,i}^\dagger \mathbf{G}_{a,i})^{-1} = \bar{\mathbf{Q}}_i\mathbf{R}_i(\mathbf{R}_i^\dagger\mathbf{R}_i)^{-1} = \bar{\mathbf{Q}}_i\mathbf{R}_i^{-\dagger}, \tag{4.113}$$

where $\mathbf{R}_i^{-\dagger}$ is a lower triangular matrix with diagonal r_{jj}^{-1}, $j = 1,\ldots,i$. It follows from (4.111), (4.112), and (4.113) that

$$\mathbf{w}_i = [\mathbf{G}_i(\mathbf{G}_{a,i}^\dagger \mathbf{G}_{a,i})^{-1}]_{:,i} = r_{ii}^{-1}\bar{\mathbf{q}}_i \tag{4.114}$$

with $\bar{\mathbf{q}}_i$ being the ith (last) column of $\bar{\mathbf{Q}}_i$. Hence it follows immediately that

$$\mathbf{W} = [\mathbf{w}_1,\ldots,\mathbf{w}_L] = \bar{\mathbf{Q}}\mathbf{D}_R^{-1}. \tag{4.115}$$

As for $\mathbf{B}$, observe from (4.19) and (4.20) that $\mathbf{G} = \bar{\mathbf{Q}}\mathbf{R}$ and $\underline{\mathbf{Q}} = \mathbf{R}^{-1}$. Thus

$$\bar{\mathbf{Q}}^\dagger\bar{\mathbf{Q}} = \mathbf{I} - \mathbf{R}^{-\dagger}\mathbf{R}^{-1}, \tag{4.116}$$

and

$$\begin{aligned}\mathbf{W}^{\dagger}\mathbf{G} &= \mathbf{D}_R^{-1}\bar{\mathbf{Q}}^{\dagger}\bar{\mathbf{Q}}\mathbf{R} \\ &= \mathbf{D}_R^{-1}(\mathbf{I} - \mathbf{R}^{-\dagger}\mathbf{R}^{-1})\mathbf{R} \\ &= \mathbf{D}_R^{-1}\mathbf{R} - \mathbf{D}_R^{-1}\mathbf{R}^{-\dagger}. \qquad (4.117)\end{aligned}$$

Note that the first term in (4.117) is upper triangular with identity diagonal and the second term is lower triangular. Hence

$$\mathbf{B} = \mathcal{U}(\mathbf{W}^{\dagger}\mathbf{G}) = \mathcal{U}(\mathbf{D}_R^{-1}\mathbf{R}) = \mathbf{D}_R^{-1}\mathbf{R} - \mathbf{I}. \qquad (4.118)$$

To derive the MSE matrix, it follows from (4.3) that the analog estimate of $\mathbf{x}$ is

$$\hat{\mathbf{x}} = (\mathbf{W}^{\dagger}\mathbf{G} - \mathbf{B})\mathbf{x} + \mathbf{W}^{\dagger}\mathbf{n}. \qquad (4.119)$$

Thus the error vector is

$$\begin{aligned}\hat{\mathbf{x}} - \mathbf{x} &= (\mathbf{W}^{\dagger}\mathbf{G} - (\mathbf{B} + \mathbf{I}))\mathbf{x} + \mathbf{W}^{\dagger}\mathbf{n}, \qquad (4.120) \\ &= -\mathbf{D}_R^{-1}\mathbf{R}^{-\dagger}\mathbf{x} + \mathbf{D}_R^{-1}\bar{\mathbf{Q}}^{\dagger}\mathbf{n}, \qquad (4.121)\end{aligned}$$

where we have used (4.117) and (4.118). The MSE matrix is

$$\begin{aligned}\mathbf{E} &= \mathbb{E}[(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^{\dagger}] \\ &= \mathbf{D}_R^{-1}\mathbf{R}^{-\dagger}\mathbf{R}^{-1}\mathbf{D}_R^{-1} + \mathbf{D}_R^{-1}\bar{\mathbf{Q}}^{\dagger}\bar{\mathbf{Q}}\mathbf{D}_R^{-1} \qquad (4.122) \\ \text{(cf. (4.116))} &= \mathbf{D}_R^{-2}, \qquad (4.123)\end{aligned}$$

which is diagonal for any channel realization.

4.C Appendix: Proof of Theorem 4.3

We first analyze how $\mathbf{\Omega}$ influences the cost function and the constraints of (4.33). Let us denote the singular values of the effective channel $\mathbf{G} \triangleq \mathbf{HP} \in \mathbb{C}^{n_R \times L}$ as $\sigma_{G,1} \geq \cdots \geq \sigma_{G,K} \geq \sigma_{G,K+1} = \cdots = \sigma_{G,L} = 0$. The singular values of the augmented matrix $\mathbf{G}_a = \begin{pmatrix}\mathbf{HP} \\ \mathbf{I}\end{pmatrix} \in \mathbb{C}^{(n_R+L)\times L}$ are easily shown to be

$$\sigma_{G_a,i} = \begin{cases}\sqrt{1 + \sigma_{G,i}^2}, & 1 \leq i \leq K, \\ 1, & K < i \leq L.\end{cases} \qquad (4.124)$$

Let $\boldsymbol{\sigma}_{G_a} \in \mathbb{R}_+^L$ be the vector consisting of $\sigma_{G_a,i}$. Clearly $\boldsymbol{\sigma}_{G_a}$ is invariant to the choice of $\boldsymbol{\Omega}$. However, $\boldsymbol{\Omega}$ determines the diagonal of $\mathbf{R}$ as shown in the following lemma:

Lemma 4.9. There exists an $\boldsymbol{\Omega}$ such that $\mathbf{G}_a = \mathbf{QR}$ if and only if

$$\{|[\mathbf{R}]_{ii}|^2\} \prec_\times \boldsymbol{\sigma}_{G_a}^2. \tag{4.125}$$

In other words, when $\boldsymbol{\Omega}$ goes over the whole Stiefel manifold [68]

$$\mathcal{S}(L;K) \triangleq \left\{\mathbf{Q} \in \mathbb{C}^{L\times K} : \mathbf{Q}^\dagger\mathbf{Q} = \mathbf{I}\right\}, \tag{4.126}$$

the achievable set of the diagonal of $\mathbf{R}$ given in (4.36) is

$$\{\mathbf{r} \in \mathbb{C}^L : |\mathbf{r}| \prec_\times \boldsymbol{\sigma}_{G_a}\}. \tag{4.127}$$

Proof. From the SVD

$$\mathbf{P} = \mathbf{U}_P \text{diag}(\sqrt{\mathbf{p}})\boldsymbol{\Omega}^\dagger = \mathbf{U}_P[\text{diag}(\sqrt{\mathbf{p}}) \,\vdots\, \mathbf{0}_{K\times(L-K)}]\boldsymbol{\Omega}_0^\dagger,$$

where $\boldsymbol{\Omega}_0 \in \mathbb{C}^{L\times L}$ is a unitary matrix whose first K columns form $\boldsymbol{\Omega}$, $\mathbf{G}_a$ may be rewritten as

$$\mathbf{G}_a = \begin{bmatrix} \mathbf{HU}_P[\text{diag}(\sqrt{\mathbf{p}}) \,\vdots\, \mathbf{0}_{K\times(L-K)}]\boldsymbol{\Omega}_0^\dagger \\ \mathbf{I}_L \end{bmatrix}. \tag{4.128}$$

We can further rewrite (4.128) as

$$\mathbf{G}_a = \begin{bmatrix} \mathbf{I}_{n_R} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_0 \end{bmatrix} \begin{bmatrix} \mathbf{HU}_P[\text{diag}(\sqrt{\mathbf{p}}) \,\vdots\, \mathbf{0}_{K\times(L-K)}] \\ \mathbf{I}_L \end{bmatrix} \boldsymbol{\Omega}_0^\dagger. \tag{4.129}$$

Note that the matrix in the second bracket in (4.129) shares the singular values with $\mathbf{G}_a$. According to Theorem 4.2, if $([\mathbf{R}]_{11},\ldots,[\mathbf{R}]_{LL}) \prec_\times \boldsymbol{\sigma}_{G_a}$, we can have the decomposition:

$$\mathbf{J} \triangleq \begin{bmatrix} \mathbf{HU}_P[\text{diag}(\sqrt{\mathbf{p}}) \,\vdots\, \mathbf{0}_{K\times(L-K)}] \\ \mathbf{I}_L \end{bmatrix} = \mathbf{Q}_J\mathbf{RP}_J^\dagger, \tag{4.130}$$

where $\mathbf{R} \in \mathbb{R}^{L\times L}$ is an upper triangular matrix with diagonal $\mathbf{r}$, $\mathbf{Q}_J \in \mathbb{C}^{(n_R+L)\times L}$ is a semi-unitary matrix, and $\mathbf{P}_J \in \mathbb{C}^{L\times L}$ is unitary. Inserting (4.130) into (4.129) yields

$$\mathbf{G}_a = \begin{bmatrix} \mathbf{I}_{n_R} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_0 \end{bmatrix} \mathbf{Q}_J \mathbf{R} \mathbf{P}_J^\dagger \mathbf{\Omega}_0^\dagger. \tag{4.131}$$

Let $\mathbf{\Omega}_0 = \mathbf{P}_J^\dagger$ (hence $\mathbf{\Omega}^\dagger$ is formed by the first K rows of $\mathbf{P}_J$) and

$$\mathbf{Q} = \begin{bmatrix} \mathbf{I}_{n_R} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_0^\dagger \end{bmatrix} \mathbf{Q}_J. \tag{4.132}$$

Then (4.131) reduces to the QR decomposition $\mathbf{G}_a = \mathbf{QR}$. Now we have proven the "if" part of this lemma.

Conversely, since the QR decomposition is a special case of the GTD, it follows immediately from the GTD Theorem that the diagonal of $\mathbf{R}$ satisfies (4.125). □

Using the above lemma, we can replace the constraint

$$\begin{pmatrix} \mathbf{HP} \\ \mathbf{I} \end{pmatrix} = \mathbf{QR}$$

in (4.33) by $\left([\mathbf{R}]_{11}^2, \ldots, [\mathbf{R}]_{LL}^2\right) \prec_\times \boldsymbol{\sigma}_{G_a}^2$. Note also that $\mathbf{PP}^\dagger$ is invariant to $\mathbf{\Omega}$. We can at this time remove $\mathbf{\Omega}$ from the optimization problem and simplify (4.33) as

$$\begin{array}{ll} \underset{\mathbf{U}_P, \mathbf{p}, [\mathbf{R}]_{ii}^2}{\text{minimize}} & f_0\left(\left\{[\mathbf{R}]_{ii}^{-2}\right\}\right) \\ \text{subject to} & \left([\mathbf{R}]_{11}^2, \ldots, [\mathbf{R}]_{LL}^2\right) \prec_\times \boldsymbol{\sigma}_{G_a}^2 \\ & \mathbf{1}^T\mathbf{p} \le P_0 \\ & \mathbf{p} \ge \mathbf{0}, \end{array} \tag{4.133}$$

where $\boldsymbol{\sigma}_{G_a}$ depends on $\mathbf{p}$ and $\mathbf{U}_P$. If we denote $r_{[i]}$ the ith largest element of $\left\{[\mathbf{R}]_{ii}^2\right\}_{i=1}^L$, (4.133) can be rewritten more explicitly as

$$\begin{array}{ll} \underset{\mathbf{U}_P, \mathbf{p}, [\mathbf{R}]_{ii}^2}{\text{minimize}} & f_0\left(\left\{[\mathbf{R}]_{ii}^{-2}\right\}\right) \\ \text{subject to} & \prod_{i=1}^k r_{[i]} \le \prod_{i=1}^k \sigma_{G_a,i}^2, \quad 1 \le i \le K-1 \\ & \prod_{i=1}^L r_{[i]} = \prod_{i=1}^K \sigma_{G_a,i}^2 \\ & \mathbf{1}^T\mathbf{p} \le P_0 \\ & \mathbf{p} \ge \mathbf{0}. \end{array} \tag{4.134}$$

Next, we show that the solution to (4.133) occurs when $\mathbf{U}_P = \mathbf{V}_H$ and the proof will be completed. Denote $\{\sigma_{H,i}\}$ and $\{\sqrt{p_i}\}$ as the singular values of $\mathbf{H}$ and $\mathbf{P}$, which are both in non-increasing ordering. According to Theorem B.4 in Appendix B

$$\begin{aligned} &\textstyle\prod_{i=1}^{k} \sigma_{G,i}^2 \leq \prod_{i=1}^{k} \sigma_{H,i}^2 p_i, \quad 1 \leq i \leq K-1 \\ &\textstyle\prod_{i=1}^{K} \sigma_{G,i}^2 = \prod_{i=1}^{K} \sigma_{H,i}^2 p_i, \end{aligned} \tag{4.135}$$

where the equality holds if and only if $\mathbf{U}_P = \mathbf{V}_H$. Denote $\mathbf{x}, \mathbf{y} \in \mathbb{R}^K$ by

$$x_i \triangleq \log(\sigma_{G,i}^2), \quad y_i \triangleq \log(\sigma_{H,i}^2 p_i), \quad 1 \leq i \leq K.$$

Then $\mathbf{x} \prec_+ \mathbf{y}$, and $\mathbf{x} = \mathbf{y}$ if and only if $\mathbf{U}_P = \mathbf{V}_H$. It is easy to prove that $\log(1+\exp(x))$ is an increasing convex function. Hence by Corollary 2.5, $\sum_{i=1}^{K} \log(1+\exp(x_i))$ is a Schur-convex function of $\mathbf{x}$. By $\mathbf{x} \prec_+ \mathbf{y}$, $x_k \leq z_k \triangleq \sum_{i=1}^{k} y_i - \sum_{i=1}^{k-1} x_i$ and

$$\sum_{i=1}^{k} \log(1+\exp(x_i)) \leq \sum_{i=1}^{k-1} \log(1+\exp(x_i)) + \log(1+\exp(z_k)). \tag{4.136}$$

Moreover, because $(x_1, \ldots, x_{k-1}, z_k) \prec_+ (y_1, \ldots, y_k)$,

$$\sum_{i=1}^{k-1} \log(1+\exp(x_i)) + \log(1+\exp(z_i)) \leq \sum_{i=1}^{k} \log(1+\exp(y_i)). \tag{4.137}$$

Hence we obtain

$$\sum_{i=1}^{k} \log(1+\exp(x_i)) \leq \sum_{i=1}^{k} \log(1+\exp(y_i)), \quad 1 \leq k \leq K, \tag{4.138}$$

and equivalently

$$\prod_{i=1}^{k} (1+\sigma_{G,i}^2) \leq \prod_{i=1}^{k} (1+\sigma_{H,i}^2 p_i), \quad 1 \leq k \leq K. \tag{4.139}$$

Since $\sigma_{G_a,i}^2 = 1 + \sigma_{G,i}^2$ (see (4.124)), it follows from (4.139) and (4.134) that given $\mathbf{p}$, the feasible set of $\{[\mathbf{R}]_{ii}^2\}_{i=1}^{L}$ is relaxed to the maximal extent when $\mathbf{U}_P = \mathbf{V}$. More precisely, if $\mathbf{p}$ and $[\mathbf{R}]_{ii}^2$ are feasible for some $\mathbf{U}_P \neq \mathbf{V}_H$, they must also be feasible for $\mathbf{U}_P = \mathbf{V}_H$. Hence the optimum solution occurs when $\mathbf{U}_P = \mathbf{V}_H$.

Using $\mathbf{P} = \mathbf{V}_H \mathrm{diag}(\sqrt{\mathbf{p}})\boldsymbol{\Omega}^\dagger$, it follows that

$$\sigma_{G,i}^2 = \sigma_{H,i}^2 p_i, \quad \text{for } 1 \leq i \leq K \tag{4.140}$$

and (4.134) is simplified to be (4.35). The theorem is proven.

4.D Appendix: Proof of Procedure in Table 4.1

The first two steps of Table 4.1 are straightforward. We only need to explain the last three steps.

By Theorem 4.3, we know that the precoder is of form $\mathbf{P} = \mathbf{V}_H$ $\mathrm{diag}(\sqrt{\mathbf{p}})\boldsymbol{\Omega}^\dagger$. We can rewrite it to be $\mathbf{P} = \mathbf{V}_H[\mathrm{diag}(\sqrt{\mathbf{p}})\vdots\mathbf{0}_{K\times(L-K)}]\boldsymbol{\Omega}_0^\dagger$, where $\boldsymbol{\Omega}_0$ is a unitary matrix. The augmented matrix

$$\begin{aligned}\begin{bmatrix}\mathbf{HP}\\ \mathbf{I}_L\end{bmatrix} &= \begin{bmatrix}\mathbf{U}_H\boldsymbol{\Sigma}_H[\mathrm{diag}(\sqrt{\mathbf{p}})\vdots\mathbf{0}_{K\times(L-K)}]\boldsymbol{\Omega}_0^\dagger\\ \mathbf{I}_L\end{bmatrix} &(4.141)\\ &= \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Omega}_0\end{bmatrix}\begin{bmatrix}\mathbf{U}_H\boldsymbol{\Sigma}_H[\mathrm{diag}(\sqrt{\mathbf{p}})\vdots\mathbf{0}_{K\times(L-K)}]\\ \mathbf{I}_L\end{bmatrix}\boldsymbol{\Omega}_0^\dagger. &(4.142)\end{aligned}$$

Define

$$\mathbf{J} \triangleq \begin{bmatrix}\mathbf{U}_H\boldsymbol{\Sigma}_H[\mathrm{diag}(\sqrt{\mathbf{p}})\vdots\mathbf{0}_{K\times(L-K)}]\\ \mathbf{I}_L\end{bmatrix}, \tag{4.143}$$

we can apply the GTD to $\mathbf{J}$ such that

$$\mathbf{J} = \mathbf{Q}_J\mathbf{R}\mathbf{P}_J^\dagger, \tag{4.144}$$

where $\mathbf{R}$ has diagonal elements given in Step 2. Such a decomposition always exists, which follows from Theorem 4.2 and the fact that the diagonal elements of $\mathbf{R}$ satisfy the constraint of (4.35)

$$(|[\mathbf{R}]_{11}|, \ldots, |[\mathbf{R}]_{LL}|) \prec_\times \left(\left\{\sqrt{1+\sigma_{H,i}^2 p_i}\right\}_{i=1}^K, 1, \ldots, 1\right),$$

where the elements of the right vector in the above equation are actually the singular values of $\mathbf{J}$.

Now let

$$\mathbf{Q} \triangleq \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_J^\dagger \end{bmatrix} \mathbf{Q}_J \tag{4.145}$$

and $\mathbf{\Omega}_0 = \mathbf{P}_J^\dagger$. We have from (4.142) and (4.144) that

$$\begin{bmatrix} \mathbf{HP} \\ \mathbf{I} \end{bmatrix} = \mathbf{QR}. \tag{4.146}$$

By Lemma 4.1, we see that the feedback filter $\mathbf{B} = \mathbf{D}_R^{-1}\mathbf{R} - \mathbf{I}$ and the feed-forward filter $\mathbf{W} = \bar{\mathbf{Q}}\mathbf{D}_R^{-1}$, where $\bar{\mathbf{Q}}$ is the submatrix consisting the first n_R rows of $\mathbf{Q}$. Note from (4.145) that the first n_R rows of $\mathbf{Q}$ and $\mathbf{Q}_J$ are the same. Table 4.1 is justified.

4.E Appendix: Proof of Power Allocation for QoS ©2006 IEEE. Reprinted with permission from IEEE

4.E.1 Proof of Lemma 4.6

Proof. We replace the inequalities in (4.79) by the equivalent constraints obtained by taking log's:

$$\sum_{i=1}^{k} \log(\psi_i) \geq \sum_{i=1}^{k} \log(\beta_i), \quad 1 \leq k \leq K.$$

The Lagrangian $\mathcal{L}$ associated with (4.79), after this modification of the constraints, is

$$\mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\mu}) = \sum_{k=1}^{K} \left(\psi_k - \mu_k \sum_{i=1}^{k} (\log(\psi_i) - \log(\beta_i)) \right).$$

By the first-order optimality conditions associated with $\boldsymbol{\psi}$, there exists $\boldsymbol{\mu} \geq \mathbf{0}$ with the property that the gradient of the Lagrangian with respect to $\boldsymbol{\psi}$ vanishes. Equating to zero the partial derivative of the Lagrangian with respect to ψ_j, we obtain the relations

$$\frac{\partial \mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\mu})}{\partial \psi_j} = 1 - \frac{1}{\psi_j} \sum_{i=j}^{K} \mu_i = 0, \quad j = 1, \ldots, K.$$

Thus

$$\psi_j = \sum_{i=j}^{K} \mu_i, \quad j = 1, \ldots, K.$$

Hence, $\psi_j - \psi_{j+1} = \mu_j \geq 0$. □

4.E.2 Proof of Lemma 4.7

Proof. If $\boldsymbol{\psi}$ is a solution to (4.78) with the property that $\psi_i > \frac{1}{\sigma_{H,i}^2}$ for all $1 \leq i \leq K$, then by the convexity of the constraints, it follows that $\boldsymbol{\psi}$ is a solution of (4.79). By Lemma 4.6, we conclude that Lemma 4.7 holds with $j = K$. Now, suppose that $\boldsymbol{\psi}$ is a solution of (4.78) with $\psi_i = 1/\sigma_{H,i}^2$ for some i. We wish to show that $\psi_k = 1/\sigma_{H,k}^2$ for all $k > i$. Suppose, to the contrary, that there exists an index $k \geq i$ with the property that $\psi_k = 1/\sigma_{H,k}^2$ and $\psi_{k+1} > 1/\sigma_{H,k+1}^2$. We show that components k and $k+1$ of $\boldsymbol{\psi}$ can be modified so as to satisfy the constraints and make the cost strictly smaller. In particular, let $\boldsymbol{\psi}(\epsilon)$ be identical with $\boldsymbol{\psi}$ except for components k and $k+1$:

$$\psi_k(\epsilon) = (1+\epsilon)\psi_k \quad \text{and} \quad \psi_{k+1}(\epsilon) = \frac{\psi_{k+1}}{1+\epsilon}. \tag{4.147}$$

For a small $\epsilon > 0$, $\boldsymbol{\psi}(\epsilon)$ satisfies the constraints of (4.78). The change $\Delta(\epsilon)$ in the cost function of (4.78) is

$$\Delta(\epsilon) = (1+\epsilon)\psi_k + \frac{\psi_{k+1}}{1+\epsilon} - \psi_k - \psi_{k+1}.$$

The derivative of $\Delta(\epsilon)$ evaluated at zero is

$$\Delta'(0) = \psi_k - \psi_{k+1}.$$

Since $1/\sigma_{H,k}^2$ is an increasing function of k and since $\psi_k = 1/\sigma_{H,k}^2$, we conclude that $\psi_{k+1} > \psi_k$ and $\Delta'(0) < 0$. Hence, for $\epsilon > 0$ near zero, $\boldsymbol{\psi}(\epsilon)$ has a smaller cost than $\boldsymbol{\psi}(0)$, which yields a contradiction. Hence, there exists an index j with the property that $\psi_i = 1/\sigma_{H,i}^2$ for all $i > j$ and $\psi_i > 1/\sigma_{H,i}^2$ for all $i \leq j$.

According to Lemma 4.6, $\psi_i \geq \psi_{i+1}$for any $i < j$. To complete the proof, we need to show that $\psi_j \leq \psi_{j+1}$. As noted previously, any solution of (4.78) satisfies

$$\prod_{i=1}^{K} \psi_i = \prod_{i=1}^{K} \beta_i,$$

which implies (cf. (4.77))

$$\prod_{i=1}^{j} \psi_i = \prod_{i=1}^{j} \beta_i \left(\prod_{i=j+1}^{K} \beta_i \sigma_{H,i}^2 \right) > \prod_{i=1}^{j} \beta_i.$$

That is, the constraint $\prod_{i=1}^{j} \psi_i \geq \prod_{i=1}^{j} \beta_i$ in (4.78) is inactive. If $\psi_j > \psi_{j+1}$, we will decrease the jth component and increase the $j+1$ component, while leaving the other components unchanged. Letting $\boldsymbol{\psi}(\delta)$ be the modified vector, we set

$$\psi_{j+1}(\delta) = (1+\delta)\psi_{j+1} \quad \text{and} \quad \psi_j(\delta) = \frac{\psi_j}{1+\delta}.$$

Since the jth constraint in (4.78) is inactive, $\boldsymbol{\psi}(\delta)$ is feasible for δ near zero. And if $\psi_j > \psi_{j+1}$, then the cost decreases as δ increases. It follows that $\psi_j \leq \psi_{j+1}$. □

4.E.3 Proof of Lemma 4.8

Proof. First suppose that $\gamma_l \geq 1/\sigma_l^2$. By the arithmetic/geometric mean inequality, the problem

$$\min \sum_{i=1}^{l} \psi_i \quad \text{subject to} \quad \prod_{i=1}^{l} \psi_i \geq \prod_{i=1}^{l} \beta_i, \quad \boldsymbol{\psi} \geq \mathbf{0}, \tag{4.148}$$

has the solution $\psi_i = \gamma_l$, $1 \leq i \leq l$. Since $\sigma_{H,i}$ is a decreasing function of i and $\gamma_l \geq 1/\sigma_{H,i}^2$, we conclude that $\psi_i = \gamma_l$ satisfies the constraints $\psi_i \geq 1/\sigma_{H,i}^2$ for $1 \leq i \leq l$. Since l attains the maximum in (4.81),

$$\gamma_l^k \geq \prod_{i=1}^{k} \beta_k$$

for all $k \leq l$. Hence, by taking $\psi_i = \gamma_l$ for $1 \leq i \leq l$, the first l inequalities in (4.78) are satisfied, with equality for $k = l$, and the first l lower bound constraints $\psi_i \geq 1/\sigma_{H,i}^2$ are satisfied.

Let $\boldsymbol{\psi}^\dagger$ denote any optimal solution of (4.78). If

$$\prod_{i=1}^{l} \psi_i^\dagger = \prod_{i=1}^{l} \beta_i, \tag{4.149}$$

then by the unique optimality of $\psi_i = \gamma_l$, $1 \le i \le l$, in (4.148), and by the fact that the inequality constraints in (4.78) are satisfied for $k \in [1, l]$, we conclude that $\psi_i^\dagger = \gamma_l$ for all $i \in [1, l]$. On the other hand, suppose that

$$\prod_{i=1}^{l} \psi_i^\dagger > \prod_{i=1}^{l} \beta_l = \gamma_l^l. \tag{4.150}$$

We show below that this leads to a contradiction; consequently, (4.149) holds and $\psi_i^\dagger = \gamma_l$ for $i \in [1, l]$.

Define the quantity

$$\gamma_* = \left(\prod_{i=1}^{l} \psi_i^\dagger \right)^{1/l}.$$

By (4.150) $\gamma_* > \gamma_l$. Again, by the arithmetic/geometric mean inequality, the solution of the problem

$$\min \sum_{i=1}^{l} \psi_i \quad \text{subject to} \quad \prod_{i=1}^{l} \psi_i \ge \gamma_*^l, \;\; \boldsymbol{\psi} \ge \mathbf{0}, \tag{4.151}$$

is $\psi_i = \gamma_*$ for $i \in [1, l]$. By (4.150), $\gamma_* > \gamma_l$ and $\boldsymbol{\psi}$ satisfies the inequality constraints in (4.78) for $k \in [1, l]$.

Let M be the smallest index with the property that

$$\prod_{i=1}^{M} \psi_i^\dagger = \prod_{i=1}^{M} \beta_i. \tag{4.152}$$

Such an index exists since $\boldsymbol{\psi}^\dagger$ is optimal, which implies that

$$\prod_{i=1}^{K} \psi_i^\dagger = \prod_{i=1}^{K} \beta_i.$$

First, suppose that $M \le j$, where j is the break point given in Lemma 4.7. By complementary slackness, $\mu_i = 0$ and $\psi_i^\dagger - \psi_{i+1}^\dagger = \mu_i$

for $1 \leq i < M$. We conclude that $\psi_i = \gamma_*$ for $1 \leq i \leq M$. By (4.152) we have

$$\gamma_*^M = \prod_{i=1}^{M} \beta_i.$$

It follows that

$$\left(\prod_{i=1}^{M} \beta_i \right)^{1/M} = \gamma_* > \gamma_l,$$

which contradicts the fact that l achieves the maximum in (4.81).

In the case $M > j$, we have $\psi_i = \gamma_*$ for $1 \leq i \leq j$. Again, this follows by complementary slackness. However, we need to stop when $i = j$ since the lower bound constraints become active for $i > j$. In Lemma 4.7, we show that $\psi_i^\dagger \geq \psi_j^\dagger = \gamma_*$ for $i \geq j$. Consequently, we have

$$\prod_{i=1}^{M} \beta_i = \prod_{i=1}^{M} \psi_i^\dagger \geq \gamma_*^M > \gamma_l^M.$$

Again, this contradicts the fact that l achieves the maximum in (4.81). This completes the analysis of the case where $\gamma_l \geq 1/\sigma_{H,l}^2$.

Now consider the case $\gamma_l < 1/\sigma_{H,l}^2$. By the definition of γ_l, we have

$$\gamma_l \geq \left(\prod_{i=1}^{K} \beta_i \right)^{1/K} \quad \text{or} \quad \gamma_l^K \geq \prod_{i=1}^{K} \beta_i. \tag{4.153}$$

If j is the break point described in Lemma 4.7, then $\psi_i^\dagger \geq \psi_j^\dagger$ for all i; it follows that

$$\prod_{i=1}^{K} \psi_i^\dagger \geq \left(\psi_j^\dagger \right)^K. \tag{4.154}$$

Since the product of the components of $\boldsymbol{\psi}^\dagger$ is equal to the product of the components of $\boldsymbol{\beta}$, from (4.153) and (4.154) we get

$$\gamma_l^K \geq \prod_{i=1}^{K} \beta_i = \prod_{i=1}^{K} \psi_i^\dagger \geq \left(\psi_j^\dagger \right)^K.$$

Hence, $\gamma_l \geq \psi_j^\dagger \geq 1/\sigma_{H,j}^2 \geq 1/\sigma_{H,i}^2$ for all $i \leq j$. In particular, if $l \leq j$, then $\gamma_l \geq 1/\sigma_{H,l}^2$, or, $l > j$ when $\gamma_l < 1/\sigma_{H,l}^2$. As a consequence, $\psi_l^\dagger = \frac{1}{\sigma_{H,l}^2}$. □

5

Extensions and Future Lines of Research

This text has developed a unified framework, based on majorization theory, for the optimal design of linear and nonlinear decision-feedback MIMO transceivers in point-to-point MIMO communication systems with perfect CSI. However, there are still many extensions and future lines of research related to this problem, some of which are briefly described in this chapter.[1]

5.1 Multiuser Systems

This text has dealt with point-to-point MIMO communication systems which includes single-user systems, multiuser systems with an orthogonal access (by which a series of single-user systems is effectively obtained), CDMA systems, and multiuser systems where joint processing is possible at both sides of the link. However, the most general case of a multiuser system without an orthogonal access and where joint processing among the users is not possible still remains an open problem. As in the single-user case, if one can assume the use

[1] This chapter corresponds to the state of the art as of mid 2007 and, hopefully, it will soon become obsolete.

of sufficiently long and good codes, then the problem simplifies drastically, especially when the design is based on the maximization of the sum-rate [174].

It is worth mentioning that a pragmatic approach to deal with multiuser systems consists of employing single-user designs for the users of the network in an iterative manner [9] or iteratively optimizing the receivers and the transmitters [138], but a global optimum may not be reached. For the special case of multiuser systems with single antennas (or multiple antennas only on one side of the link exploited via beamforming), the reader is referred to [135] for a tutorial on QoS-based resource allocation.

Consider a MIMO multiple-access channel (MAC) (see Figure 5.1) with U users. The received signal can be written as

$$\mathbf{y} = \sum_{u=1}^{U} \mathbf{H}_u \mathbf{s}_u + \mathbf{n}, \tag{5.1}$$

where $\mathbf{H}_u$ is the $n_R \times n_u$ channel matrix between the uth user and the receiver and $\mathbf{s}_u$ is the signal transmitted by the uth user given by $\mathbf{s}_u = \mathbf{P}_u \mathbf{x}_u$, where $\mathbf{P}_u$ is the linear precoder and $\mathbf{x}_u$ contains the L_u data symbols to be transmitted. The receiver will estimate the data from each user with a linear receiver $\mathbf{W}_u$ as $\hat{\mathbf{x}}_u = \mathbf{W}_u^\dagger \mathbf{y}$. The MSE matrix of the uth user is defined as

$$\begin{aligned} \mathbf{E}_u &\triangleq \mathbb{E}[(\hat{\mathbf{x}}_u - \mathbf{x}_u)(\hat{\mathbf{x}}_u - \mathbf{x}_u)^\dagger] && (5.2) \\ &= \left(\mathbf{W}_u^\dagger \mathbf{H}_u \mathbf{P}_u - \mathbf{I}\right)\left(\mathbf{P}_u^\dagger \mathbf{H}_u^\dagger \mathbf{W}_u - \mathbf{I}\right) + \mathbf{W}_u^\dagger \mathbf{R}_u \mathbf{W}_u, && (5.3) \end{aligned}$$

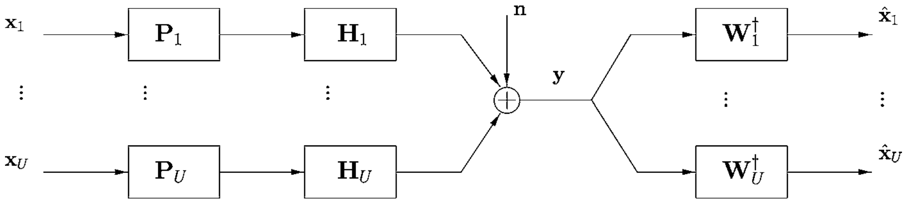


Fig. 5.1 Scheme of a MIMO MAC communication system with linear transceivers.

where $\mathbf{R}_u$ is the interference-plus-noise covariance matrix seen by the uth user:

$$\mathbf{R}_u = \sum_{l \neq u} \mathbf{H}_l \mathbf{P}_l \mathbf{P}_l^\dagger \mathbf{H}_l^\dagger + \mathbf{R}_n. \tag{5.4}$$

The design problem can be formulated as the minimization of a general cost function of the MSEs of the different users subject to a power constraint per user as given next.[2]

Problem 1 (MIMO–MAC transceivers):

$$\begin{array}{lll} \underset{\{\mathbf{P}_u, \mathbf{W}_u\}}{\text{minimize}} & f_0\left(\{\mathbf{d}(\mathbf{E}_u)\}_{u=1}^U\right) & \\ \text{subject to} & \operatorname{Tr}\left(\mathbf{P}_u \mathbf{P}_u^\dagger\right) \leq P_u & 1 \leq u \leq U, \end{array} \tag{5.5}$$

where P_u is the maximum power for the uth user.

The additional difficulty with respect to the single-user formulation in Chapter 3 is the fact that the MSE matrix contains now the signals from the other users as interference, coupling the whole problem. The design of each of the receivers, however, is still decoupled (assuming that the cost function is increasing in the MSEs) and is again given by the Wiener filter:

$$\mathbf{W}_u = \left(\mathbf{H}_u \mathbf{P}_u \mathbf{P}_u^\dagger \mathbf{H}_u^\dagger + \mathbf{R}_u\right)^{-1} \mathbf{H}_u \mathbf{P}_u \tag{5.6}$$

and the resulting MSE matrix is (similarly to (3.20))

$$\mathbf{E}_u = \mathbf{I} - \mathbf{P}_u^\dagger \mathbf{H}_u^\dagger \left(\mathbf{H}_u \mathbf{P}_u \mathbf{P}_u^\dagger \mathbf{H}_u^\dagger + \mathbf{R}_u\right)^{-1} \mathbf{H}_u \mathbf{P}_u = \left(\mathbf{I} + \mathbf{P}^\dagger \mathbf{H}^\dagger \mathbf{R}_u^{-1} \mathbf{H} \mathbf{P}\right)^{-1}. \tag{5.7}$$

For the particular case in which the design is based on minimizing the sum of MSEs of all users and substreams $\sum_{u=1}^U \operatorname{Tr}(\mathbf{E}_u)$ and, for each user, the number of substreams is not less than the number of transmit antennas, i.e., $L_u \geq n_u$, Problem 1 was solved in [78, 94] as shown next.

[2] A similar formulation can be considered for the broadcast (downlink) channel.

First rewrite the objective function in terms of the transmitters $\{\mathbf{P}_u\}$ as

$$\begin{aligned}
&\sum_{u=1}^{U} \mathrm{Tr}\left(\mathbf{I} - \mathbf{P}_u^\dagger \mathbf{H}_u^\dagger \big(\mathbf{H}_u \mathbf{P}_u \mathbf{P}_u^\dagger \mathbf{H}_u^\dagger + \mathbf{R}_u\big)^{-1} \mathbf{H}_u \mathbf{P}_u\right) \\
&\quad = \sum_{u=1}^{U} L_u - \sum_{u=1}^{U} \mathrm{Tr}\left(\left(\sum_{l=1}^{U} \mathbf{H}_l \mathbf{P}_l \mathbf{P}_l^\dagger \mathbf{H}_l^\dagger + \mathbf{R}_n\right)^{-1} \mathbf{H}_u \mathbf{P}_u \mathbf{P}_u^\dagger \mathbf{H}_u^\dagger\right) \\
&\quad = \sum_{u=1}^{U} L_u - n_R + \mathrm{Tr}\left(\sum_{u=1}^{U} \mathbf{H}_u \mathbf{P}_u \mathbf{P}_u^\dagger \mathbf{H}_u^\dagger + \mathbf{R}_n\right)^{-1} \mathbf{R}_n. \qquad (5.8)
\end{aligned}$$

Ignoring the constant term, the problem can be reformulated as

$$\begin{array}{ll}
\underset{\{\mathbf{P}_u\}}{\text{minimize}} & \mathrm{Tr}\left(\sum_{u=1}^{U} \mathbf{H}_u \mathbf{P}_u \mathbf{P}_u^\dagger \mathbf{H}_u^\dagger + \mathbf{R}_n\right)^{-1} \mathbf{R}_n \\
\text{subject to} & \mathrm{Tr}\left(\mathbf{P}_u \mathbf{P}_u^\dagger\right) \le P_u \quad 1 \le u \le U,
\end{array} \qquad (5.9)$$

which is still a complicated nonconvex problem. At this point, we can define the transmit covariance matrix of the uth user as

$$\mathbf{\Phi}_u \triangleq \mathbf{P}_u \mathbf{P}_u^\dagger \qquad (5.10)$$

with the implication that the rank of $\mathbf{\Phi}_u$ is upper bounded by the number of substreams L_u which, in general, may be smaller than the number of transmit antennas n_u (i.e., the dimension of $\mathbf{\Phi}_u$). Such a constraint is nonconvex and would complicate the problem. In [94], problem (5.9) was solved for the case where $L_u \ge n_u$. In that case, matrix $\mathbf{\Phi}_u$ does not have a rank constraint and simply needs to satisfy $\mathrm{Tr}(\mathbf{\Phi}_u) \le P_u$ and $\mathbf{\Phi}_u \ge 0$. The problem can then be formulated as

$$\begin{array}{lll}
\underset{\{\mathbf{\Phi}_u\},\mathbf{\Psi}}{\text{minimize}} & \mathrm{Tr}\left(\mathbf{\Psi} \mathbf{R}_n\right) & \\
\text{subject to} & \mathbf{\Psi} \ge \left(\sum_{u=1}^{U} \mathbf{H}_u \mathbf{\Phi}_u \mathbf{H}_u^\dagger + \mathbf{R}_n\right)^{-1} & \\
& \mathrm{Tr}\left(\mathbf{\Phi}_u\right) \le P_u & 1 \le u \le U \\
& \mathbf{\Phi}_u \ge 0. &
\end{array} \qquad (5.11)$$

Now, using the Schur-complement (cf. Appendix B), we can rewrite the constraint $\mathbf{\Psi} \ge \left(\sum_{u=1}^{U} \mathbf{H}_u \mathbf{\Phi}_u \mathbf{H}_u^\dagger + \mathbf{R}_n\right)^{-1}$ as the following linear matrix inequality:

$$\begin{bmatrix} \mathbf{\Psi} & \mathbf{I} \\ \mathbf{I} & \sum_{u=1}^{U} \mathbf{H}_u \mathbf{\Phi}_u \mathbf{H}_u^\dagger + \mathbf{R}_n \end{bmatrix} \ge \mathbf{0}. \qquad (5.12)$$

Hence, the problem can be finally reformulated in convex form as a semidefinite program (SDP) which can be optimally solved [20].

A similar formulation to Problem 1 can be taken for the broadcast channel, i.e., the downlink channel where a common transmitter sends information to distributed receivers. For the particular case where the cost function is the sum of MSEs of all users and substreams, and under the assumption $L_u \geq n_u$, the problem was solved in [49] for the case of common information sent to all the users and in [136] for the case of independent information sent to the users via a downlink–uplink duality.

Decision feedback MIMO transceivers can be similarly considered in the formulation of Problem 1.

5.2 Robust Designs for Imperfect CSI

This text has assumed perfect CSI at both sides of the link. While this assumption may be acceptable in wired systems or even wireless systems with low mobility, in many practical scenarios knowledge of the CSI is doomed to be imperfect due to estimation errors, feedback quantization, and/or feedback delay. We next show how to extend the formulation considered in this text to obtain robust designs to cope with imperfect CSI.

The basic idea is to model the current channel state $\mathbf{H}$ as

$$\mathbf{H} = \hat{\mathbf{H}} + \boldsymbol{\Delta}, \tag{5.13}$$

where $\hat{\mathbf{H}}$ is the channel estimation and $\boldsymbol{\Delta}$ denotes the error. A naive approach consists of using the estimation $\hat{\mathbf{H}}$ and proceed naively as if that was the real channel. The robust approach, on the other hand, takes into account the existence of the error $\boldsymbol{\Delta}$. To be more general, a different estimation should be considered at the transmitter $\hat{\mathbf{H}}_T$ and at the receiver $\hat{\mathbf{H}}_R$ as illustrated in Figure 5.2; if we let $\hat{\mathbf{H}}_R = \hat{\mathbf{H}}_T$ we get the same channel estimation at both sides, whereas if we let $\hat{\mathbf{H}}_R = \mathbf{H}$ we obtain the case of perfect CSIR.

There are two completely different philosophies to model the error: the *worst-case approach* (e.g., [10, 114]) and the *Bayesian approach* (e.g. [98, 123]). The former assumes that the error belongs to some set,

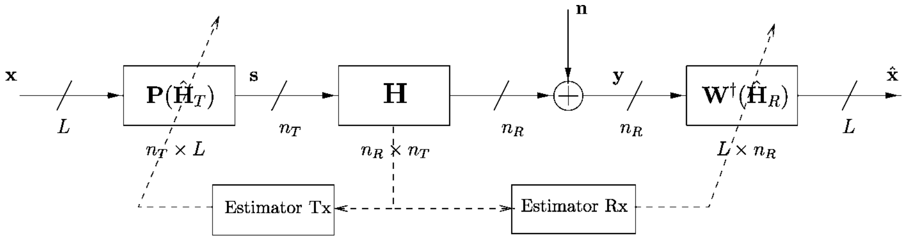


Fig. 5.2 Scheme of a robust MIMO communication system with a linear transceiver.

for example, that it has a bounded norm: $\|\mathbf{\Delta}\| \leq \epsilon$, and the design is based on the worst possible error (compare with (3.13)):

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W}}{\text{minimize}} & \sup_{\|\mathbf{\Delta}\|\leq\epsilon} f_0\left(\{\text{MSE}_i\}\right) \\ \text{subject to} & \text{Tr}\left(\mathbf{P}\mathbf{P}^\dagger\right) \leq P_0. \end{array} \tag{5.14}$$

Some results along this line can be found in [108, 114]. However, this approach tends to be overly pessimistic in practice.

The Bayesian approach models the error statistically according to some distribution such as Gaussian with zero mean and some given correlation, i.e., $\text{vec}(\Delta) \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_T \otimes \mathbf{R}_R)$. The typical approach is to consider the average performance (*stochastic approach*) of the cost function as formulated next.

Problem 2 (Stochastic MIMO transceiver):

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W}}{\text{minimize}} & \mathbb{E}_{\mathbf{\Delta}}\left[f_0\left(\{\text{MSE}_i\}\right)\right] \\ \text{subject to} & \text{Tr}\left(\mathbf{P}\mathbf{P}^\dagger\right) \leq P_0. \end{array} \tag{5.15}$$

This problem is extremely difficult as it is generally impossible to find a closed-form expression of $\mathbb{E}_{\mathbf{\Delta}}\left[f_0\left(\{\text{MSE}_i\}\right)\right]$ for an arbitrary cost function. An alternative more tractable formulation is based on considering a cost function of the averaged MSEs as shown next.

Problem 3 (Stochastic MIMO transceiver):

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W}}{\text{minimize}} & f_0\left(\{\mathbb{E}_{\mathbf{\Delta}}\left[\text{MSE}_i\right]\}\right) \\ \text{subject to} & \text{Tr}\left(\mathbf{P}\mathbf{P}^\dagger\right) \leq P_0. \end{array} \tag{5.16}$$

Both problems coincide when the cost function is linear (in fact, for a linear cost function, both problems are rather simple).

Observe that Problems 2 and 3 are implicitly assuming imperfect CSI at both sides of the link. If instead CSI is perfect at the receiver, then the optimal receiver is, of course, the Wiener filter (see Section 3.3) and the remaining optimization is only with respect to the transmitter **P**.

Initial results on Problem 3 were derived in [177]. For example, in the case of imperfect CSI at both sides, it can be shown that

$$\mathbb{E}\left[\mathbf{H}\mathbf{P}\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\right] = \hat{\mathbf{H}}\mathbf{P}\mathbf{P}^{\dagger}\hat{\mathbf{H}}^{\dagger} + \operatorname{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\mathbf{R}_T\right)\mathbf{R}_R. \tag{5.17}$$

Therefore, the expected value of the MSE matrix is (from (3.9))

$$\begin{aligned}\mathbb{E}\left[\mathbf{E}\right] \;=\; & \mathbf{W}^{\dagger}\left(\hat{\mathbf{H}}\mathbf{P}\mathbf{P}^{\dagger}\hat{\mathbf{H}}^{\dagger} + \operatorname{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\mathbf{R}_T\right)\mathbf{R}_R + \mathbf{I}\right)\mathbf{W} \\ & + \mathbf{I} - \mathbf{W}^{\dagger}\hat{\mathbf{H}}\mathbf{P} - \mathbf{P}^{\dagger}\hat{\mathbf{H}}^{\dagger}\mathbf{W}\end{aligned} \tag{5.18}$$

from which the optimal receiver is the modified Wiener filter

$$\mathbf{W} = \tilde{\mathbf{R}}_n^{-1}\hat{\mathbf{H}}\mathbf{P}\left(\mathbf{I} + \mathbf{P}^{\dagger}\hat{\mathbf{H}}^{\dagger}\tilde{\mathbf{R}}_n^{-1}\hat{\mathbf{H}}\mathbf{P}\right)^{-1}, \tag{5.19}$$

where $\tilde{\mathbf{R}}_n = \mathbf{I} + \operatorname{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\mathbf{R}_T\right)\mathbf{R}_R$ is the equivalent noise covariance matrix, and the resulting averaged MSE matrix is

$$\mathbf{E}^{\mathrm{ave}} = \left(\mathbf{I} + \mathbf{P}^{\dagger}\hat{\mathbf{H}}^{\dagger}\tilde{\mathbf{R}}_n^{-1}\hat{\mathbf{H}}\mathbf{P}\right)^{-1}. \tag{5.20}$$

The design of the transmitter **P** can be done similarly as in Chapter 3, where now the effective noise covariance matrix has increased from **I** to $\mathbf{I} + \operatorname{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\mathbf{R}_T\right)\mathbf{R}_R$ (for the case $\mathbf{R}_T = \mathbf{I}$ and with designs with a power constraint, the term $\operatorname{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\mathbf{R}_T\right)$ becomes the fixed quantity P_0 and then $\tilde{\mathbf{R}}_n = \mathbf{I} + P_0\mathbf{R}_R$).

An alternative emerging approach to deal with uncertainties is the so-called *chance programming* or probabilistic modeling [103]. It is actually a Bayesian approach that considers the performance that can be achieved with some given probability of $1 - p$ (or outage probability of p) rather than the average (e.g., [103, 126]) as formulated next.

Problem 4 (Outage MIMO transceiver):

$$\begin{array}{ll}\underset{t,\mathbf{P},\mathbf{W}}{\text{minimize}} & t \\ \text{subject to} & \Pr_{\boldsymbol{\Delta}}\left[f_0\left(\{\mathrm{MSE}_i\}\right) \leq t\right] \geq 1 - p \\ & \operatorname{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \leq P_0.\end{array} \tag{5.21}$$

This formulation presents a good tradeoff between the worst-case approach, which is too pessimistic, and the average performance approach, which is too optimistic and loose.

The formulations in Problems 2–4 can be similarly considered for the case of decision feedback MIMO transceivers.

5.3 ML Decoding

This text has considered linear and decision-feedback receivers, both of which have a low complexity and are convenient for practical implementation. An alternative with better performance but prohibitive complexity is the ML receiver. Interestingly, due to the recent developments on efficient implementation of ML decoding such as sphere decoding [34] and semidefinite relaxation [95], it is now feasible to consider the use of an ML receiver in some cases.

The design of a linear precoder assuming an ML receiver is a difficult problem and has not been solved thus far. Some approximate and partial solutions exist based on minimizing the error probability and the minimum distance of the constellation as described next.

Problem 5 (Minimum-P_e ML MIMO transceiver):

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W}}{\text{minimize}} & P_e\left(\mathbf{P},\mathbf{W}\right) \\ \text{subject to} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \leq P_0, \end{array} \tag{5.22}$$

where $P_e\left(\mathbf{P},\mathbf{W}\right)$ denotes the error probability for a given constellation when using linear precoder $\mathbf{P}$ and linear equalizer $\mathbf{W}$.

Since the minimization of the probability of error under ML decoding is difficult to deal with, because it is rarely solvable in closed form, it is reasonable to consider a design based on the maximization of the minimum distance between hypotheses as an indirect way to reduce the probability of error.

Problem 6 (Maximum-distance ML MIMO transceiver):

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W}}{\text{maximize}} & \min_{i\neq j}\left(\mathbf{x}^{(i)}-\mathbf{x}^{(j)}\right)^{\dagger}\boldsymbol{\Gamma}\left(\mathbf{P},\mathbf{W}\right)\left(\mathbf{x}^{(i)}-\mathbf{x}^{(j)}\right) \\ \text{subject to} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \leq P_0, \end{array} \tag{5.23}$$

where $\mathbf{x}^{(i)}$ is the ith symbol vector of the given transmit vector constellation and $\mathbf{\Gamma}(\mathbf{P},\mathbf{W}) = \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{W}\left(\mathbf{W}^{\dagger}\mathbf{R}_n\mathbf{W}\right)^{-1}\mathbf{W}^{\dagger}\mathbf{H}\mathbf{P}$ is the SNR-like matrix.

Unfortunately, this problem is still very complicated as the solution depends on the symbol alphabet. The optimal solution for the particular case of transmitting $L = 2$ BPSK/QPSK symbols over a MIMO channel was derived in [31]. For example, for two BPSK transmitted symbols, the optimal linear precoder admits a very simple expression:

$$\mathbf{P} = \sqrt{\frac{P_0}{2}}\mathbf{V}_H \begin{bmatrix} 1 & i \\ 0 & 0 \end{bmatrix}, \tag{5.24}$$

where $\mathbf{V}_H$ is a (semi-)unitary matrix with columns equal to the right singular vectors of the channel matrix corresponding to the L largest singular values. Unfortunately, the approach in [31] does not seem to extend easily to the more general case of transmitting L symbols from *arbitrary* constellations as it is based on a detailed analysis of all the possible combinations of transmit symbols.

As a practical solution, it was proposed in [132] to maximize a lower bound on the minimum distance based on

$$\begin{aligned}\min_{i\neq j}&(\mathbf{x}^{(i)} - \mathbf{x}^{(j)})^{\dagger}\mathbf{\Gamma}(\mathbf{P},\mathbf{W})(\mathbf{x}^{(i)} - \mathbf{x}^{(j)}) \\ &\geq \lambda_{\min}(\mathbf{\Gamma}(\mathbf{P},\mathbf{W}))\min_{i\neq j}||\mathbf{x}^{(i)} - \mathbf{x}^{(j)}||^2.\end{aligned} \tag{5.25}$$

The problem formulation becomes

$$\begin{array}{ll}\underset{\mathbf{P},\mathbf{W}}{\text{maximize}} & \lambda_{\min}(\mathbf{\Gamma}(\mathbf{P},\mathbf{W})) \\ \text{subject to} & \mathrm{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \leq P_0.\end{array} \tag{5.26}$$

The optimum receive matrix $\mathbf{W}$ can be any matrix composed of an invertible matrix and a matched filter to the received equivalent channel matrix $\mathbf{HP}$, for example, the Wiener filter $\mathbf{W} = \mathbf{HP}\left(\mathbf{I} + \mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{HP}\right)^{-1}$. An optimum transmit matrix is $\mathbf{P} = \mathbf{V}_H\mathbf{\Sigma}$, where $\mathbf{\Sigma} = \mathrm{diag}\left(\sqrt{\mathbf{p}}\right)$ is a diagonal matrix containing the square-root of the power allocation $\mathbf{p}$ given by

$$p_i = P_0 \frac{\lambda_{H,i}^{-1}}{\sum_j \lambda_{H,j}^{-1}} \quad 1 \leq i \leq L. \tag{5.27}$$

5.4 Information-Theoretic Approach

A fundamental way to design the MIMO transceiver is by maximizing the mutual information. The solution to such a problem is known since Shannon (cf. Section 1.4), but it carries the implication of ideal Gaussian coding. The requirement of Gaussian signaling can be relaxed by imposing some *a priori* realistic finite-order constellation and allowing the system to use a powerful code on top of the constellation.

Problem 7 (Maximum-I MIMO transceiver):

$$\begin{array}{ll} \underset{\mathbf{P},\mathbf{W}}{\text{maximize}} & I\left(\mathbf{P},\mathbf{W}\right) \\ \text{subject to} & \operatorname{Tr}\left(\mathbf{P}\mathbf{P}^{\dagger}\right) \leq P_0, \end{array} \tag{5.28}$$

where $I\left(\mathbf{P},\mathbf{W}\right)$ denotes the mutual information for a given constellation when using linear precoder $\mathbf{P}$ and linear equalizer $\mathbf{W}$.

This problem is extremely challenging and remains unsolved. For the particular case of a diagonal MIMO channel and transceiver, i.e., a diagonal transmission, the problem was solved in [92]. Of course, for the case of Gaussian signaling, the mutual information becomes the well-known expression $\log\det\left(\mathbf{I}+\mathbf{P}^{\dagger}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{P}\right)$ and the solution is by now a classical.

A

Convex Optimization Theory

In the last two decades, several fundamental and practical results have been obtained in convex optimization theory [12, 13, 20]. The engineering community not only has benefited from these recent advances by finding applications, but has also fueled the mathematical development of both the theory and efficient algorithms. The two classical mathematical references on the subject are [124] and [93]. More recent engineering-oriented excellent references are [12], [13], and [20].

Traditionally, it was a common belief that linear problems were easy to solve as opposed to nonlinear problems. However, as stated by Rockafellar in a 1993 survey [125]: “the great watershed in optimization is not between linearity and nonlinearity, but convexity and nonconvexity” [20]. In a nutshell, convex problems can be solved optimally either in closed form (by means of the optimality conditions derived from Lagrange duality) or numerically (with very efficient algorithms that exhibit a polynomial convergence). As a consequence, roughly speaking, one can say that once a problem has been expressed in convex form, it has been solved.

Unfortunately, most engineering problems are not convex when directly formulated. However, many of them have a potential hidden convexity that engineers have to unveil in order to be able to use all the machinery of convex optimization theory.

This appendix introduces the basic ideas of convex optimization.

A.1 Convex Problems

An optimization problem with arbitrary equality and inequality constraints can always be written in the following standard form [20]:

$$\begin{array}{lll} \underset{\mathbf{x}}{\text{minimize}} & f_0(\mathbf{x}) & \\ \text{subject to} & f_i(\mathbf{x}) \leq 0 & 1 \leq i \leq m, \\ & h_i(\mathbf{x}) = 0 & 1 \leq i \leq p, \end{array} \tag{A.1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the optimization variable, f_0 is the *cost* or *objective function*, $f_1, \ldots, f_m$ are the m inequality constraint functions, and $h_1, \ldots, h_p$ are the p equality constraint functions. If there are no constraints, we say that the problem is *unconstrained.*

If the objective and inequality constraint functions are convex[1] and the equality constraint functions are linear (or, more generally, affine), the problem is then a *convex optimization problem* (or *convex program*).

The set of points for which the objective and all constraint functions are defined, i.e.,

$$D = \bigcap_{i=0}^{m} \operatorname{dom} f_i \cap \bigcap_{i=1}^{p} \operatorname{dom} h_i$$

is called the *domain* of the optimization problem (A.1). A point $\mathbf{x} \in D$ is *feasible* if it satisfies all the constraints $f_i(\mathbf{x}) \leq 0$ and $h_i(\mathbf{x}) = 0$. The problem (A.1) is said to be feasible if there exists at least one feasible point and *infeasible* otherwise. The *optimal value* (minimal value) is denoted by $f^\star$ and is achieved at an optimal solution $\mathbf{x}^\star$, i.e., $f^\star = f_0(\mathbf{x}^\star)$ (if the problem is infeasible, it is commonly denoted by $f^\star = +\infty$).

Many analysis and design problems arising in engineering can be cast (or recast) in the form of a convex optimization problem. In general, some manipulations are required to convert the problem into a

[1] A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if, for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ and $\theta \in [0,1]$, $\theta\mathbf{x} + (1-\theta)\mathbf{y} \in \operatorname{dom} f$ (i.e., the domain is a convex set) and $f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$.

convex one (unfortunately, this is not always possible). The interest of expressing a problem in convex form is that, although an analytical solution may not exist and the problem may seem difficult to solve (it may have hundreds of variables and a nonlinear, nondifferentiable objective function), it can still be solved (numerically) very efficiently both in theory and practice [20]. Another interesting feature of expressing the problem in convex form is that additional constraints can be straightforwardly added as long as they are convex.

Convex programming has been used in related areas such as FIR filter design [35, 168], antenna array pattern synthesis [89], power control for interference-limited wireless networks [81], and beamforming design in a multiuser scenario with a multi-antenna base station [10].

A.2 Classes of Convex Problems

When the functions f_i and h_i in (A.1) are linear (affine), the problem is called a *linear program* (LP) and is much simpler to solve. If the objective function is quadratic and the constraint functions are linear (affine), then it is called a *quadratic program* (QP); if, in addition, the inequality constraints are also quadratic, it is called *quadratically constrained quadratic program* (QCQP). QPs include LPs as special case.

A problem that is closely related to quadratic programming is the *second-order cone program* (SOCP) [20, 91] that includes constraints of the form:

$$\|\mathbf{A}\mathbf{x} + \mathbf{b}\| \leq \mathbf{c}^T\mathbf{x} + d, \tag{A.2}$$

where $\mathbf{A} \in \mathbb{R}^{k \times n}$, $\mathbf{b} \in \mathbb{R}^k$, $\mathbf{c} \in \mathbb{R}^n$, and $d \in \mathbb{R}$ are given and fixed. Note that (A.2) defines a convex set because it is an affine transformation of the second-order cone $\mathcal{C}^n = \{(\mathbf{x}, t) \in \mathbb{R}^n \mid \|\mathbf{x}\| \leq t\}$, which is convex since both $\|\mathbf{x}\|$ and $-t$ are convex. If $\mathbf{c} = \mathbf{0}$, then (A.2) reduces to a quadratic constraint (by squaring both sides).

A more general problem than an SOCP is a *semidefinite program* (SDP) [20, 155] that has matrix inequality constraints of the form:

$$x_1\mathbf{F}_1 + \cdots + x_n\mathbf{F}_n + \mathbf{G} \leq \mathbf{0}, \tag{A.3}$$

where $\mathbf{F}_1, \ldots, \mathbf{F}_n, \mathbf{G} \in \mathcal{S}^k$ ($\mathcal{S}^k$ is the set of Hermitian $k \times k$ matrices) and $\mathbf{A} \geq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semidefinite.

A very useful generalization of the standard convex optimization problem (A.1) is obtained by allowing the inequality constraints to be vector valued and using generalized inequalities [20]:

$$\begin{array}{lll} \underset{\mathbf{x}}{\text{minimize}} & f_0(\mathbf{x}) & \\ \text{subject to} & \mathbf{f}_i(\mathbf{x}) \preceq_{\mathcal{K}_i} \mathbf{0} & 1 \leq i \leq m, \\ & \mathbf{h}_i(\mathbf{x}) = \mathbf{0} & 1 \leq i \leq p, \end{array} \tag{A.4}$$

where the generalized inequalities[2] $\preceq_{\mathcal{K}_i}$ are defined by the proper cones $\mathcal{K}_i$ ($\mathbf{a} \preceq_{\mathcal{K}} \mathbf{b} \Leftrightarrow \mathbf{b} - \mathbf{a} \in \mathcal{K}$) [20] and f_i are $\mathcal{K}_i$-convex.[3]

Among the simplest convex optimization problems with generalized inequalities are *cone programs* (CP) (or *conic form problems*), which have a linear objective and one inequality constraint function [7, 20]:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \mathbf{c}^T\mathbf{x} \\ \text{subject to} & \mathbf{F}\mathbf{x} + \mathbf{g} \preceq_{\mathcal{K}} \mathbf{0} \\ & \mathbf{A}\mathbf{x} = \mathbf{b}. \end{array} \tag{A.5}$$

CPs particularize nicely to LPs, SOCPs, and SDPs as follows: (i) if $\mathcal{K} = \mathbb{R}^n_+$ (nonnegative orthant), the partial ordering $\preceq_{\mathcal{K}}$ is the usual componentwise inequality between vectors and (A.5) reduces to an LP; (ii) if $\mathcal{K} = \mathcal{C}^n$ (second-order cone), $\preceq_{\mathcal{K}}$ corresponds to a constraint of the form (A.2) and the problem (A.5) becomes an SOCP; (iii) if $\mathcal{K} = \mathcal{S}^n_+$ (positive semidefinite cone), the generalized inequality $\preceq_{\mathcal{K}}$ reduces to the usual matrix inequality as in (A.3) and the problem (A.5) simplifies to an SDP.

There is yet another very interesting and useful class of problems: *geometric programs* (GP), which are not convex in their natural form but can be transformed into convex problems [19, 20, 23].

[2] A generalized inequality is a partial ordering on $\mathbb{R}^n$ that has many of the properties of the standard ordering on $\mathbb{R}$. A common example is the matrix inequality defined by the cone of positive semidefinite $n \times n$ matrices $\mathcal{S}^n_+$.

[3] A function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^{k_i}$ is $\mathcal{K}_i$-convex if the domain is a convex set and, for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ and $\theta \in [0,1]$, $\mathbf{f}(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \preceq_{\mathcal{K}_i} \theta\mathbf{f}(\mathbf{x}) + (1-\theta)\mathbf{f}(\mathbf{y})$.

A.3 Reformulating a Problem in Convex Form

Convex problems can be optimally solved in practice, either in closed form or numerically. However, the natural formulation of most engineering problems is not convex. In many cases, fortunately, there is a hidden convexity that can be unveiled by properly reformulating the problem. The main task of an engineer is in fact to cast the problem in convex form and, if possible, in any of the well-known classes of convex problems (so that specific and optimized algorithms can be used).

Unfortunately, there is not a systematic way to reformulate a problem in convex form. In fact, it is rather an art that can only be learned by examples. There are two main ways to reformulate a problem in convex form. The main one is to devise a convex problem equivalent to the original nonconvex one by using a series of smart changes of variables. As an example, consider the minimization of $1/\left(1+x^2\right)$ subject to $x^2 \geq 1$, which is a nonconvex problem (both the cost function and the constraint are nonconvex). The problem can be rewritten in convex form, after the change of variable $y = x^2$, as the minimization of $1/\left(1+y\right)$ subject to $y \geq 1$ (and the optimal x can be recovered from the optimal y as $x = \sqrt{y}$). The class of geometric problems is a very important example of nonconvex problems that can be reformulated in convex form by a change of variable [19, 20, 23].

Nevertheless, it is not really necessary to devise a convex problem that is exactly equivalent to the original one. In fact, it suffices if they both have the same set of optimal solutions (related by some mapping). In other words, both problems have to be equivalent only within the set of optimal solutions but not otherwise. Of course, the difficulty is how to obtain such a "magic" convex problem without knowing beforehand the set of optimal solutions. One common way to do this is by relaxing the problem (removing some of the constraints) such that it becomes convex, in a way that the "relaxed" optimal solutions can be shown to satisfy the removed constraints as well. A remarkable example of this approach is [10] for multiuser beamforming.

A.4 Lagrange Duality Theory and KKT Optimality Conditions

Lagrange duality theory is a very rich and mature theory that links the original minimization problem (A.1), termed primal problem, with a dual maximization problem. In some occasions, it is simpler to solve the dual problem than the primal one. A fundamental result in duality theory is given by the optimality Karush–Kuhn–Tucker (KKT) conditions that any primal-dual solution must satisfy. By exploring the KKT conditions, it is possible in many cases to obtain a closed-form solution to the original problem. In the following, the basic results on duality theory including the KKT conditions are stated (for details, the reader is referred to [12, 13, 20]).

The basic idea in Lagrange duality is to take the constraints of (A.1) into account by augmenting the objective function with a weighted sum of the constraint functions. The *Lagrangian* of (A.1) is defined as

$$L\left(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}\right) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x}), \tag{A.6}$$

where λ_i and ν_i are the *Lagrange multipliers* associated with the ith inequality constraint $f_i(\mathbf{x}) \leq 0$ and with the ith equality constraint $h_i(\mathbf{x}) = 0$, respectively.

The optimization variable $\mathbf{x}$ is called the *primal variable* and the Lagrange multipliers $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ are also termed the *dual variables.* The original objective function $f_0(\mathbf{x})$ is referred to as the *primal objective*, whereas the *dual objective* $g\left(\boldsymbol{\lambda}, \boldsymbol{\nu}\right)$ is defined as the minimum value of the Lagrangian over $\mathbf{x}$:

$$g\left(\boldsymbol{\lambda}, \boldsymbol{\nu}\right) = \inf_{\mathbf{x} \in D} L\left(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}\right), \tag{A.7}$$

which is concave even if the original problem is not convex because it is the pointwise infimum of a family of affine functions of $(\boldsymbol{\lambda}, \boldsymbol{\nu})$. Note that the infimum in (A.7) is with respect to all $\mathbf{x} \in D$ (not necessarily feasible points). The dual variables $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ are *dual feasible* if $\boldsymbol{\lambda} \geq \mathbf{0}$.

The dual function $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is a lower bound on the optimal value $f^\star$ of the problem (A.1). Indeed, for any feasible $(\mathbf{x}, (\boldsymbol{\lambda}, \boldsymbol{\nu}))$:

$$\begin{aligned} f_0(\mathbf{x}) &\geq f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x}) && \text{(A.8)} \\ &\geq \inf_{\mathbf{z} \in D} \left(f_0(\mathbf{z}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{z}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{z}) \right) && \text{(A.9)} \\ &= g(\boldsymbol{\lambda}, \boldsymbol{\nu}), && \text{(A.10)} \end{aligned}$$

where we have used the fact that $f_i(\mathbf{x}) \leq 0$ and $h_i(\mathbf{x}) = 0$ for any feasible $\mathbf{x}$ and $\lambda_i \geq 0$ for any feasible λ_i in the first inequality. Thus, for the set of feasible $(\mathbf{x}, (\boldsymbol{\lambda}, \boldsymbol{\nu}))$, it follows that

$$\min_{\mathbf{x}} f_0(\mathbf{x}) \geq \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}} g(\boldsymbol{\lambda}, \boldsymbol{\nu}), \tag{A.11}$$

which holds even if the original problem is not convex. The difference between the optimal primal objective $f^\star$ and the optimal dual objective $g^\star$ is called the *duality gap*, which is always nonnegative $f^\star - g^\star \geq 0$. This property is called weak duality. If (A.11) is satisfied with equality we say that strong duality holds.

A central result in convex analysis [12, 13, 20, 93, 124] is that when the problem is convex, under some technical conditions (called constraint qualifications), the duality gap reduces to zero at the optimal (strong duality holds), i.e., (A.11) is achieved with equality for some $(\mathbf{x}^\star, (\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star))$. Therefore, one way to solve the original problem is to solve instead its associated dual problem:

$$\begin{array}{ll} \underset{\boldsymbol{\lambda}, \boldsymbol{\nu}}{\text{maximize}} & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} & \boldsymbol{\lambda} \geq \mathbf{0}, \end{array} \tag{A.12}$$

which is always a convex optimization problem even if the original problem is not convex (the objective to be maximized $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is concave and the constraint is convex). It is interesting to note that a primal-dual feasible pair $(\mathbf{x}, (\boldsymbol{\lambda}, \boldsymbol{\nu}))$ localizes the optimal value of the primal (and dual) problem in an interval:

$$f^\star \in [g(\boldsymbol{\lambda}, \boldsymbol{\nu}), f_0(\mathbf{x})]. \tag{A.13}$$

This can be used in optimization algorithms to provide nonheuristic stopping criteria.

Let $\mathbf{x}^\star$ and $(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$ be the primal and dual variables at the optimum. If we substitute them in the chain of inequalities (A.8)–(A.10), we see that each of the inequalities must be satisfied with equality (assuming that strong duality holds). To have equality in (A.8), it must be that $\lambda_i f_i(\mathbf{x}) = 0$ (this is the so-called complementary slackness condition). Moreover, since the inequality in (A.9) must also be satisfied with equality, the infimum is achieved at $\mathbf{x}^\star$; in other words, the gradient of the Lagrangian with respect to $\mathbf{x}$ must be zero at $(\mathbf{x}^\star, (\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star))$, i.e., $\nabla_{\mathbf{x}} L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) = \mathbf{0}$. These two results, together with the constraints on the primal and dual variables, form the (KKT) conditions:

$$h_i(\mathbf{x}^\star) = 0, \quad f_i(\mathbf{x}^\star) \leq 0, \tag{A.14}$$

$$\lambda_i^\star \geq 0, \tag{A.15}$$

$$\nabla_{\mathbf{x}} f_0(\mathbf{x}^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla_{\mathbf{x}} f_i(\mathbf{x}^\star) + \sum_{i=1}^{p} \nu_i^\star \nabla_{\mathbf{x}} h_i(\mathbf{x}^\star) = \mathbf{0}, \tag{A.16}$$

$$\text{(complementary slackness)} \quad \lambda_i^\star f_i(\mathbf{x}^\star) = 0. \tag{A.17}$$

Under some technical conditions (called constraint qualifications), the KKT conditions are necessary and sufficient for optimality. One simple version of the constraint qualifications is Slater's condition, which is satisfied when there exists $\mathbf{x}$ such that $f_i(\mathbf{x}) < 0,\ 1 \leq i \leq m$ and $h_i(\mathbf{x}) = 0,\ 1 \leq i \leq p$ (such a point is sometimes called *strictly feasible* since the inequality constraints hold with strict inequality) [12, 20].

In practice, the KKT conditions are very useful to obtain optimal primal-dual solutions analytically.

A.5 Sensitivity Analysis

The optimal dual variables (Lagrange multipliers) of a convex optimization problem give useful information about the sensitivity of the optimal value with respect to perturbations of the constraints [12, 13, 20]. Consider the following perturbed version of the original convex problem (A.1):

$$\begin{array}{lll} \underset{\mathbf{x}}{\text{minimize}} & f_0(\mathbf{x}) & \\ \text{subject to} & f_i(\mathbf{x}) \le u_i & 1 \le i \le m, \\ & h_i(\mathbf{x}) = v_i & 1 \le i \le p. \end{array} \tag{A.18}$$

This problem coincides with the original problem (A.1) if $u_i = 0$ and $v_i = 0$. When u_i is positive it means that we have relaxed the ith inequality constraint; when u_i is negative, it means that we have tightened the constraint. We define $f^\star(\mathbf{u},\mathbf{v})$ as the optimal value of the perturbed problem with the perturbations $\mathbf{u}$ and $\mathbf{v}$. Note that $f^\star(\mathbf{0},\mathbf{0}) = f^\star$.

Suppose that $f^\star(\mathbf{u},\mathbf{v})$ is differentiable at $\mathbf{u} = \mathbf{0}$, $\mathbf{v} = \mathbf{0}$. Then, provided that strong duality holds, the optimal dual variables $\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star$ are related to the gradient of $f^\star(\mathbf{u},\mathbf{v})$ at $\mathbf{u} = \mathbf{0}$, $\mathbf{v} = \mathbf{0}$ by [20]

$$\lambda_i^\star = -\frac{\partial f^\star(\mathbf{0},\mathbf{0})}{\partial u_i}, \quad \nu_i^\star = -\frac{\partial f^\star(\mathbf{0},\mathbf{0})}{\partial v_i}. \tag{A.19}$$

This means that tightening the ith constraint a small amount (i.e., taking u_i small and negative) yields an increase in $f^\star$ of approximately $-\lambda_i^\star u_i$ and, similarly, loosening the ith constraint a small amount (i.e., taking u_i small and positive) yields a decrease in $f^\star$ of approximately $\lambda_i^\star u_i$.

The sensitivity result of (A.19) allows us to assign a numerical value to how *active* a constraint is at the optimum $\mathbf{x}^\star$. If $f_i(\mathbf{x}^\star) < 0$, then the constraint is inactive and it follows that the constraint can be tightened or loosened a small amount without affecting the optimal value (this agrees with the fact that $\lambda_i^\star$ must be zero by the complementary slackness condition). Suppose now that $f_i(\mathbf{x}^\star) = 0$, i.e., the ith constraint is active at the optimum. The ith optimal Lagrange multiplier tells us how active the constraint is: if $\lambda_i^\star$ is small, it means that the constraint can be loosened or tightened a small amount without much effect on the optimal value; if $\lambda_i^\star$ is large, it means that if the constraint is loosened or tightened a small amount, the effect on the optimal value will be great.

A.6 Efficient Numerical Algorithms to Solve Convex Problems

During the last two decades, there has been a tremendous advance in developing efficient algorithms for solving wide classes of convex optimization problems. The most recent breakthrough in convex optimization theory is probably the development of interior-point methods for nonlinear convex problems. This was well established by Nesterov and Nemirovski in 1994 [104], where they extended the theory of linear programming interior-point methods (Karmarkar, 1984) to nonlinear convex optimization problems (based on the convergence theory of Newton's method for self-concordant functions).

The traditional optimization methods are based on gradient descent algorithms, which suffer from slow convergence and sensitivity to the algorithm initialization and stepsize selection. The recently developed methods for convex problems enjoy excellent convergence properties (polynomial convergence) and do not suffer from the usual problems of the traditional methods. In addition, it is simple to employ nonheuristic stopping criteria based on a desired resolution, since the difference between the cost value at each iteration and the optimum value can be upper-bounded using duality theory as in (A.13) [12, 20].

Many different software implementations have been recently developed and many of them are publicly available for free. It is worth pointing out that the existing packages not only provide the optimal primal variables of the problem but also the optimal dual variables. Currently, one of the most popular software optimization packages is SeDuMi [145], which is a Matlab toolbox for solving optimization problems over symmetric cones. Another convenient software package is cvx which is a Matlab-based modeling system for convex optimization: cvx turns Matlab into a modeling language, allowing constraints and objectives to be specified using standard Matlab expression syntax [53].

A.7 Primal and Dual Decompositions

The basic idea of the so-called *decomposition techniques* is to decompose the original large problem into distributively solvable *subproblems*

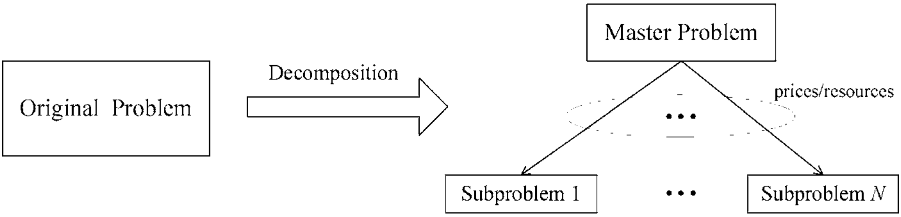


Fig. A.1 Decomposition of a problem into several subproblems controlled by a master problem through prices (dual decomposition) or direct resource allocation (primal decomposition).

which are then coordinated by a high level *master problem* by means of some form of signaling (see Figure A.1) [12, 14, 88]. Most of the existing decomposition techniques can be classified into *primal decomposition* and *dual decomposition* methods. The former is based on decomposing the original primal problem, whereas the latter is based on decomposing the Lagrangian dual problem [12, 142].

Primal decomposition methods correspond to a *direct resource allocation* where the master problem allocates the existing resources directly to each subproblem. Dual decomposition methods correspond to a *resource allocation via pricing* where the master problem sets the price for the resources to each subproblem. We next overview the basic primal and dual decomposition techniques (the interested reader is referred to [107] for more details on decompositions).

A.7.1 Primal Decomposition

A primal decomposition is appropriate when the problem has a coupling variable such that, when fixed to some value, the rest of the optimization problem decouples into several subproblems. Consider, for example, the following problem:

$$\begin{array}{lll} \underset{\mathbf{y},\{\mathbf{x}_i\}}{\text{maximize}} & \sum_i f_i(\mathbf{x}_i) & \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i & \forall i \\ & \mathbf{A}_i\mathbf{x}_i \leq \mathbf{y} & \\ & \mathbf{y} \in \mathcal{Y}. & \end{array} \tag{A.20}$$

Clearly, if variable $\mathbf{y}$ were fixed, then the problem would decouple. Therefore, it makes sense to separate the optimization in (A.20) into two levels of optimization. At the lower level, we have the subproblems, one for each i, in which (A.20) decouples when $\mathbf{y}$ is fixed:

$$\begin{array}{ll} \underset{\mathbf{x}_i}{\text{maximize}} & f_i(\mathbf{x}_i) \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i \\ & \mathbf{A}_i \mathbf{x}_i \leq \mathbf{y}. \end{array} \tag{A.21}$$

At the higher level, we have the master problem in charge of updating the coupling variable $\mathbf{y}$ by solving:

$$\begin{array}{ll} \underset{\mathbf{y}}{\text{maximize}} & \sum_i f_i^\star(\mathbf{y}) \\ \text{subject to} & \mathbf{y} \in \mathcal{Y}, \end{array} \tag{A.22}$$

where $f_i^\star(\mathbf{y})$ is the optimal objective value of problem (A.21) for a given $\mathbf{y}$. If the original problem (A.20) is a convex optimization problem, then the subproblems as well as the master problem are all convex programs.

If the function $\sum_i f_i^\star(\mathbf{y})$ is differentiable, then the master problem (A.22) can be solved with a gradient method. In general, however, the objective function $\sum_i f_i^\star(\mathbf{y})$ may be nondifferentiable, and the subgradient method is a convenient approach which only requires the knowledge a subgradient for each $f_i^\star(\mathbf{y})$, given by [12, Sec. 6.4.2][88, Ch. 9]

$$\mathbf{s}_i(\mathbf{y}) = \boldsymbol{\lambda}_i^\star(\mathbf{y}), \tag{A.23}$$

where $\boldsymbol{\lambda}_i^\star(\mathbf{y})$ is the optimal Lagrange multiplier corresponding to the constraint $\mathbf{A}_i\mathbf{x}_i \leq \mathbf{y}$ in problem (A.21). The global subgradient is then $\mathbf{s}(\mathbf{y}) = \sum_i \mathbf{s}_i(\mathbf{y}) = \sum_i \boldsymbol{\lambda}_i^\star(\mathbf{y})$. The subproblems in (A.21) can be locally and independently solved with the knowledge of $\mathbf{y}$.

A.7.2 Dual Decomposition

A dual decomposition is appropriate when the problem has a coupling constraint such that, when relaxed, the optimization problem decouples into several subproblems. Consider, for example, the following problem:

$$\begin{array}{lll} \underset{\{\mathbf{x}_i\}}{\text{maximize}} & \sum_i f_i\left(\mathbf{x}_i\right) & \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i & \forall i \\ & \sum_i \mathbf{h}_i\left(\mathbf{x}_i\right) \leq \mathbf{c}. & \end{array} \tag{A.24}$$

Clearly, if the constraint $\sum_i \mathbf{h}_i\left(\mathbf{x}_i\right) \leq \mathbf{c}$ were absent, then the problem would decouple. Therefore, it makes sense to form the Lagrangian by relaxing the coupling constraint in (A.24) as

$$\begin{array}{ll} \underset{\{\mathbf{x}_i\}}{\text{maximize}} & \sum_i f_i\left(\mathbf{x}_i\right) - \boldsymbol{\lambda}^T\left(\sum_i \mathbf{h}_i\left(\mathbf{x}_i\right) - \mathbf{c}\right) \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i \quad \forall i, \end{array} \tag{A.25}$$

such that the optimization separates into two levels of optimization. At the lower level, we have the subproblems (i.e., the Lagrangians), one for each i, in which (A.25) decouples:

$$\begin{array}{ll} \underset{\mathbf{x}_i}{\text{maximize}} & f_i\left(\mathbf{x}_i\right) - \boldsymbol{\lambda}^T \mathbf{h}_i\left(\mathbf{x}_i\right) \\ \text{subject to} & \mathbf{x}_i \in \mathcal{X}_i. \end{array} \tag{A.26}$$

At the higher level, we have the master dual problem in charge of updating the dual variable $\boldsymbol{\lambda}$ by solving the dual problem:

$$\begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{minimize}} & g\left(\boldsymbol{\lambda}\right) = \sum_i g_i\left(\boldsymbol{\lambda}\right) + \boldsymbol{\lambda}^T \mathbf{c} \\ \text{subject to} & \boldsymbol{\lambda} \geq \mathbf{0}, \end{array} \tag{A.27}$$

where $g_i\left(\boldsymbol{\lambda}\right)$ is the dual function obtained as the maximum value of the Lagrangian solved in (A.26) for a given $\boldsymbol{\lambda}$. This approach is in fact solving the dual problem instead of the original primal one. Hence, it will only give appropriate results if strong duality holds (e.g., when the original problem is convex and there exist strictly feasible solutions).

If the dual function $g\left(\boldsymbol{\lambda}\right)$ is differentiable, then the master dual problem in (A.27) can be solved with a gradient method. In general, however, it may not be differentiable, and the subgradient method becomes again a convenient approach which only requires the knowledge a subgradient for each $g_i\left(\boldsymbol{\lambda}\right)$, given by [12, Sec. 6.1]

$$\mathbf{s}_i\left(\boldsymbol{\lambda}\right) = -\mathbf{h}_i\left(\mathbf{x}_i^{\star}\left(\boldsymbol{\lambda}\right)\right), \tag{A.28}$$

where $\mathbf{x}_i^\star(\boldsymbol{\lambda})$ is the optimal solution of problem (A.26) for a given $\boldsymbol{\lambda}$. The global subgradient is then $\mathbf{s}(\boldsymbol{\lambda}) = \sum_i \mathbf{s}_i(\mathbf{y}) + \mathbf{c} = \mathbf{c} - \sum_i \mathbf{h}_i(\mathbf{x}_i^\star(\boldsymbol{\lambda}))$. The subproblems in (A.26) can be locally and independently solved with knowledge of $\boldsymbol{\lambda}$.

A.7.3 Gradient and Subgradient Methods

After performing a decomposition, the objective function of the resulting master problem may or may not be differentiable. For differentiable/nondifferentiable functions a gradient/subgradient method is very convenient because of its simplicity, little requirements of memory usage, and amenability for parallel implementation [12, 88, 142].

For a convex (concave) function f, a subgradient at point $\mathbf{x}_0$ is defined as any vector $\mathbf{s}$ that satisfies

$$f(\mathbf{x}) \geq (\leq) f(\mathbf{x}_0) + \mathbf{s}^T(\mathbf{x} - \mathbf{x}_0) \quad \forall \mathbf{x}. \tag{A.29}$$

Consider the following general concave maximization over a convex set:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{maximize}} & sf_0(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{X}. \end{array} \tag{A.30}$$

Both the gradient and subgradient projection methods generate a sequence of feasible points $\{\mathbf{x}(t)\}$ as

$$\mathbf{x}(t+1) = [\mathbf{x}(t) + \alpha(t)\mathbf{s}(t)]_{\mathcal{X}}, \tag{A.31}$$

where $\mathbf{s}(t)$ is a gradient of f_0 evaluated at the point $\mathbf{x}(t)$ if f_0 is differentiable and a subgradient otherwise, $[\cdot]_{\mathcal{X}}$ denotes the projection onto the feasible set $\mathcal{X}$, and $\alpha(t)$ is a positive stepsize. It is interesting to point out that each iteration of the subgradient method may not improve the objective value as happens with a gradient method. What makes the subgradient method work is that, for sufficiently small stepsize, the distance of the current solution $\mathbf{x}(t)$ to the optimal solution $\mathbf{x}^\star$ decreases.

There are many results on convergence of the gradient/subgradient method with different choices of stepsizes [12, 13, 142]. For example, for a diminishing stepsize rule $\alpha(t) = \frac{1+m}{t+m}$, where m is a fixed nonnegative number, the algorithm is guaranteed to converge to the optimal value (assuming bounded gradients/subgradients) [13]. For a constant stepsize $\alpha(t) = \alpha$, more convenient for distributed algorithms, the gradient algorithm converges to the optimal value provided that the stepsize is sufficiently small (assuming that the gradient is Lipschitz) [12], whereas for the subgradient algorithm the best value converges to within some range of the optimal value (assuming bounded subgradients) [13].

B

Matrix Results

B.1 Generalized Triangular Decompositions

The nonlinear MIMO transceiver designs in Chapter 4 are based on the generalized triangular decomposition (GTD) of the form $\mathbf{H} = \mathbf{QRP}^\dagger$, where $\mathbf{H} \in \mathbb{C}^{m \times n}$ is a rank K matrix, $\mathbf{R}$ is a K by K upper triangular matrix, and both $\mathbf{Q} \in \mathbb{C}^{m \times K}$ and $\mathbf{P} \in \mathbb{C}^{n \times K}$ are orthonormal matrices. This form of matrix decomposition is quite generic. Special instances of this decomposition are, in chronological order:

(a) the singular value decomposition (SVD) [6, 77]

$$\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\dagger,$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix containing the singular values on the diagonal,

(b) the Schur decomposition [137]

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

where $\mathbf{R}$ contains the eigenvalues of $\mathbf{H}$ on the diagonal (here $\mathbf{P} = \mathbf{Q}$),

(c) The QR factorization [48, 67]

$$\mathbf{H} = \mathbf{Q}\mathbf{R},$$

(here $\mathbf{P} = \mathbf{I}$),

(d) The complete orthogonal decomposition [57]

$$\mathbf{H} = \mathbf{Q}_2\mathbf{R}_2\mathbf{Q}_1^\dagger,$$

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 $\mathbf{R}_1^\dagger$,

(e) the geometric mean decomposition (GMD) [69, 84, 175]

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

where $\mathbf{R}$ has equal diagonal elements equal to the geometric mean of the positive singular values.

B.1.1 Existence of GTD

On the generic form $\mathbf{H} = \mathbf{Q}\mathbf{R}\mathbf{P}^\dagger$, we pose the following question. Given $\mathbf{H}$, what kind of upper triangular matrix $\mathbf{R}$ (identified by its diagonal) is feasible by changing the orthonormal matrices $\mathbf{P}$ and $\mathbf{Q}$? The following theorem answers this question.

Theorem B.1. (GTD Theorem). Let $\mathbf{H} \in \mathbb{C}^{m\times n}$ have rank K with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K > 0$. There exists an upper triangular matrix $\mathbf{R} \in \mathbb{C}^{K\times K}$ and matrices $\mathbf{Q}$ and $\mathbf{P}$ with orthonormal columns such that $\mathbf{H} = \mathbf{Q}\mathbf{R}\mathbf{P}^\dagger$ if and only if the diagonal elements of $\mathbf{R}$ satisfy $|\mathbf{r}| \prec_\times \boldsymbol{\sigma}$,[1] where $|\mathbf{r}|$ is a vector with the absolute values of $\mathbf{r}$ element-wise.

The proof for the GTD Theorem is quite simple based on the following two theorems.

[1] See Definition 2.6 for the definition of multiplicative majorization.

The first result is due to Weyl [166] (see also [65, p. 171]):

Theorem B.2. If $\mathbf{A} \in \mathbb{C}^{K \times K}$ with eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_K|$ and with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K$, then

$$|\boldsymbol{\lambda}| \prec_\times \boldsymbol{\sigma}.$$

The second result is due to Horn [64] (see also [65, p. 220]):

Theorem B.3. If $\mathbf{r} \in \mathbb{C}^K$ and $\boldsymbol{\sigma} \in \mathbb{R}^K$ satisfy

$$|\boldsymbol{r}| \prec_\times \boldsymbol{\sigma}, \tag{B.1}$$

then there exists an upper triangular matrix $\mathbf{R} \in \mathbb{C}^{K \times K}$ with singular values σ_i, $1 \leq i \leq K$, and with $\mathbf{r}$ on the diagonal of $\mathbf{R}$.

We now combine Theorems B.2 and B.3 to prove the GTD Theorem.

Proof. [Proof of Theorem B.1] If $\mathbf{H} = \mathbf{QRP}^\dagger$, then the eigenvalues of $\mathbf{R}$ are its diagonal elements and the singular values of $\mathbf{R}$ coincide with those of $\mathbf{H}$. By Theorem B.2, $|\mathbf{r}| \prec_\times \boldsymbol{\sigma}$. Conversely, suppose that $|\mathbf{r}| \prec_\times \boldsymbol{\sigma}$ holds. Let $\mathbf{H} = \mathbf{U\Sigma V}^\dagger$ be the singular value decomposition (SVD), where $\boldsymbol{\Sigma} \in \mathbb{R}^{K \times K}$. By Theorem B.3, there exists an upper triangular matrix $\mathbf{R} \in \mathbb{C}^{K \times K}$ with the r_i on the diagonal and with singular values σ_i, $1 \leq i \leq K$. Let $\mathbf{R} = \mathbf{U}_0\boldsymbol{\Sigma}\mathbf{V}_0^\dagger$ be the SVD of $\mathbf{R}$. Substituting $\boldsymbol{\Sigma} = \mathbf{U}_0^\dagger\mathbf{R}\mathbf{U}_0$ in the SVD for $\mathbf{H}$, we have

$$\mathbf{H} = (\mathbf{U}\mathbf{U}_0^\dagger)\mathbf{R}(\mathbf{V}\mathbf{V}_0^\dagger)^\dagger.$$

In other words, $\mathbf{H} = \mathbf{QRP}^\dagger$ where $\mathbf{Q} = \mathbf{U}\mathbf{U}_0^\dagger$ and $\mathbf{P} = \mathbf{V}\mathbf{V}_0^\dagger$. □

B.1.2 The GTD Algorithm

In the following, we present an efficient algorithm for computing the GTD $\mathbf{H} = \mathbf{QRP}^\dagger$, where the diagonal $\mathbf{r}$ of $\mathbf{R}$ satisfies (B.1).

Let $\mathbf{U\Sigma V}^\dagger$ be the SVD of $\mathbf{H}$, where $\boldsymbol{\Sigma}$ is a K by K diagonal matrix with the diagonal containing the positive singular values. We let $\mathbf{R}^{(L)} \in \mathbb{C}^{K \times K}$ denote an upper triangular matrix with the following properties:

(a) $r_{ij}^{(L)} = 0$ when $i > j$ or $j > i \geq L$. In other words, the trailing principal submatrix of $\mathbf{R}^{(L)}$, starting at row L and column L, is diagonal.

(b) If $\mathbf{r}^{(L)}$ denotes the diagonal of $\mathbf{R}^{(L)}$, then the first $L-1$ elements of $\mathbf{r}$ and $\mathbf{r}^{(L)}$ are equal. In other words, the leading diagonal elements of $\mathbf{R}^{(L)}$ match the prescribed leading elements of the vector $\mathbf{r}$.

(c) $|\mathbf{r}|_{L:K} \prec_{\times} |\mathbf{r}|_{L:K}^{(L)}$, where $\mathbf{r}_{L:K}$ denotes the subvector of $\mathbf{r}$ consisting of components L through K. In other words, the trailing diagonal elements of $\mathbf{R}^{(L)}$ multiplicatively majorize the trailing elements of the prescribed vector $\mathbf{r}$.

Initially, we set $\mathbf{R}^{(1)} = \mathbf{\Sigma}$. Clearly, (a)–(c) hold for $L = 1$. Proceeding by induction, suppose we have generated upper triangular matrices $\mathbf{R}^{(L)}$, $L = 1, 2, \ldots, k$, satisfying (a)–(c), and unitary matrices $\mathbf{Q}_L$ and $\mathbf{P}_L$, such that $\mathbf{R}^{(L+1)} = \mathbf{Q}_L^{\dagger}\mathbf{R}^{(L)}\mathbf{P}_L$ for $1 \leq L < k$. We now show how to construct unitary matrices $\mathbf{Q}_k$ and $\mathbf{P}_k$ such that $\mathbf{R}^{(k+1)} = \mathbf{Q}_k^{\dagger}\mathbf{R}^{(k)}\mathbf{P}_k$, where $\mathbf{R}^{(k+1)}$ satisfies (a)–(c) for $L = k+1$.

Let p and q be defined as follows:

$$p = \arg \min_i \{|r_i^{(k)}| : k \leq i \leq K,\ |r_i^{(k)}| \geq |r_k|\}, \tag{B.2}$$

$$q = \arg \max_i \{|r_i^{(k)}| : k \leq i \leq K,\ |r_i^{(k)}| \leq |r_k|,\ i \neq p\}, \tag{B.3}$$

where $r_i^{(k)}$ is the ith element of $\mathbf{r}^{(k)}$. Since $|\mathbf{r}|_{k:K} \prec_{\times} |\mathbf{r}|_{k:K}^{(k)}$, there exists p and q satisfying (B.2) and (B.3). Let $\mathbf{\Pi}$ be the matrix corresponding to the symmetric permutation $\mathbf{\Pi}^{\dagger}\mathbf{R}^{(k)}\mathbf{\Pi}$ which moves the diagonal elements $r_{pp}^{(k)}$ and $r_{qq}^{(k)}$ to the kth and $(k+1)$-st diagonal positions, respectively. Let $\delta_1 = r_{pp}^{(k)}$ and $\delta_2 = r_{qq}^{(k)}$ denote the new diagonal elements at locations k and $k+1$ associated with the permuted matrix $\mathbf{\Pi}^{\dagger}\mathbf{R}^{(k)}\mathbf{\Pi}$.

Next, we construct unitary matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ by modifying the elements in the identity matrix that lie at the intersection of rows k and $k+1$ and columns k and $k+1$. We multiply the permuted matrix $\mathbf{\Pi}^{\dagger}\mathbf{R}^{(k)}\mathbf{\Pi}$ on the left by $\mathbf{G}_2^{\dagger}$ and on the right by $\mathbf{G}_1$. These multiplications will change the elements in the 2 by 2 submatrix at the intersection of rows k and $k+1$ with columns k and $k+1$. Our choice for the elements of $\mathbf{G}_1$ and $\mathbf{G}_2$ is shown below, where we focus on the relevant 2 by 2 submatrices of $\mathbf{G}_2^\dagger$, $\mathbf{\Pi}^\dagger\mathbf{R}^{(k)}\mathbf{\Pi}$, and $\mathbf{G}_1$:

$$\underset{(\mathbf{G}_2^\dagger)}{\frac{r_k}{|r_k|^2}\begin{bmatrix} c\delta_1^\dagger & s\delta_2^\dagger \\ -s\delta_2 & c\delta_1 \end{bmatrix}} \quad \underset{(\mathbf{\Pi}^\dagger\mathbf{R}^{(k)}\mathbf{\Pi})}{\begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix}} \quad \underset{(\mathbf{G}_1)}{\begin{bmatrix} c & -s \\ s & c \end{bmatrix}} = \underset{(\mathbf{R}^{(k+1)})}{\begin{bmatrix} r_k & x \\ 0 & y \end{bmatrix}} \tag{B.4}$$

If $|\delta_1| = |\delta_2| = |r_k|$, we take $c = 1$ and $s = 0$; if $|\delta_1| \neq |\delta_2|$, we take

$$c = \sqrt{\frac{|r_k|^2 - |\delta_2|^2}{|\delta_1|^2 - |\delta_2|^2}} \quad \text{and} \quad s = \sqrt{1 - c^2}. \tag{B.5}$$

In either case,

$$x = \frac{sc(|\delta_2|^2 - |\delta_1|^2)r_k}{|r_k|^2} \quad \text{and} \quad y = \frac{\delta_1\delta_2 r_k}{|r_k|^2}. \tag{B.6}$$

Figure B.1 depicts the transformation from $\mathbf{\Pi}^\dagger\mathbf{R}^{(k)}\mathbf{\Pi}$ to $\mathbf{G}_2^\dagger\mathbf{\Pi}^\dagger\mathbf{R}^{(k)}\mathbf{\Pi}\mathbf{G}_1$. The dashed box is the 2 by 2 submatrix displayed in (B.4). Notice that c and s, defined in (B.5), are real scalars chosen so that

$$c^2 + s^2 = 1 \quad \text{and} \quad c^2|\delta_1|^2 + s^2|\delta_2|^2 = |r_k|^2. \tag{B.7}$$

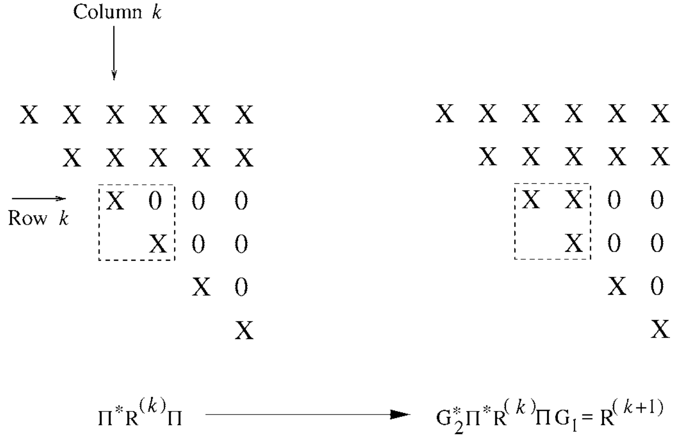


Fig. B.1 The operation displayed in (B.4).

With these identities, the validity of (B.4) follows by direct computation. By the choice of p and q, we have

$$|\delta_2| \leq |r_k| \leq |\delta_1|. \tag{B.8}$$

If $|\delta_1| \neq |\delta_2|$, it follows from (B.8) that c and s are real nonnegative scalars. It can be checked that the 2 by 2 matrices in (B.4) associated with $\mathbf{G}_1$ and $\mathbf{G}_2^\dagger$ are both unitary. Consequently, both $\mathbf{G}_1$ and $\mathbf{G}_2$ are unitary. We define

$$\mathbf{R}^{(k+1)} = (\mathbf{\Pi G}_2)^\dagger \mathbf{R}^{(k)} (\mathbf{\Pi G}_1) = \mathbf{Q}_k^\dagger \mathbf{R}^{(k)} \mathbf{P}_k,$$

where $\mathbf{Q}_k = \mathbf{\Pi G}_2$ and $\mathbf{P}_k = \mathbf{\Pi G}_1$. By (B.4) and Figure B.1, $\mathbf{R}^{(k+1)}$ has properties (a) and (b) for $L = k+1$. Readers are referred to [71] for the verification of property (c).

Hence, after $K-1$ steps we obtain an upper triangular matrix $\mathbf{R}^{(K)}$, with $\mathbf{r}_{1:K-1}$ occupying the first $K-1$ diagonal elements, and unitary matrices $\mathbf{Q}_i$ and $\mathbf{P}_i$, $i = 1, 2, \ldots, K-1$, such that

$$\mathbf{R}^{(K)} = (\mathbf{Q}_{k-1}^\dagger \ldots \mathbf{Q}_2^\dagger \mathbf{Q}_1^\dagger) \mathbf{\Sigma} (\mathbf{P}_1 \mathbf{P}_2 \ldots \mathbf{P}_{k-1}). \tag{B.9}$$

Equating determinants in (B.9) and utilizing the identity $r_i^{(k)} = r_i$ for $1 \leq i \leq K-1$, we have

$$\prod_{i=1}^{K} |r_i^{(K)}| = \frac{|r_K^{(K)}|}{|r_K|} \left(\prod_{i=1}^{K} |r_i| \right) = \prod_{i=1}^{K} \sigma_i = \prod_{i=1}^{K} |r_i|,$$

where the last equality is due to the assumption $|\mathbf{r}| \prec_\times \boldsymbol{\sigma}$. It follows that $|r_K^{(K)}| = |r_K|$. Let $\mathbf{C}$ be the diagonal matrix obtained by replacing the (K,K) element of the identity matrix by $r_K^{(K)}/r_K$. The matrix $\mathbf{C}$ is unitary since $|r_k|/|r_K^{(K)}| = 1$. The matrix

$$\mathbf{R} = \mathbf{C}^\dagger \mathbf{R}^{(K)} \tag{B.10}$$

has diagonal equal to $\mathbf{r}$ due to the choice of $\mathbf{C}$.

Combining (B.9) and (B.10) with the SVD $\mathbf{H} = \mathbf{U \Sigma V}^\dagger$ gives

$$\mathbf{H} = \mathbf{U Q}_1 \mathbf{Q}_2 \cdots \mathbf{Q}_{k-1} \mathbf{CRP}_{k-1}^\dagger \cdots \mathbf{P}_2^\dagger \mathbf{P}_1^\dagger \mathbf{V}^\dagger.$$

Hence, we have obtained the GTD with

$$\mathbf{Q} = \mathbf{U} \left(\prod_{i=1}^{K-1} \mathbf{Q}_i \right) \mathbf{C} \quad \text{and} \quad \mathbf{P} = \mathbf{V} \left(\prod_{i=1}^{K-1} \mathbf{P}_i \right).$$

Finally, note that if $\mathbf{r}$ is real, then $\mathbf{G}_1$ and $\mathbf{G}_2$ are real, which implies $\mathbf{R}$ is real.

We summarize the steps of the GTD algorithm as follows:

1. Let $\mathbf{H} = \mathbf{U\Sigma V}^\dagger$ be the SVD of $\mathbf{H}$, and suppose we are given $\mathbf{r} \in \mathbb{C}^K$ with $|\mathbf{r}| \prec_\times \boldsymbol{\sigma}$. Initialize $\mathbf{Q} = \mathbf{U}$, $\mathbf{P} = \mathbf{V}$, $\mathbf{R} = \mathbf{\Sigma}$, and $k = 1$.
2. Let p and q be defined as follows:
$$p = \arg\ \min_i \{|[\mathbf{R}]_{ii}| : k \le i \le K,\ |[\mathbf{R}]_{ii}| \ge |r_k|\},$$
$$q = \arg\ \max_i \{|[\mathbf{R}]_{ii}| : k \le i \le K,\ |[\mathbf{R}]_{ii}| \le |r_k|,\ i \neq p\}.$$
In $\mathbf{R}$, $\mathbf{P}$, and $\mathbf{Q}$, perform the following exchanges:
$$\begin{aligned} ([\mathbf{R}]_{kk}, [\mathbf{R}]_{k+1,k+1}) &\leftrightarrow ([\mathbf{R}]_{pp}, [\mathbf{R}]_{qq}) \\ (\mathbf{R}_{1:k-1,k}, \mathbf{R}_{1:k-1,k+1}) &\leftrightarrow (\mathbf{R}_{1:k-1,p}, \mathbf{R}_{1:k-1,q}) \\ (\mathbf{P}_{:,k}, \mathbf{P}_{:,k+1}) &\leftrightarrow (\mathbf{P}_{:,p}, \mathbf{P}_{:,q}) \\ (\mathbf{Q}_{:,k}, \mathbf{Q}_{:,k+1}) &\leftrightarrow (\mathbf{Q}_{:,p}, \mathbf{Q}_{:,q}) \end{aligned}$$
3. Construct the matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ shown in (B.4). Replace $\mathbf{R}$ by $\mathbf{G}_2^\dagger \mathbf{R} \mathbf{G}_1$, replace $\mathbf{Q}$ by $\mathbf{Q}\mathbf{G}_2$, and replace $\mathbf{P}$ by $\mathbf{P}\mathbf{G}_1$.
4. If $k = K - 1$, then go to step 5. Otherwise, replace k by $k + 1$ and go to step 2.
5. Multiply column K of $\mathbf{Q}$ by $[\mathbf{R}]_{KK}/r_K$; replace $[\mathbf{R}]_{KK}$ by r_K. The product $\mathbf{QRP}^\dagger$ is the GTD of $\mathbf{H}$ based on $\mathbf{r}$.

A MATLAB implementation of the GTD algorithm is posted on the web site of William Hager (http://www.math.ufl.edu/~hager/). Given the SVD, this algorithm for the GTD requires $O((m+n)K)$ flops. For comparison, reduction of $\mathbf{H}$ to bidiagonal form by the Golub–Kahan bidiagonalization scheme [51] (also see [52, 56, 167]), often the first step in the computation of the SVD, requires $O(mnK)$ flops.

B.1.3 Application to Inverse Eigenvalue Problem

Besides its applications to MIMO transceiver designs, GTD can also be applied to solve inverse eigenvalue problems, surveyed extensively in [25]. In [26], Chu presents a recursive procedure for constructing matrices with prescribed eigenvalues and singular values. His algorithm, which he calls SVD_EIG, is based on Horn's divide and conquer proof of the sufficiency of Weyl's product inequalities. In general, the output of SVD_EIG is not upper triangular. Consequently, this routine could not be used to generate the GTD. Chu notes that to achieve an upper triangular matrix would require an algorithm "one order more expensive than the divide-and-conquer algorithm."

Given a vector of singular values $\boldsymbol{\sigma} \in \mathbb{R}^n$ and a vector of eigenvalues $\boldsymbol{\lambda} \in \mathbb{C}^n$, with $\boldsymbol{\lambda} \preceq \boldsymbol{\sigma}$, we can use the GTD to generate a matrix $\mathbf{R}$ with $\boldsymbol{\lambda}$ on the diagonal and with singular values $\boldsymbol{\sigma}$. Both MATLAB routines and SVD_EIG [26] require $O(n^2)$ flops, so in an asymptotic sense, the approaches are equivalent. However, in the numerical experiments given in [71], the GTD algorithm is orders of magnitude faster than SVD_EIG and requires less amount of memory. In general, the GTD is numerically more accurate.

B.2 Miscellaneous Matrix Results

In this section, we include for convenience a few basic algebra results that are standard fare in most textbooks (e.g., [66, 83, 96, 133, 143]) and that will be repeatedly used throughout this text.

B.2.1 Basic Results on the Trace and Determinant

The following relation, commonly referred to as the circularity of the trace, is widely used:

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA}).$$

The following is a basic result on the determinant for conformable matrices [96, 143]:

$$|\mathbf{I} + \mathbf{AB}| = |\mathbf{I} + \mathbf{BA}|.$$

Another useful result is

$$\mathbf{A} \geq \mathbf{B} \Rightarrow |\mathbf{A}| \geq |\mathbf{B}|.$$

B.2.2 Singular Values of Product of Matrices

Theorem B.4. Let $\mathbf{A} \in \mathbb{C}^{M \times N}$ and $\mathbf{B} \in \mathbb{C}^{N \times K}$ be two rank K matrices with singular values $\sigma_{A,1} > \sigma_{A,2} > \cdots > \sigma_{A,K} > 0$ and $\sigma_{B,1} > \sigma_{B,2} > \cdots > \sigma_{B,K} > 0$. The product $\mathbf{AB}$ has nonzero singular values $\sigma_{AB,1} > \sigma_{AB,2} > \cdots > \sigma_{AB,K} > 0$. Then

$$\begin{aligned} &\textstyle\prod_{i=1}^{k} \sigma_{AB,i} \leq \prod_{i=1}^{k} \sigma_{A,i}\sigma_{B,i}, \quad 1 \leq k \leq K-1 \\ &\textstyle\prod_{i=1}^{K} \sigma_{AB,i} = \prod_{i=1}^{K} \sigma_{A,i}\sigma_{B,i}. \end{aligned} \tag{B.11}$$

Matrix Inversion Lemma

The general expression of the matrix inversion lemma is [66, 83, 133, 143]

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1}\right)^{-1}\mathbf{DA}^{-1}. \tag{B.12}$$

A particular case is the Woodbury's Identity:

$$\left(\mathbf{R} + \gamma^2\mathbf{cc}^\dagger\right)^{-1} = \mathbf{R}^{-1} - \frac{\gamma^2}{1 + \gamma^2\mathbf{c}^\dagger\mathbf{R}^{-1}\mathbf{c}}\mathbf{R}^{-1}\mathbf{cc}^\dagger\mathbf{R}^{-1}. \tag{B.13}$$

Cauchy–Schwarz's inequality

The Cauchy–Schwarz's inequality in vector form is [96, 66]

$$|\mathbf{y}^\dagger\mathbf{x}| \leq \|\mathbf{y}\|_2 \|\mathbf{x}\|_2 \tag{B.14}$$

with equality if and only if $\mathbf{y} = \alpha\mathbf{x}$, i.e., if $\mathbf{y}$ and $\mathbf{x}$ are linearly dependent.

Hadamard's inequality

Given an $n \times n$ positive semidefinite matrix $\mathbf{R}$, the following holds [33, 66, 96]:

$$|\mathbf{R}| \leq \prod_{i=1}^{n} [\mathbf{R}]_{ii}$$

with equality if and only if matrix $\mathbf{R}$ is diagonal (except in the trivial case in which $\mathbf{R}$ is singular).

Jensen's inequality [20, 33]

If f is a convex function, $x_1,\ldots,x_k \in \operatorname{dom} f$, and $\theta_1,\ldots,\theta_k \geq 0$ with $\theta_1 + \cdots + \theta_k = 1$, then

$$f\left(\theta_1 x_1 + \cdots + \theta_k x_k\right) \leq \theta_1 f\left(x_1\right) + \cdots + \theta_k f\left(x_k\right).$$

Moreover, if f is strictly convex, then equality implies that the x_i's for which $\theta_i > 0$ are equal.

The inequality extends to infinite sums, integrals, and expected values. For example, if x is a random variable such that $x \in \operatorname{dom} f$ with probability one and f is convex, then

$$f\left(\mathbb{E}\left[x\right]\right) \leq \mathbb{E}\left[f\left(x\right)\right],$$

provided the expectations exist. Moreover, if f is strictly convex, then equality implies that x is a constant.

Schur complement [20]

Consider the following partitioned matrix:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^{\dagger} & \mathbf{C} \end{bmatrix}. \tag{B.15}$$

If $\mathbf{A}$ is invertible, the Schur complement is defined as

$$\mathbf{S} = \mathbf{C} - \mathbf{B}^{\dagger}\mathbf{A}^{-1}\mathbf{B}. \tag{B.16}$$

The following characterizations of positive definiteness or semidefiniteness of the block matrix $\mathbf{M}$ hold:

- $\mathbf{M} > \mathbf{0}$ if and only if $\mathbf{A} > \mathbf{0}$ and $\mathbf{S} > \mathbf{0}$.
- if $\mathbf{A} > \mathbf{0}$, then $\mathbf{M} \geq \mathbf{0}$ if and only if $\mathbf{S} \geq \mathbf{0}$.

Observe that $\mathbf{M} \geq \mathbf{0}$ is equivalent (via a permutation) to

$$\begin{bmatrix} \mathbf{C} & \mathbf{B}^{\dagger} \\ \mathbf{B} & \mathbf{A} \end{bmatrix} \geq 0. \tag{B.17}$$

Acknowledgments

Daniel P. Palomar would like to thank his Ph.D. advisor, Miguel Angel Lagunas, and his mentor at Stanford University, John Cioffi, for their inspiring support at a time when he was starting an unexpectedly amusing journey on MIMO transceiver design and majorization theory. He would also like to acknowledge Stephen Boyd for introducing him into the wonderful world of convex optimization theory. In addition, he would like to thank his subsequent collaborators on the same topic: Javier R. Fonollosa, Mats Bengtsson, Björn Ottersten, and Sergio Barbarossa.

Yi Jiang would like to thank his Ph.D. advisor, Jian Li, for her invaluable support and inspiration during his Ph.D. work on these fascinating research problems. He would also like to thank Professor William W. Hager for the inspiring discussions on relating the MIMO transceiver design to the matrix decomposition. He is also grateful to his Postdoc advisor, Mahesh K. Varanasi, for providing him the opportunity of investigating the information-theoretic aspects of the MIMO transceiver design.

References

[1] S. M. Alamouti, “A simple transmit diversity technique for wireless communications,” *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451–1458, October 1998.

[2] N. Amitay and J. Salz, “Linear equalization theory in digital data transmission over dually polarized fading radio channels,” *At&T Bell Labs Technical Journal*, vol. 63, no. 10, pp. 2215–2259, December 1984.

[3] J. B. Andersen, “Array gain and capacity for known random channels with multiple element arrays at both ends,” *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 11, pp. 2172–2178, November 2000.

[4] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis.* Wiley, third ed., 2003.

[5] S. L. Ariyavisitakul, “Turbo space-time processing to improve wireless channel capacity,” *IEEE Transactions on Communications*, vol. 48, no. 8, pp. 1347–1358, August 2000.

[6] E. Beltrami, “Sulle funzioni bilineari,” *Giornale De Matematiche*, vol. 11, pp. 98–106, 1873.

[7] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications.* Society for Industrial and Applied Mathematics, 2001.

[8] S. Benedetto and E. Biglieri, *Principles of Digital Transmission: With Wireless Applications.* New York, NY, USA: Kluwer Academic Publishers, 1999.

[9] M. Bengtsson, “A pragmatic approach to multi-user spatial multiplexing,” in *Proceedings of 2nd IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM-2002)*, (Rosslyn, VA), August 4–6 2002.

[10] M. Bengtsson and B. Ottersten, “Optimal and suboptimal transmit beamforming,” in *Handbook of Antennas in Wireless Communications*, (L. C. Godara, ed.), Boca Raton, FL: CRC Press, 2001.

[11] T. Berger and D. W. Tufts, “Optimum pulse amplitude modulation. Part I: Transmitter-receiver design and bounds from information theory,” *IEEE Transactions on Information Theory*, vol. IT-13, no. 2, pp. 196–208, April 1967.

[12] D. P. Bertsekas, *Nonlinear Programming.* Belmont, MA, USA: Athena Scientific, second ed., 1999.

[13] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization.* Belmont, MA, USA: Athena Scientific, 2003.

[14] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods.* Englewood Cliffs, NJ: Prentice Hall, 1989.

[15] R. Bhatia, *Matrix Analysis.* New York: Springer-Verlag, 1997.

[16] E. Biglieri, *Coding for Wireless Channels.* New York, NY: Springer, 2005.

[17] J. A. C. Bingham, “Multicarrier modulation for data transmission: An idea whose time has come,” *IEEE Communication Magazine*, vol. 28, no. 5, pp. 5–14, May 1990.

[18] H. Boche and E. A. Jorswieck, “On the ergodic capacity as a function of the correlation properties in systems with multiple transmit antennas without CSI at the transmitter,” *IEEE Transactions on Communications*, vol. 52, no. 10, pp. 1654–1657, October 2004.

[19] S. Boyd, S. J. Kim, L. Vandenberghe, and A. Hassibi, *A Tutorial on Geometric Programming.* Optimization and Engineering, Springer, April 2007.

[20] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge, UK: Cambridge University Press, 2004.

[21] L. H. Brandenburg and A. D. Wyner, “Capacity of the Gaussian channel with memory: The multivariate case,” *The Bell System Technical Journal*, vol. 53, no. 5, pp. 745–778, May–June 1974.

[22] A. R. Calderbank, “The art of signaling: Fifty years of coding theory,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2561–2595, October 1998.

[23] M. Chiang, “Geometric programming for communication systems,” *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 1–2, pp. 1–154, August 2005.

[24] K. Cho and D. Yoon, “On the general BER expression of one and two-dimensional amplitude modulations,” *IEEE Transactions on Communications*, vol. 50, no. 7, pp. 1074–1080, July 2002.

[25] M. T. Chu, “Inverse eigenvalue problems,” *SIAM Review*, vol. 40, no. 1, pp. 1–39 (electronic), 1998.

[26] M. T. Chu, “A fast recursive algorithm for constructing matrices with prescribed eigenvalues and singular values,” *SIAM Journal of Numerical Analysis*, vol. 37, pp. 1004–1020, 2000.

[27] S. T. Chung and A. J. Goldsmith, “Degrees of freedom in adaptive modulation: A unified view,” *IEEE Transactions on Communications*, vol. 49, no. 9, pp. 1561–1571, September 2001.

[28] J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu, and G. D. Forney, Jr., "MMSE Decision-feedback equalizers and coding. Parts I-II: Equalization results and coding results," *IEEE Transactions on Communications*, vol. 43, no. 10, pp. 2582–2604, October 1995.

[29] J. M. Cioffi and G. D. Forney, Jr., "Generalized decision-feedback equalization for packet transmission with ISI and Gaussian noise," in *Communications, Computation, Control and Signal Processing*, (A. Paulraj, V. Roychowdhury, and C. D. Schaper, eds.), ch. 4, Boston, MA, USA: Kluwer Academic Publishers, 1997.

[30] A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1639–1667, June 2002.

[31] L. Collin, O. Berder, P. Rostaing, and G. Burel, "Optimal minimum distance-based precoder for MIMO spatial multiplexing systems," *IEEE Transactions on Signal Processing*, vol. 52, no. 3, pp. 617–627, March 2004.

[32] M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. 29, pp. 439–441, May 1983.

[33] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York, NY, USA: Wiley, 1991.

[34] M. O. Damen, H. El Gamal, and G. Caire, "On maximum-likelihood detection and the search for the closest lattice point," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2389–2402, October 2003.

[35] T. N. Davidson, Z.-Q. Luo, and J. F. Sturm, "Linear matrix inequality formulation of spectral mask constraints with applications to FIR filter design," *IEEE Transactions on Signal Processing*, vol. 50, no. 11, pp. 2702–2715, November 2002.

[36] Y. Ding, T. N. Davidson, Z.-Q. Luo, and K. M. Wong, "Minimum BER block precoders for zero-forcing equalization," *IEEE Transactions on Signal Processing*, vol. 51, no. 9, pp. 2410–2423, September 2003.

[37] H. El Gamal, G. Caire, and M. O. Damen, "Lattice coding and decoding achieve the optimal diversity-multiplexing tradeoff of MIMO channels," *IEEE Transactions on Information Theory*, vol. 50, no. 6, pp. 968–985, June 2004.

[38] P. Elia, K. R. Kumar, S. A. Pawar, P. V. Kumar, and H.-F. Lu, "Explicit construction of space-time block codes achieving the Diversity-multiplexing gain tradeoff," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3869–3884, September 2006.

[39] U. Erez, S. Shamai, and R. Zamir, "Capacity and lattice-strategies for cancelling known interference," *IEEE Transactions on Information Theory*, vol. 51, no. 11, pp. 3820–3833, November 2005.

[40] U. Erez and S. ten Brink, "A close-to-capacity dirty paper coding scheme," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3417–3432, October 2005.

[41] G. D. Forney, Jr., "On the role of MMSE estimation in approaching the information-theoretic limits of linear Gaussian channels: Shannon meets Wiener," *41st Allerton Conference on Communication, Control, and Computing, Monticello, IL*, p. 2003, October 2003.

[42] G. D. Forney, Jr., “Shannon meets Wiener II: On MMSE estimation in successive decoding schemes,” *Proceedings of 2004 Allerton Conference on Communication, Control, and Computing, Monticello, IL*, p. 2004, September 2004.

[43] G. Foschini and M. Gans, “On limits of wireless communications in a fading environment when using multiple antennas,” *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, 1998.

[44] G. J. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky, “Simplified processing for high spectral efficiency wireless communication employing multiple-element arrays,” *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, March 1999.

[45] G. J. Foschini, “Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas,” *Bell Labs Technical Journal*, vol. 1, no. 2, pp. 41–59, Autumn 1996.

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

[47] G. Ginis and J. Cioffi, “Vectored transmission for digital subscriber line systems,” *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 5, pp. 1085–1104, June 2002.

[48] W. Givens, “Computation of plane unitary rotations transforming a general matrix to triangular form,” *Journal of SIAM*, vol. 6, no. 1, pp. 26–50, March 1958.

[49] R. H. Gohary, T. N. Davidson, and Z.-Q. Luo, “An efficient design method for vector broadcast systems with common information,” in *Proceedings of IEEE 2003 Global Communications Conference (GLOBECOM-2003)*, (San Francisco, CA), December 1–5 2003.

[50] A. Goldsmith, *Wireless Communications.* New York: Cambridge University Press, 2005.

[51] G. H. Golub and W. Kahan, “Calculating the singular values and pseudo-inverse of a matrix,” *SIAM Journal on Numerical Analysis*, vol. 2, pp. 205–224, 1965.

[52] G. H. Golub and C. F. V. Loan, *Matrix Computations.* Baltimore, MD: The Johns Hopkins University Press, third ed., 1996.

[53] M. Grant, S. Boyd, and Y. Ye, “Disciplined convex programming,” in *Global Optimization: from Theory to Implementation, Nonconvex Optimization and Its Applications*, (L. Liberti and N. Maculan, eds.), pp. 155–210, New York: Springer Science+Business Media, Inc., 2006.

[54] R. M. Gray, “On the asymptotic eigenvalue distribution of Toeplitz matrices,” *IEEE Transactions on Information Theory*, vol. IT-18, no. 6, pp. 725–730, November 1972.

[55] T. Guess, “Optimal sequence for CDMA with decision-feedback receivers,” *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 886–900, April 2003.

[56] W. W. Hager, *Applied Numerical Linear Algebra.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

[57] R. J. Hanson and C. L. Lawson, “Extensions and applications of the householder algorithm for solving linear least squares problems,” *Mathematics of Computation*, vol. 23, pp. 787–812, March 1969.

[58] H. Harashima and H. Miyakawa, “Matched-transmission technique for channels with intersymbol interference,” *IEEE Transactions on Communications*, pp. 774–780, August 1972.

[59] G. H. Hardy, J. E. Littlewood, and G. Pólya, “Some simple inequalities satisfied by convex functions,” *Messenger Mathematics*, vol. 58, pp. 145–152, 1929.

[60] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities.* London and New York: Cambridge University Press, second ed., 1952.

[61] B. Hassibi, “A fast square-root implementation for BLAST,” *Thirty-Fourth Asilomar Conference on Signals, Systems and Computers*, pp. 1255–1259, November 2000.

[62] W. Hirt and J. L. Massey, “Capacity of the discrete-time Gaussian channel with intersymbol interference,” *IEEE Transactions on Information Theory*, vol. 34, no. 3, pp. 380–388, May 1988.

[63] M. L. Honig, K. Steiglitz, and B. Gopinath, “Multichannel signal processing for data communications in the presence of crosstalk,” *IEEE Transactions on Communications*, vol. 38, no. 4, pp. 551–558, April 1990.

[64] A. Horn, “On the eigenvalues of a matrix with prescribed singular values,” *Proceedings of American Mathematical Society*, vol. 5, pp. 4–7, 1954.

[65] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis.* Cambridge: Cambridge University Press, 1991.

[66] R. A. Horn and C. R. Johnson, *Matrix Analysis.* New York: Cambridge University Press, 1985.

[67] A. Householder, “Unitary triangularization of a nonsymmetric matrix,” *Journal of ACM 6*, vol. 6, pp. 339–342, 1958.

[68] A. T. James, “Normal multivariate analysis and the orthogonal group,” *The Annals of Mathematical Statistics*, vol. 25, no. 1, pp. 40–75, March 1954.

[69] Y. Jiang, W. Hager, and J. Li, “The geometric mean decomposition,” *Linear Algebra and Its Applications*, vol. 396, pp. 373–384, February 2005.

[70] Y. Jiang, W. Hager, and J. Li, “The generalized triangular decomposition,” *Mathematics of Computation*, November 2006.

[71] Y. Jiang, W. Hager, and J. Li, “Tunable channel decomposition for MIMO communications using channel state information,” *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4405–4418, November 2006.

[72] Y. Jiang, J. Li, and W. Hager, “Joint transceiver design for MIMO communications using geometric mean decomposition,” *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3791–3803, October 2005.

[73] Y. Jiang, J. Li, and W. Hager, “Uniform channel decomposition for MIMO communications,” *IEEE Transactions on Signal Processing*, vol. 53, no. 11, pp. 4283–4294, November 2005.

[74] Y. Jiang, D. P. Palomar, and M. K. Varanasi, “Precoder optimization for nonlinear MIMO transceiver based on arbitrary cost function,” in *Proceedings of 41st IEEE Annual Conference on Information Sciences and Systems (CISS-2007)*, The John Hopkins University, Baltimore, MD. pp. 14–16, March 2007.

[75] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Englewood Cliffs, NJ, USA: Prentice Hall, 1993.

[76] G. Jöngren, M. Skoglund, and B. Ottersen, "Combining beamforming and orthogonal space-time block coding," *IEEE Transactions on Information Theory*, vol. 48, no. 3, pp. 611–627, March 2002.

[77] C. Jordan, "Mémoire sur les formes bilinéaires," *Journal of Mathematiques pure et Appliquees*, vol. 19, pp. 35–54, 1874.

[78] E. Jorswieck and H. Boche, "Transmission strategies for the MIMO MAC with MMSE receiver: average MSE optimization and achievable individual MSE region," *IEEE Transactions on Signal Processing*, vol. 51, no. 11, pp. 2872–2881, November 2003.

[79] E. A. Jorswieck and H. Boche, "Optimal transmission strategies and impact of correlation in multiantenna systems with different types of channel state information," *IEEE Transactions on Signal Processing*, vol. 52, no. 12, pp. 3440–3453, December 2004.

[80] I. Kalet, "The multitone channel," *IEEE Transactions on Communications*, vol. 37, no. 2, pp. 119–124, February 1989.

[81] S. Kandukuri and S. Boyd, "Optimal power control in interference-limited fading wireless channels with outage-probability specifications," *IEEE Transactions on Communications*, vol. 1, no. 1, pp. 46–55, January 2002.

[82] M. Kavehrad and J. Salz, "Cross-polarization cancellation and equalization in digital transmission over dually polarized multipath fading channels," *At&T Technical Journal*, vol. 64, no. 10, pp. 2211–2245, December 1985.

[83] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[84] P. Kosowski and A. Smoktunowicz, "On constructing unit triangular matrices with prescribed singular values," *Computing*, vol. 64, pp. 279–285, 2000.

[85] H. Krim and M. Viberg, "Two decades of array signal processing research," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, July 1996.

[86] P. Lancaster, *Theory of Matrices*. New York: Academic Press, 1969.

[87] E. G. Larsson and P. Stoica, *Space-Time Block Coding for Wireless Communications*. Cambridge, UK: Cambridge University Press, 2003.

[88] L. S. Lasdon, *Optimization Theory for Large Systems*. New York: Macmillian, 1970.

[89] H. Lebret and S. Boyd, "Antenna array pattern synthesis via convex optimization," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 526–532, March 1997.

[90] K. H. Lee and D. P. Petersen, "Optimal linear coding for vector channels," *IEEE Transactions on Communications*, vol. COM-24, no. 12, pp. 1283–1290, December 1976.

[91] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra and Applications*, vol. 284, no. 1–3, pp. 193–228, July 1998.

[92] A. Lozano, A. Tulino, and S. Verdú, "Optimum power allocation for parallel Gaussian channels with arbitrary input distributions," *IEEE Transactions on Information Theory*, vol. 52, no. 7, pp. 3033–3051, July 2006.

[93] D. G. Luenberger, *Optimization by Vector Space Methods.* New York: Wiley, 1969.

[94] Z.-Q. Luo, T. N. Davidson, G. B. Giannakis, and K. M. Wong, "Transceiver optimization for block-based multiple access through ISI channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 4, pp. 1037–1052, April 2004.

[95] W. K. Ma, T. N. Davidson, K. M. Wong, Z. Q. Luo, and P. C. Ching, "Quasi-maximum-likelihood multiuser detection using semi-definite relaxation with application to synchronous CDMA," *IEEE Transactions on Signal Processing*, vol. 50, no. 4, pp. 912–922, April 2002.

[96] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics.* New York: Wiley, 1999.

[97] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications.* New York: Academic Press, 1979.

[98] J. Milanovic, T. N. Davidson, Z.-Q. Luo, and K. M. Wong, "Design of robust redundant precoding filter banks with zero-forcing equalizers for unknown frequency-selective channels," in *Proceedings of 2000 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2000)*, Istanbul, Turkey. pp. 2761–2764, June 2000.

[99] R. A. Monzingo and T. W. Miller, *Introduction to Adaptive Arrays.* New York, NY, USA: Wiley, 1980.

[100] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Transactions on Circuits and Systems*, vol. CAS-23, no. 9, pp. 551–562, September 1976.

[101] A. Narula, M. J. Lopez, M. D. Trott, and G. W. Wornell, "Efficient use of side information in multiple-antenna data transmission over fading channels," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1423–1435, October 1998.

[102] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1293–1302, July 1993.

[103] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 969–996, November 2006.

[104] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming.* Vol. 13, Philadelphia, PA: SIAM, Studies in Applied Mathematics, 1994.

[105] F. Oggier, G. Rekaya, J.-C. Belfiore, and E. Viterbo, "Perfect space-time block codes," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3903–3912, September 2006.

[106] E. N. Onggosanusi, A. M. Sayeed, and B. D. Van Veen, "Efficient signaling schemes for wideband space-time wireless channels using channel state information," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 1, pp. 1–13, January 2003.

[107] D. P. Palomar and M. Chiang, "Alternative distributed algorithms for network utility maximization: Framework and applications," *IEEE Transactions on Automatic Control*, to appear 2007.

[108] D. P. Palomar, *A Unified Framework for Communications through MIMO Channels*. PhD thesis, Technical University of Catalonia (UPC), Barcelona, Spain, May 2003.

[109] D. P. Palomar, "Convex primal decomposition for multicarrier linear MIMO transceivers," *IEEE Transactions on Signal Processing*, vol. 53, no. 12, December 2005.

[110] D. P. Palomar, M. Bengtsson, and B. Ottersten, "Minimum BER linear transceivers for MIMO channels via primal decomposition," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2866–2882, August 2005.

[111] 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 Transactions on Signal Processing*, vol. 51, no. 9, pp. 2381–2401, September 2003.

[112] D. P. Palomar and J. R. Fonollosa, "Practical algorithms for a family of waterfilling solutions," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 686–695, February 2005.

[113] D. P. Palomar and M. A. Lagunas, "Simplified joint transmit-receive space-time equalization on spatially correlated MIMO channels: A beamforming approach," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 730–743, June 2003.

[114] D. P. Palomar, M. A. Lagunas, and J. M. Cioffi, "Optimum linear joint transmit-receive processing for MIMO channels with QoS constraints," *IEEE Transactions on Signal Processing*, vol. 52, no. 5, pp. 1179–1197, May 2004.

[115] A. Paulraj, R. Nabat, and D. Gore, *Introduction to Space-Time Wireless Communications*. Cambridge, UK: Cambridge University Press, 2003.

[116] W. W. Peterson and J. E. J. Weldon, *Error-Correcting Codes*. Cambridge, Massachusetts, and London, England: The MIT Press, second ed., 1978.

[117] B. Picinbono, "On Circularity," *IEEE Transactions on Signal Processing*, vol. 42, no. 12, pp. 3473–3482, December 1994.

[118] B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *IEEE Transactions on Signal Processing*, vol. 43, no. 8, pp. 2030–2033, August 1995.

[119] H. V. Poor and S. Verdú, "Probability of error in MMSE multiuser detection," *IEEE Transactions on Information Theory*, vol. 43, no. 3, pp. 858–871, May 1997.

[120] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, third ed., 1995.

[121] S. U. H. Qureshi, "Adaptive Equalization," *Proceedings of the IEEE*, vol. 73, no. 9, pp. 1349—1387, September 1985.

[122] G. G. Raleigh and J. M. Cioffi, "Spatio-temporal coding for wireless communication," *IEEE Transactions on Communications*, vol. 46, no. 3, pp. 357–366, March 1998.

[123] F. Rey, M. Lamarca, and G. Vázquez, "Robust power allocation algorithms for MIMO OFDM systems with imperfect CSI," *IEEE Transactions on Signal Processing*, vol. 53, no. 3, pp. 1070–1085, March 2005.

[124] R. T. Rockafellar, *Convex Analysis.* Princeton, NJ: Princeton Univ. Press, second ed., 1970.

[125] R. T. Rockafellar, “Lagrange multipliers and optimality,” *SIAM Review*, vol. 35, no. 2, pp. 183–238, 1993.

[126] Y. Rong, S. A. Vorobyov, and A. B. Gershman, “Robust linear receivers for multi-access space-time block coded MIMO systems: A probabilistically constrained approach,” *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1560–1570, August 2006.

[127] M. Rupf and J. L. Massey, “Optimal sequence multisets for synchronous code-division multiple-access channels,” *IEEE Transactions on Information Theory*, vol. 40, pp. 1261–1266, July 1994.

[128] J. Salz, “Optimum mean-square decision-feedback equalization,” *Bell System of Techncial Journal*, vol. 52, pp. 1341–1373, October 1973.

[129] J. Salz, “Digital transmission over cross-coupled linear channels,” *At&T Technical Journal*, vol. 64, no. 6, pp. 1147–1159, July–August 1985.

[130] H. Sampath, P. Stoica, and A. Paulraj, “Generalized linear precoder and decoder design for MIMO channels using the weighted MMSE criterion,” *IEEE Transactions on Communications*, vol. 49, no. 12, pp. 2198–2206, December 2001.

[131] A. Scaglione, G. B. Giannakis, and S. Barbarossa, “Redundant filterbank precoders and equalizers. Part I: Unification and optimal designs,” *IEEE Transactions on Signal Processing*, vol. 47, no. 7, pp. 1988–2006, July 1999.

[132] A. Scaglione, P. Stoica, S. Barbarossa, G. B. Giannakis, and H. Sampath, “Optimal designs for space-time linear precoders and decoders,” *IEEE Transactions on Signal Processing*, vol. 50, no. 5, pp. 1051–1064, May 2002.

[133] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis.* Reading, MA: Addison-Wesley, 1991.

[134] P. J. Schreier and L. L. Scharf, “Second-order analysis of improper complex random vectors and processes,” *IEEE Transactions on Signal Processing*, vol. 51, no. 3, pp. 714–725, March 2003.

[135] M. Schubert and H. Boche, “QoS-based resource allocation and transceiver optimization,” *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 6, pp. 383–529, 2006.

[136] M. Schubert, S. Shuying, E. A. Jorswieck, and H. Boche, “Downlink sum-MSE transceiver optimization for linear multi-user MIMO systems,” in *Proceedings of 39th Asilomar Conference on Signals, Systems and Computers*, (Pacific Grove, CA), pp. 1424–1428, October 28–November 1 2005.

[137] I. Schur, “On the characteristic roots of a linear substitution with an application to the theory of integral equations,” *Mathematique Annalen*, vol. 66, pp. 488–510, 1909.

[138] S. Serbetli and A. Yener, “Transceiver optimization for multiuser MIMO systems,” *IEEE Transactions on Signal Processing*, vol. 52, no. 1, pp. 214–226, January 2004.

[139] S. Shamai and L. Laroia, “The inter-symbol interference channel: Lower bounds on capacity and channel precoding loss,” *IEEE Transactions on Information Theory*, vol. 42, no. 9, pp. 1388–1404, September 1996.

[140] C. E. Shannon, “Communication in the presence of noise,” *Proceeding IRE*, vol. 37, no. 1, pp. 10–21, January 1949.

[141] M. B. Shenouda and T. N. Davidson, “A framework for designing MIMO systems with decision feedback equalization or Tomlinson-Harashima precoding,” in *Proceedings of 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’07)*, (Honolulu, Hawaii), pp. 15–20, April 2007.

[142] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions.* Berlin, Germany: Springer-Verlag, 1985.

[143] T. Söderström and P. Stoica, *System Identification.* London, UK: Prentice Hall International, 1989.

[144] T. Starr, J. M. Cioffi, and P. J. Silverman, *Understanding Digital Subscriber Line Technology.* Upper Saddle River, NJ: Prentice Hall, 1999.

[145] J. F. Sturm, “Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones,” *Optimization Methods and Software*, vol. 11 and 12, pp. 625–653, 1999.

[146] V. Tarokh, N. Seshadri, and A. R. Calderbank, “Space-time codes for high data rate wireless communication: Performance criterion and code construction,” *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 744–765, March 1998.

[147] I. E. Telatar, “Capacity of multi-antenna Gaussian channels,” *AT&T Bell Labs, Internal Tech. Memo*, June 1995.

[148] I. E. Telatar, “Capacity of multi-antenna Gaussian channels,” *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, November–December 1999.

[149] M. Tomlinson, “New automatic equaliser employing modulo arithmetic,” *Electronic Letters*, vol. 7, pp. 138–139, March 1971.

[150] H. V. Trees, *Detection, Estimation, and Modulation Theory. Part IV: Optimum Array Processing.* New York: Wiley, 2002.

[151] D. N. C. Tse and P. Viswanath, *Fundamentals of Wireless Communications.* Cambridge, U.K.: Cambridge University Press, 2005.

[152] B. S. Tsybakov, “Capacity of a vector Gaussian channel without memory,” *Problems of Information Transmission*, vol. 1, pp. 26–40, 1965.

[153] B. S. Tsybakov, “Capacity of a discrete-time Gaussian channel with a filter,” *Problems of Information Transmission*, vol. 6, pp. 253–256, July/September 1970.

[154] B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, April 1988.

[155] L. Vandenberghe and S. Boyd, “Semidefinite programming,” *SIAM Review*, vol. 38, no. 1, pp. 49–95, March 1996.

[156] M. K. Varanasi and T. Guess, “Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel,” *Proceedings of the Thirty-First Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1405–1409, November 2–5 1997.

[157] S. Verdú, *Multiuser Detection.* New York, NY, USA: Cambridge University Press, 1998.

[158] S. Vishwanath, N. Jindal, and A. Goldsmith, “Duality, achievable rates and sum-rate capacity of Gaussian MIMO broadcast channels,” *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2658–2668, October 2003.

[159] E. Visotsky and U. Madhow, “Space-time transmit precoding with imperfect feedback,” *IEEE Transactions on Information Theory*, vol. 47, no. 6, pp. 2632–2639, September 2001.

[160] P. Viswanath and D. N. C. Tse, “Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality,” *IEEE Transactions on Information Theory*, vol. 49, pp. 1912–1921, August 2003.

[161] P. Viswanath and V. Anantharam, “Optimal sequences and sum capacity of synchronous CDMA systems,” *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 1984–1991, September 1999.

[162] P. Viswanath and V. Anantharam, “Optimal sequences for CDMA under colored noise: A schur-saddle function property,” *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1295–1318, June 2002.

[163] P. Viswanath, V. Anantharam, and D. N. C. Tse, “Optimal sequences, power control, and user capacity of synchronous CDMA systems with linear MMSE multiuser receivers,” *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 1968–1983, September 1999.

[164] E. Viterbo and J. Boutros, “A universal lattice decoder for fading channels,” *IEEE Transactions on Information Theory*, vol. 45, no. 7, pp. 1639–1642, July 2000.

[165] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, “Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem,” *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 313–324, February 2003.

[166] H. Weyl, “Inequalities between two kinds of eigenvalues of a linear transformation,” *Proceedings of the Natuaral Academic Science USA*, vol. 35, pp. 408–411, 1949.

[167] J. H. Wilkinson and C. Reinsch, “Linear Algebra,” in *Handbook for Automatic Computation*, (F. L. Bauer, ed.), Berlin: Springer-Verlag, 1971.

[168] S.-P. Wu, S. Boyd, and L. Vandenberghe, “FIR filter design via semidefinite programming and spectral factorization,” in *Proceedings of IEEE Conference on Decision and Control*, pp. 271–276, December 1996.

[169] F. Xu, T. N. Davidson, J. K. Zhang, and K. M. Wong, “Design of block transceivers with decision feedback detection,” *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp. 965–978, March 2006.

[170] J. Yang and S. Roy, “Joint transmitter-receiver optimization for multi-input multi-output systems with decision feedback,” *IEEE Transactions on Information Theory*, vol. 40, no. 5, pp. 1334–1347, September 1994.

[171] J. Yang and S. Roy, “On joint transmitter and receiver optimization for multiple-input-multiple-output (MIMO) transmission systems,” *IEEE Transactions on Communications*, vol. 42, no. 12, pp. 3221–3231, December 1994.

[172] W. Yu and J. M. Cioffi, “Sum capacity of a Gaussian vector broadcast channel,” *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 1875–1892, September 2004.

[173] W. Yu, D. P. Varodayan, and J. M. Cioffi, “Trellis and convolutional precoding for transmitter-based interference presubtraction,” *IEEE Transactions on Communications*, vol. 53, no. 7, pp. 1220–1230, July 2005.

[174] W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi, “Iterative water-filling for Gaussian vector multiple-access channels,” *IEEE Transactions on Information Theory*, vol. 50, no. 1, pp. 145–152, January 2004.

[175] J.-K. Zhang, A. Kavčić, X. Ma, and K. M. Wong, “Design of Unitary Precoders for ISI Channels,” in *Proceedings IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 2265–2268, Orlando, Florida, 2002.

[176] J.-K. Zhang, A. Kavčić, and K. M. Wong, “Equal-diagonal QR decomposition and its application to precoder design for successive-cancellation detection,” *IEEE Transactions on Information Theory*, vol. 51, no. 1, pp. 154–172, January 2005.

[177] X. Zhang, D. P. Palomar, and B. Ottersten, “Robust design of linear MIMO transceivers,” *submitted to IEEE Transactions on Signal Processing*, July 2006.

[178] L. Zheng and D. N. C. Tse, “Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels,” *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.