Optimal Transceiver Design for Non-Regenerative MIMO Relay Systems

Chao Zhao

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

October 2013

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

2013/10/30

Abstract

Multiple-input multiple-output (MIMO) relaying can increase system throughput, overcome shadowing and expand network coverage more efficiently than its single-antenna counterpart. Non-regenerative (amplify-and-forward) strategies, in which the relays apply linear transformation matrices to their received baseband signals before retransmitting them, are favored in many applications due to low processing delays and implementation complexity. In this regard, transceiver design is crucial to fulfilling the great potential of MIMO relay communication systems. In this thesis, we explore this general problem from two different perspectives: coherent combining and adaptation.

Within the first perspective, we design linear transceivers for a one-source–multiple-relays–one-destination system in which the source sends information to the destination through multiple parallel relay stations, such that the signals from these relays are coherently combined at the destination to benefit from distributed array gain. Two approaches are proposed: a low-complexity structured hybrid framework and a minimum mean square error (MSE) optimization approach. In the first approach, the non-regenerative MIMO relaying matrix at each relay is generated by cascading two substructures, akin to an equalizer for the backward channel and a precoder for the forward channel. For each of them, we introduced one-dimensional parametric families of candidate matrix transformations. This hybrid framework allows for the classification and comparison of all possible combinations of these substructures, including several previously investigated methods and their generalizations. The design parameters can further be optimized based on individual channel realizations or on channel statistics; in the latter case, the optimum parameters can be well approximated by linear functions of the signal-to-noise ratios (SNRs). This hybrid framework achieves a good balance between performance and complexity. In the second approach, the relaying matrices are designed to minimize the MSE between the transmitted and received signal symbols. Two types of constraints on the transmit power of the relays are considered separately: *weighted sum* and *per-relay* power constraints. Under the weighted sum power constraint, we are able to derive a closed-form expression for the optimal solution, by introducing a complex scaling factor at the destination and using Lagrangian duality. Under the per-relay power constraints, we propose a power balancing algorithm that converts the problem into an equivalent one with a weighted sum power constraint. In addition, we investigate the joint design of the MIMO equalizer at the

destination and the relaying matrices, using block coordinate descent or steepest descent. The bit-error rate (BER) simulation results demonstrate better performance than previous methods.

Within the second perspective, we propose a unified framework for adaptive transceiver optimization for non-regenerative MIMO relay networks. Transceiver designs based on channel state information (CSI) implicitly assume that the underlying wireless channels remain almost constant within each transmission block. This implies that both the channels and the corresponding optimal transceivers evolve gradually across successive blocks. To benefit from this property, we propose a new inter-block adaptive approach based on the minimum MSE criterion, in which the optimum from the previous block is used as the initial search point for the current block. By optimizing the relaying matrices in the first place, we make this adaptation easy to implement by means of iterative algorithms such as the gradient descent. In addition, the proposed framework can accommodate various network topologies by imposing structural constraints on the system model, and leads to new and more efficient algorithms with better performance for certain topologies. We explain in detail how to handle these constraints for different system configurations. Numerical results demonstrate that inter-block adaptation can lead to a significant reduction in computational complexity.

Sommaire

Conception Optimale de Transcepteurs pour les Systèmes à Relais Non-Régénératifs MIMO

Le relayage multi-entrées multi-sorties (MIMO) au moyen de réseaux d'antennes permet d'accroître la capacité des systèmes sans fil, pallier aux effets d'ombrage et augmenter la couverture de réseaux plus efficacement que sa contrepartie n'utilisant qu'une seule antenne. Les stratégies non-régénératives (amplification-et-suivie), dans lesquelles les relais appliquent des matrices de transformation linéaire à leur signaux d'entrée avant de les retransmettre, sont préférées dans de nombre d'applications, en raison de leur faibles délai de traitement et complexité de mise en uvre. À cet égard, la conception de transcepteurs est cruciale afin de pleinement exploiter le grand potentiel qu'offrent les relais MIMO dans les systèmes de communications sans fil. Dans cette thèse, nous explorons ce problème général à partir de deux perspectives différentes: la combinaison cohérente et l'adaptation.

Dans la première perspective, nous concevons des architectures de transcepteur pour un système de type source-simplerelais-multiplesdestination-simple (1S-MR-1D) dans lequel la source envoie de l'information à la destination par le biais de plusieurs stations relais en parallèle, de telle sorte que les signaux en provenance des relais se combinent de manière cohérente à la destination afin de bénéficier du gain d'un réseau antennes distribuées. cette fin, deux approches sont proposées : un schème reposant sur une structure hybride à complexité réduite et une approche d'optimisation basée sur la minimisation de l'erreur quadratique moyenne (MSE). Dans la première approche, les matrices de transformation MIMO non-régénératives utilisées à chacun des relais sont obtenues en cascadant deux sous-structures, qui s'apparentent à un égalisateur pour le canal en réception et à un pré-codeur pour le canal en transmission, respectivement. Pour chacune de ces sous-structures, nous introduisons une famille paramétrique uni-dimensionnelle de transformations matricielles. Ce schème hybride permet la classification et la comparaison de toutes les combinaisons possibles de ces sous-structures, qui incluent plusieurs méthodes déjà existantes de même que leur généralisation. Les paramètres de conception peuvent de plus être optimisés, soit pour des réalisations individuelles des canaux de transmission, soit en se basant sur leurs statistiques. Dans ce dernier cas, les paramètres optimaux peuvent être approximés par des fonctions linéaires des rapports signal-sur-bruit. Le schème hybride permet d'atteindre

un bon équilibre entre la performance et la complexité. Dans la deuxième approche, les matrices de relayage sont conçues de façon à minimiser la MSE entre les symboles transmis et reçus. On considère séparément deux types de contraintes sur la puissance en transmission des relais: la contrainte dite de somme pondérée et la contrainte par relais. Sous la contrainte de somme pondérée, nous développons des expressions mathématiques explicites pour la solution optimale via l'introduction d'un facteur de gain complexe à la destination et l'utilisation de la dualité Lagrangienne. Sous la contrainte de puissance par relais, nous proposons un algorithme de balancement qui permet de convertir le problème d'optimisation en un problème équivalent avec contrainte de type somme pondérée. De plus, nous étudions le problème de la conception jointe de l'égalisateur MIMO à la destination et des matrices de relais, en considérant l'optimisation par descente selon coordonnées successives ou selon la plus forte pente. Les taux d'erreurs binaires obtenus par simulation démontrent une performance supérieure à celle de méthodes existantes.

Dans la deuxième perspective, nous proposons un cadre unifié d'optimisation des transcepteurs adaptatifs pour les réseaux de relais MIMO non-régénératifs. La conception de transmetteur basée sur l'information de l'état du canal (CSI) suppose implicitement que les canaux sans fil demeurent constants durant chaque bloc de transmission. Cela implique que les canaux et les transcepteurs optimaux correspondants évoluent graduellement au passage des blocs. Afin de bénéficier de cette propriété, nous proposons une nouvelle approche d'adaptation inter-bloc basée sur le critère de minimisation de la MSE. Dans cette approche, la solution obtenue du bloc précédent est utilisée comme point de départ dans la recherche d'une solution optimale pour le bloc actuel. Fortuitement, il est possible d'optimiser les matrices de relayage de façon analytique en premier lieu, ce qui facilite grandement l'adaptation des paramètres restants au moyen d'algorithmes itératifs tels que celui de la descente de gradient. De plus, le cadre d'optimisation que nous proposons peut être adapté à des topologies de réseau variées par l'imposition de contraintes structurelles sur le modèle. Nous expliquons en détail comment réaliser de telles contraintes pour différentes configurations de système. Les résultats de simulations numériques démontrent que l'adaptation inter-bloc peut conduire à une réduction importante de la complexité numérique.

Acknowledgments

Let me express my sincere gratitude to my supervisor, Prof. Benoit Champagne, for his generosity, supportiveness, open-mindedness and humor. This thesis would never have happened without his guidance, help and encouragement. I would also thank him for my opportunities of giving tutorials, attending academic conferences, presenting my research to industrial partners and performing paper reviews. Moreover, what I have learned from him is beyond academics.

I am also grateful for the financial support provided by McGill University through the McGill International Doctoral Award, and by Prof. Champagne via research grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Government of Quebec through the PROMPT program and InterDigital Canada Ltée.

I would also like to thank Prof. Harry Leib and Prof. Xiao-Wen Chang for their roles as members of my supervisory committee. Their invaluable advices greatly helped to improve the technical quality of my research work. I also extend my thanks to the faculty and staff at Department of Electrical and Computer Engineering.

I have always been enjoying my life in the Telecommunication and Signal Processing group. Many thanks to my colleagues for the past five years, including Arash, Ayman, Bo, Chao-Cheng, Eric, Jiaxin, Khalid, Neda, Reza, Siamak, Xiao and Yunhao. Fang and Siavash: you are so special to me because we started and graduated together. It was a great honor to be your company.

This journey would have been much more difficult without my best friends, Yang, Di, Meng, Feng, Tao, Jian, Zhe, Charlie and Joyce. I would also like to thank Viktor Frankl for his great book "Man's search for meaning".

I am also forever indebted to my mom and dad for their love and support. They have given me everything they could, and they have been encouraging me to make my own decisions since I was a child.

Contents

List of Figures

List of Tables

List of Acronyms

1S-1R-1D	one-source–one-relay–one-destination
1S-MR-1D	one-source–multiple-relays–one-destination
1S-1R-MD	one-source–one-relay–multiple-destinations
3GPP	Third Generation Partnership Project
AF	amplify-and-forward
AGC	automatic gain control
BC	broadcast channel
BD	block diagonalization
BER	bit-error rate
BS	base station
CCI	co-channel interference
CDI	channel distribution information
CF	compress-and-forward
CoMP	coordinated multi-point
CSI	channel state information
DF	decode-and-forward
DFE	decision feedback equalizer
DMT	discrete multi-tone
DPC	dirty paper coding
EVD	eigenvalue decomposition
FDD	frequency-division duplex
GMD	geometrical mean decomposition
GP	geometric programming
IC	interference channel

JMMSE	Joint MMSE
KKT	Karush-Kuhn-Tucker
LDPC	low density parity check
LHS	left hand side
LMMSE	linear MMSE
LTE	Long Term Evolution
MAC	multiple access channel
MF	matched filtering
MIMO	multiple-input multiple-output
MMSE	minimum mean square error
MSE	mean square error
OFDM	orthogonal frequency division multiplexing
QCQP	quadratically constrained quadratic program
QoS	Quality-of-Service
RF	radio frequency
RHS	right hand side
SAF	simplistic amplify-and-forward
SDMA	spatial division multiple access
SDP	semidefinite programming
SER	symbol-error rate
SINR	signal-to-interference plus noise ratio
SISO	single-input single-output
SNR	signal-to-noise ratio
SOCP	second-order cone programming
SVD	singular value decomposition
TCM	trellis-coded modulation
TDD	time-division duplex
THP	Thomlinson-Harashima precoding
V-BLAST	Vertical-Bell-Laboratories-Layered-Space-Time
WLAN	wireless local area network
w.l.o.g.	without loss of generality
w.r.t.	with respect to
ZF	zero-forcing

Chapter 1

Introduction

Emerging applications such as multimedia and cloud computing are putting very stringent requirement of very high data transmission rates on mobile wireless devices. Since spectral efficiency for point-to-point transmissions is already close to the theoretical limit, the quest for faster wireless communications has shifted focus to novel heterogeneous network topologies. Wireless relays using multiple-input multiple-output (MIMO) technology are indispensable components of these networks, and the optimization of their transceiver subsystems is crucial to fulfilling the great potential of MIMO relay communications. Therefore, in this thesis, we explore two important aspects of transceiver design for MIMO relay systems: combining and adaptation. The former stands for coherent combining at the destination, of signals transmitted from multiple relays, in order to benefit from a distributed array gain; the latter refers to appropriate exploitation of time-domain properties of wireless channels to reduce algorithmic complexity. This chapter presents the background, rationale and objectives of our research, together with a summary of our contributions.

1.1 The Pursuit of High Transmission Rates

The wireless telecommunication industry has been benefitting from tremendous growth in transmission rates and constant reduction in device size and power consumption, which in turn has enabled sophisticated multimedia applications on mobile devices. On the one hand, thanks to Moore's law, we can "fabricate" state-of-the-art digital signal processing technologies into chips with smaller die area, less power consumption, lower production cost and higher computational capability. On the other hand, limited spectrum resources, re-

ceiver noises and time-varying wireless channels are constantly challenging mathematicians and engineers to develop new coding schemes, communication architectures, and signal processing algorithms to improve performance of communication systems. In addition, advances in battery capacity, antenna technology and radio frequency (RF) circuit design, although not quite as revolutionary, have been contributing as well.

Within the modern wireless paradigm, increasing system throughput remains as the central problem of communication system design. Mobile subscribers, service providers and telecommunication engineers all want higher data rates in a cost-effective manner. Transmitting over wider frequency bands is a straightforward solution but licensed spectrum resources are extremely scarce and expensive. This leads to a natural approach of making the most of available frequency bands, namely, to develop communication architectures with high *spectral efficiency*. However, similar to bandwidth, spectral efficiency cannot go to infinity as well: it is upper bounded by the well-known Shannon limit. To make things worse, Turbo codes and low density parity check (LDPC) codes have come very close to this limit [1].

With not so much room left for channel coding schemes to accomplish towards the Shannon limit, a new *spatial* dimension was introduced by deploying multiple antennas at either side of a communication link. The use of transmit or receive beamforming contributes to an array gain — M-fold improvement in signal-to-noise ratio (SNR). MIMO technology goes further by equipping both the transmitter and the receiver with multiple antennas, which is arguably the most significant breakthrough in communication theory and technology during the last two decades. It can bring diversity gain, array gain and spatial-multiplexing gain.

- MIMO improves the reliability of a communication link through *spatial diversity* [2]. There is a high probability that one of the antennas is not in deep fading. Unlike time diversity and frequency diversity, spatial diversity improves link reliability without wasting time or bandwidth. By using proper space-time coding, one can almost always obtain a diversity gain [3].

- MIMO benefits from an *array gain* in SNR due to coherent combing of signals across multiple antennas. This requires knowledge of the channel state information (CSI) at the transmitter or the receiver.

- MIMO provides a *spatial-multiplexing gain* attributable to the additional degrees of

freedom [3, 4]. This is done by simultaneously transmitting multiple independent symbol streams that can be efficiently separated by the receiver.

It is worth mentioning that these three gains cannot be maximized at the same time. The limited number of available degrees of freedom is the reason for the fundamental trade-off between diversity and multiplexing [5]. In the last two decades, various theoretic limits, coding schemes, transceiver architectures (linear and nonlinear), channel estimation and feedback algorithms have been investigated exhaustively to achieve capacity and/or increase reliability of MIMO communications [6]. MIMO has been widely deployed in state-of-the-art systems such as Long Term Evolution (LTE) and 802.11n/ac wireless local area network (WLAN).

Apart from improving spectral efficiency for point-to-point communications, a parallel evolution at the *network* level is also under way. First and foremost, arguably the most important invention in wireless system architectures is the concept of a "cell". Cellular networks allow frequency reuse across neighbouring cells, which makes it possible to increase spectral efficiency *per unit area* indefinitely (in theory) by using smaller and smaller cells [2]. However, this requires a large number of complex and expensive base stations. Instead, emerging standards such as LTE-Advanced are embracing *heterogeneous networks* that use a mix of macro, pico, femto and relay base stations. Macro base stations typically transmit at high power level, whereas pico, femto and relay nodes transmit at substantially lower power levels. The low-power nodes are deployed to extend coverage and improve capacity in hot spots. They have smaller physical size and therefore offer flexible site acquisitions. The bottom line is that heterogeneous networks enable flexible and low-cost deployments while providing a uniform broadband experience to users anywhere within the network [7].

Spectral efficiency can be further improved by allocating the whole frequency spectrum among different communication systems dynamically and efficiently. Licensed frequency bands may not be used during certain periods by their owners and secondary users may use them. This dynamic spectrum detection and allocation strategy, which falls into the topic of cognitive radio, "squeezes" some additional bits out of the crowded frequency spectrum.

1.2 MIMO Wireless Relaying

As indispensable components of heterogeneous networks, wireless relays may serve different purposes [8]. To begin with, wireless relaying can overcome the impairments caused by multipath fading, shadowing and path losses in cellular communication systems. The deployment of relays can extend coverage and improve the Quality-of-Service (QoS) of those users near the cell boundary effectively and economically. This is critical for providing uniform broadband experience to users anywhere within the network, thereby increasing spectral efficiency per unit area. Another advantage is that relay stations provide better spatial resolution because they are closer to the users than base stations. Consequently, co-channel interferences are lower and more users can be served simultaneously using the same time slots and frequency bands, contributing to higher system throughput. In cooperative communications, user equipments with better link quality and longer battery life can serve as relays for those with weaker links or in deep fading [9]. In addition, in wireless sensor networks, multiple distributed sensors can serve as relay nodes for a source-destination pair, thereby reducing power consumption, increasing data rates or extending coverage [10]. These aspects of wireless relaying are summarized in Fig. 1.1.

Performance of a wireless relay network depends on two major factors — network configuration and relaying strategy. Network configuration refers to how many sources, relays and destinations are involved in the communication process (network topology), how many antennas each node has, and how the wireless propagation environment evolves with time and space. An entire cellular network can be regarded as a collection of smaller-scale *primitive* subnetworks with simple topologies. Only by studying these primitives can the more complicated networks be understood with insights. For notational convenience, we describe primitive network layouts using the number of sources, relays and destinations. For example, 1S-MR-1D refers to one-source–multiple-relays–one-destination.

For a particular configuration, system performance then depends on how information is retransmitted at the relays. Relaying strategies can be classified in different ways. Firstly, they can be *non-regenerative* such as amplify-and-forward (AF), or *regenerative* such as decode-and-forward (DF) [11]. Non-regenerative AF relaying applies linear processing to the received signals and then send the transformed signals to the destination(s). This linear processing is usually represented by scalars for single-antenna relays, or matrices for multi-antenna arrays. In contrast, a DF relay decodes binary bits from its received signals, and

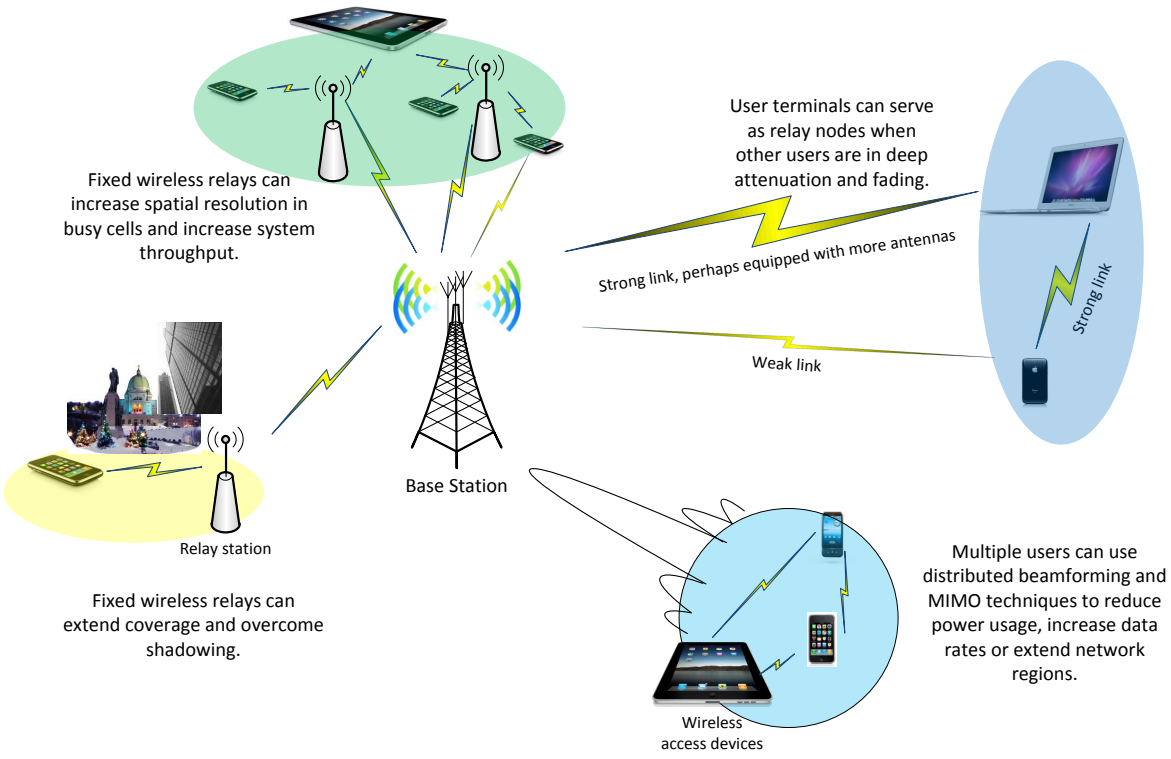

Fig. 1.1 Benefits of wireless relaying

then encodes, modulates and retransmits to the destination(s). Regenerative strategies are more sophisticated but suffer from longer delays, higher overall costs (especially for MIMO relays) and security problems. Therefore, we concentrate on non-regenerative relaying strategies in this thesis.

Secondly, relay stations can work in *full-duplex* or *half-duplex* mode. A half-duplex relay receives in the first time slot and transmits in the next one; full-duplex relays are generally difficult to implement because high-power transmitted signals from the relays would saturate the co-located receivers. The penalty in the overall spectral efficiency due to half-duplex relaying can be well compensated by the capacity gains obtained by wireless relaying. Thirdly, relaying can be one-way or two-way. In the latter approach, the source and the destination simultaneously transmit to the relay in the first hop, and the relay

broadcast signals back in the second hop [12]. In this thesis, we consider half-duplex one-way relaying.

The introduction of MIMO technology into the relaying framework, through the use of multiple antennas at the sources, relays or destinations, brings further advantages in terms of achievable performance [13], yet creates new challenges for system designers. For DF, optimal transceiver design for a relay station are separated into two parts: the MIMO receiver for the first hop transmission and the MIMO transmitter for the second hop. Implementation of these MIMO decoder and encoder increases complexity, cost and processing delay at the relays. For AF, relaying strategies are now represented by matrices, and this shift from scalars to matrices complicates transceiver design significantly. Compared with MIMO systems *without* relay, there now exist two independent noise components, the additive noise induced at the relays after the first hop, and the additive noise at the destination after the second hop. The former is affected by the relay processing matrices and a good relaying strategy must prevent over-amplification of this noise component. In the past few years, MIMO relaying has been attracting considerable interest among researchers and engineers.

1.3 Objectives and Contributions

1.3.1 Rationale and Objectives

In non-regenerative MIMO relay communication systems, wireless channels and additive noises and interferences are uncontrollable factors that affect the performance of the complete data link. Therefore, the goal of physical layer system design is to choose various processing modules of the whole relay communication system, in order to optimize some suitable performance criterion which takes into account the random and time-varying natures of the channels, noises and interferences. This design problem is complicated by the existence of both linear and nonlinear processing modules along the transmission chain. The former include components as source precoders, relay processing matrices and destination equalizers, whereas the latter include modulation, channel coding, space-time coding, interleaving, nonlinear precoding and hard/soft-decision decoding. To optimize all these components simultaneously is almost impossible and does not provide many insights as well. Furthermore, in case of slight changes in the radio environment (e.g. channel or noise properties), the entire system may need to be redesigned at the great expense.

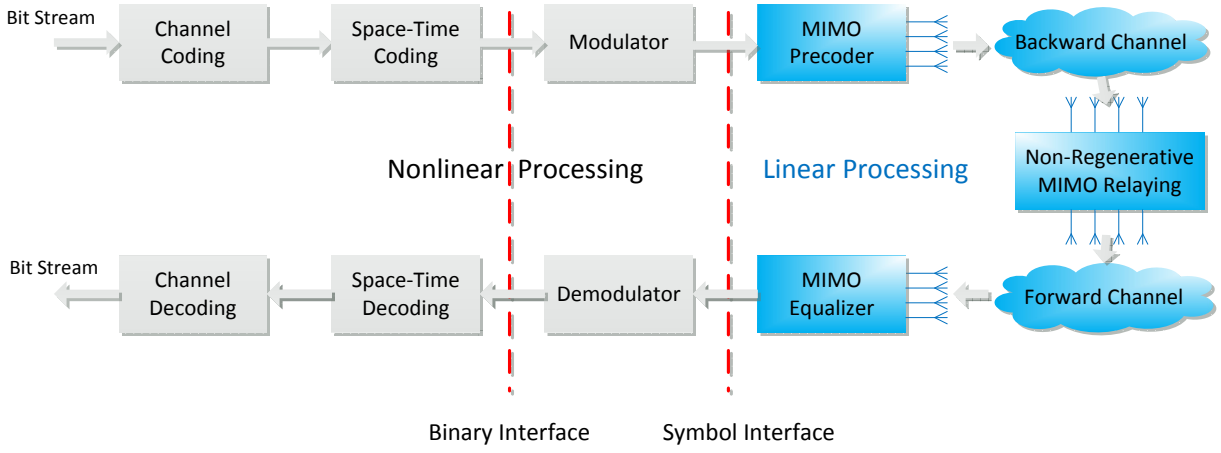

Fig. 1.2 A layered approach to nonregenerative MIMO relaying

A more popular and efficient approach, instead, is to break the complete physical layer communication link into so-called *sub-layers* and optimize these sub-layers individually.[1] Fig. 1.2 illustrates the sub-layers of a typical non-regenerative MIMO relay system. At the source user, data bits to be sent are processed by a pipeline of nonlinear processing modules such as channel coder, space-time coder (including interleaver) and modulator.[2] The resulting multiple symbol streams are mapped by a linear precoder, and then transmitted via multiple antennas to the relay. After propagating through the backward (source-to-relay) wireless channel, these signals are linearly processed by the relay and retransmitted to the destination through the forward (relay-to-destination) channel. At the destination, the received signals go through a reverse pipeline of linear equalizer and nonlinear modules including demodulator, space-time decoder and hard/soft-decision decoder. We may view this communication system as an overlay of different sub-layers, with the interfaces between them well defined. The innermost *linear processing* sub-layer includes linear MIMO precoding, linear relay processing and linear MIMO equalization, interfacing with the next outer sub-layer via symbol vectors. The next sub-layer consists of the modulator and demodulator modules, which interfaces with the next outer sub-layer via the exchange of binary data. The two outermost sub-layers are the space-time coder/decoder sub-layer

[1] In telecommunication engineering, the term "layer" usually refers to the seven layers of the Open System Interconnection (OSI) model, to which the physcal layer belongs [14]; we use the term sub-layers to avoid confusion with the OSI terminology.

[2] For simplicity, some trivial steps such as serial-to-parallel (S/P) conversion are not shown.

and the channel coder/decoder sub-layer, where all interactions now take place through binary interfaces.[3] In fact, discrete source coding/decoding, quantizer/table lookup and sampler/analog filter are nonlinear operations which could define several new sub-layers.[4] These components are well understood and pose few new challenges within the context of wireless relaying. Therefore, we do not consider these components as parts of the relay communication system and assume that the input and output are both bit streams.

The ultimate goal of this research is to improve *performance* of non-regenerative MIMO relay systems by designing transceiver architectures with *low complexity*. In particular, we concentrate on the design of the linear processing sub-layer, namely, the transformation matrices for the precoder, relay and equalizer. Channel coding, space-time coding and nonlinear transceiver architectures can improve system performance further and their designs have been extensively studied in the last decade. In addition, even though optimized linear components alone cannot guarantee global optimality, considering all sub-layers simultaneously would significantly complicates the problem at hand and provides few insights. Instead, we will turn to numerical simulations of the complete link including those nonlinear components to verify real-time performance.

To optimize the linear transceiver, one needs to choose a relevant performance measure in the first place. This objective function can be of an information-theoretic nature, such as mutual information or system throughput, or derived from statistical signal processing perspectives, e.g., mean square error (MSE), signal-to-interference plus noise ratio (SINR) and bit-error rate (BER). The transceiver maximizing the mutual information requires impractical Gaussian codes to achieve this maximum rate. Instead, statistical measures such as MSE and BER are more representative of real-time system performance. With this in mind, we shall emphasize the latter category of objective functions, MSE in particular.

In this thesis, we are particularly interested in two specific research goals: *combining* and *adaptation*. The first goal stands for coherent combining at the destination, of signals transmitted from multiple relays, in order to benefit from a distributed array gain. For the one-source–multiple-relays–one-destination (1S-MR-1D) system shown in Fig. 1.3, this combining is of utmost importance, which is different from the one-source–one-relay–one-destination (1S-1R-1D) system. In the latter, the optimal transceiver is well established in

[3] If schemes such as trellis-coded modulation (TCM) are used, the boundaries between some of these layers may be blurred and some layers may even merge together.

[4] In fact, some information such as a packet sent over a network is inherently digital and hence these components are unnecessary.

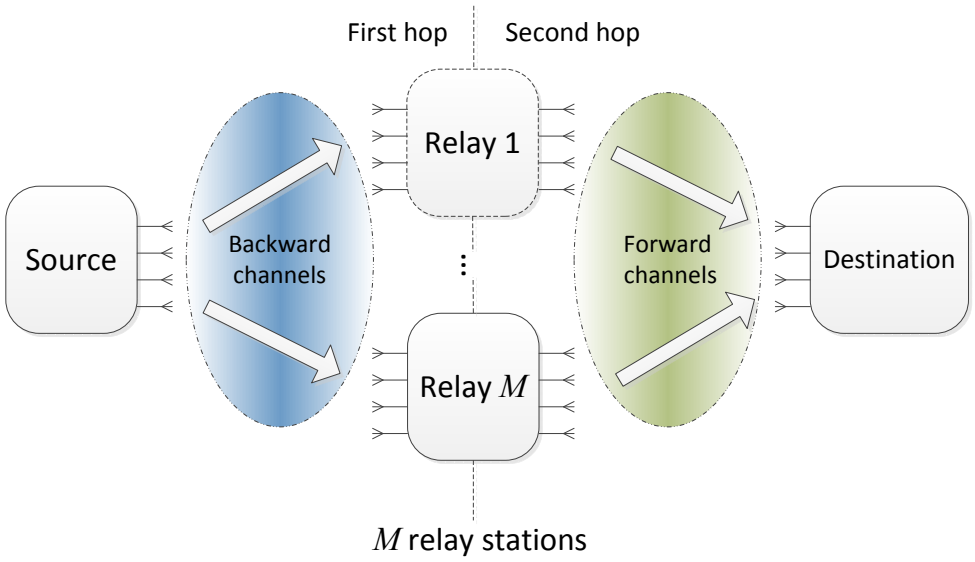

Fig. 1.3 A combining-type 1S-MR-1D system

terms of various performance criteria [15]. Interestingly, a majority of these criteria lead to a common singular value decomposition (SVD) structure. This SVD scheme, however, does not extend to the former 1S-MR-1D system since, due to the physically separated nature of the multiple relays, their combined transformation matrix inherits a block-diagonal form. Therefore, we shall study optimal design of the multiple relaying matrices for such *combining-type* 1S-MR-1D systems, aiming for the previously mentioned coherent combing but without over-amplifying the noises induced at the relay receivers. The existence of multiple antennas makes it difficult to balance these two aspects and henceforth we denote a majority part of this thesis to this research goal.

The second goal, adaptation, refers to appropriate exploitation of time-domain properties of wireless channels to reduce algorithmic complexity. Optimal transceiver design generally requires knowledge of the underlying wireless channels. In practice, the entire transmission period is divided into blocks. In each block, the channels are estimated and then used for transceiver optimization, followed by the actual data transmission. The block length is selected such that the channels stay almost constant within each block. Otherwise, model mismatch would deteriorate performance significantly. This implies that both the channels and the corresponding optimal transceivers evolve gradually across successive blocks. This property has been overlooked in previous development and evaluation

of transceiver designs for MIMO relay systems. Our purpose is to exploit it in order to simplify iterative algorithms that seemed complex when the successive blocks are viewed in isolation. This is especially important for multiuser MIMO relay networks because almost all the existing works turn to iterative numerical algorithms. In the following, we summarize the main contributions of this thesis.

1.3.2 Main Contributions

Transceiver design for combining-type relaying

We propose two different transceiver design approaches in order to leverage the distributed array gain for combining-type 1S-MR-1D systems.

In Chapter 3, we shall propose a *low-complexity hybrid framework* in which the MIMO relaying matrix at each relay is generated by cascading two substructures, akin to an equalizer for the backward channel and a precoder for the forward channel. For each of these two substructures, we introduce two one-dimensional parametric families of candidate matrix transformations. The first family, non-cooperative by nature, depends only on the backward or forward channel of the same relay. Specifically, this family includes zero-forcing (ZF), linear minimum mean square error (MMSE) and matched filtering (MF) as special cases, as well as other intermediate situations, thereby providing a continuous trade-off between interference and noise suppression. The second family, this one of a cooperative nature, makes use of additional information derived from the channels of other relays. This hybrid framework allows for the classification and comparison of all possible combinations of these substructure, including several previously investigated methods and their generalizations. The design parameters can be optimized based on individual channel realizations or on channel statistics; in the latter case, the optimum parameters can be well approximated by linear functions of the SNRs. We show that the optimal parameters differ significantly from those corresponding to the ZF, MF and linear MMSE. The proposed methods achieve a good balance between performance and complexity: they outperform existing low-complexity strategies by a large margin in terms of both capacity and BER, and at the same time, are significantly simpler than previous near-optimal iterative algorithms.

In Chapter 4, we shall propose an *MMSE-based optimization approach*. The purpose is to minimize the MSE between the transmitted signals from the source and the received signals at the destination. Two types of constraints on the transmit power of the relays

are considered separately: 1) a weighted sum power constraint, and 2) per-relay power constraints. As opposed to using general-purpose interior-point methods, we exploit the inherent structure of the problems to develop more efficient algorithms. Under the weighted sum power constraint, the optimal solution is expressed as a function of a Lagrangian parameter. By introducing a complex scaling factor at the destination, we derive a closed-form expression for this parameter, thereby avoiding the need to solve an implicit nonlinear equation numerically. Under the per-relay power constraints, we show that the optimal solution is similar to that under the weighted sum power constraint if particular weights are chosen. We then propose an iterative power balancing algorithm to compute these weights. In addition, under both types of constraints, we investigate the joint design of a MIMO equalizer at the destination and the relaying matrices, using block coordinate descent or steepest descent iterations. The above optimal designs do not require global CSI availability: each relay only needs to know its own backward and forward channel, together with a little extra information. BER simulation results demonstrate that all the proposed designs, under either type of constraints, with or without the equalizer, perform much better than previous methods and the hybrid methods. Our work also provides an interesting insight: under the per-relay power constraints, the optimal strategy sometimes does not use the maximum power at some relays. Forcing equality in the per-relay power constraints would result in loss of optimality. Another interesting point is that, no matter how low the SNR is at a particular relay, this relay does not have to be turned off completely.

A unified framework for adaptive transceiver design

In Chapter 5, we shall propose a unified framework for adaptive transceiver optimization which is applicable to a wide variety of MIMO relay networks. It leads to new and more efficient algorithms with better performance for certain network topologies. This framework also makes it convenient to exploit the above mentioned relationship between successive transmission blocks via inter-block adaptation. First, we formulate a general system model which can accommodate various network topologies by imposing appropriate structural constraints on the source precoder, the relaying matrix and the destination equalizer. Next, we derive the optimal MMSE relaying matrix as a function of the other two matrices, thereby removing this matrix and its associated power constraint from the optimization problem. This is the common step for point-to-point and multiuser systems. Subsequently,

we study optimization of either the precoder or the equalizer under different structural constraints and propose an alternating algorithm for the joint design. In this algorithm, the optimal equalizer from the previous block is chosen as the initial search point for the current block. This inter-block adaptation speeds up convergence and henceforth reduces computational complexity significantly. The proposed framework is further explained and validated numerically within the context of different system configurations. For example, for relay-assisted broadcast channel (BC) with single-antenna users, the proposed framework leads to a new diagonal scaling scheme which provides more flexibility by allowing different users to apply their own amplitude scaling and phase rotation before decoding.

List of publications

The major contributions of this thesis have been published as the following papers in peer-reviewed journals and conference proceedings.

Journal papers

J1 C. Zhao and B. Champagne, "A low-complexity hybrid framework for combining-type non-regenerative MIMO relaying," *Wireless Personal Commun.*, vol. 72, no. 1, pp. 635-652, Sep. 2013

J2 C. Zhao and B. Champagne, "Joint design of multiple non-regenerative MIMO relaying matrices with power constraints," *IEEE Trans. Signal Process.*, vol. 99, no. 19, pp. 4861-4873, Oct. 2013,

J3 C. Zhao and B. Champagne, "Inter-Block Adaptive Transceiver Design for Non-Regenerative MIMO Relay Networks," *to be submitted to IEEE Trans. Wireless Commun.*, 10 pages, 2013

Conference papers

C1 C. Zhao and B. Champagne, "Non-regenerative MIMO relaying strategies – from single to multiple cooperative relays," in *Proc. 2nd Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Suzhou, China, Oct. 2010, 6 pages.

C2 C. Zhao and B. Champagne, "MMSE-based non-regenerative parallel MIMO relaying with simplified receiver," in *Proc. IEEE Global Telecomm. Conf. (GLOBECOM)*, Houston, TX, USA, Dec. 2011, 5 pages.

C3 C. Zhao and B. Champagne, "Linear transceiver design for relay-assisted broadcast systems with diagonal scaling," in *Proc. Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, Vancouver, Canada, May 2013, 5 pages.

1.4 Organization and Notation

The organization of this thesis is as follows. Chapter 2 provides a comprehensive literature survey of previous works on MIMO relaying. The hybrid relaying framework and the MMSE-based transceiver optimization for combining-type 1S-MR-1D systems are developed in Chapter 3 and Chapter 4, respectively. Chapter 5 presents the inter-block adaptive transceiver design, followed by the conclusions in Chapter 6.

The following notations are used throughout this thesis. Italic, boldface lowercase and boldface uppercase letters represent scalars, vectors and matrices; superscripts $\bar{()}$, $()^T$, $()^H$ and $()^\dagger$ denote conjugate, transpose, Hermitian transpose and Moore-Penrose pseudo-inverse, respectively; tr() refers to the trace of a matrix; $\|\cdot\|_2$ ($\|\cdot\|_{\mathrm{F}}$) stands for the Euclidean (or Frobenius) norm of a vector; col() stacks many column vectors into a single vector, vec() stacks the columns of a matrix into a vector and unvec() is its inverse operator; diag() forms a diagonal (or block-diagonal) matrix from multiple scalars (or matrices); $\otimes$ represents the Kronecker product; $\mathbf{I}_n$ is an identity matrix of dimension n; E{} refers to mathematical expectation; $\mathbb{R}$ and $\mathbb{C}$ denote the sets of real and complex numbers; $\mathcal{R}()$ and $\mathcal{N}()$ are the column space and the null space of a matrix; dim() is the dimension of a space; $\succeq$ and $\preceq$ represent positive semidefinite ordering of matrices.

Chapter 2

Literature Survey

The idea of wireless relaying dates back to the three-terminal relay channel model proposed in the 1970s [11, 16, 17]. However, it did not receive much attention until the new millennium, when relay-based heterogeneous cellular networks were becoming indispensable. In this chapter, we attempt to provide a comprehensive survey of previous works on non-regenerative MIMO relaying. For convenience and clarity, we first discuss in Sec. 2.1 the guidelines of classifying these works, together with background information that can be helpful to understanding the following literature review. 1S-1R-1D, 1S-MR-1D and multiuser MIMO relaying are then reviewed in Sec. 2.2, 2.3 and 2.4, respectively.

2.1 General Classifications

Thanks to its great potential, wireless relaying has been attracting strong interest in the recent literature. Researchers are looking at almost all the facets — from theoretical limits to practical implementations. Existing publications target dissimilar network configurations, make different assumptions about systems and channels, choose distinct performance measures and employ various mathematical methodologies. This richness makes it difficult to present a cohesive and insightful survey. Therefore, it is important to first present some background information and guidelines that can help to classify these works and build connections among them.

The literature survey is organized by network topology on the top level. As explained in Sec. 1.2, network topology refers to the numbers of sources, relays and destinations involved in the communication process. In conventional terminology, network topology can

also mean point-to-point or multiuser networks, depending on the numbers of sources and destinations. More specifically,

- A point-to-point link connects one source and one destination with the help of a single relay (1S-1R-1D).
- A point-to-point link can also be assisted by multiple parallel relays (1S-MR-1D), *cf.* Fig. 1.3.
- Relay-assisted multiuser networks come in different forms:
  - In a multiple access channel (MAC), multiple source users are simultaneously sending information to a single destination (MS-1R-1D and MS-MR-1D).
  - In a broadcast channel (BC), a single source transmits to multiple users through the aid of a single or multiple relays (1S-1R-MD, 1S-MR-MD).
  - In an interference channel (IC), multiple sources communicates with multiple destinations simultaneously (MS-1R-1D and MS-MR-MD).

Since the above network topologies raise different challenges for system design, all of them have been studied to some extent in the literature and are reviewed in Sec. 2.2, 2.3 and 2.4, respectively. It is worth mentioning that additional topologies exist when the direct source-destination link is not negligible or the communication is assisted by multiple relays in a multi-hop fashion. These networks were also investigated but in this thesis, we consider only two-hop relaying without direct link.

In the literature survey, we may also refer to other aspects of system configurations such as the number of antennas and transmission bandwidth. A single-input single-output (SISO) relay system deploys a single antenna at each of its nodes, including the sources, relays and destinations. In a MIMO relay system, each node is equipped with multiple antennas. For simplicity, we also include in this category those systems with single-antenna sources or destinations. Source precoding, non-regenerative relaying and destination equalization are all represented by complex scalars in SISO systems, or by matrices in MIMO systems. Consequently, transceiver design for the latter is more challenging. Our focus is on these MIMO relay systems and we will refer to selected papers on SISO only if they are closely related to the corresponding MIMO problems.

In terms of bandwidth, communication systems can be narrowband or broadband, corresponding to frequency non-selective and frequency selective fading channels, respectively. In the latter scenario, the channels are represented by linear time-varying channel response functions. Multi-carrier communication architectures, such as orthogonal frequency division multiplexing (OFDM), are commonly used to split the wide spectrum into multiple narrow frequency bands, which makes it possible to view the channel experienced by each carrier as frequency non-selective. In this sense, transceiver design for narrowband systems can be readily applied to broadband systems by treating each subcarrier independently. A carrier-cooperative approach further allocates power between these subcarriers and performs slightly better [15, 18].

Within each individual network topology, we can organize the literatures in terms of research goals and methodologies. Various aspects of MIMO relaying have been explored to some extent, including theoretical limits, transceiver design, practical implementations and supporting mechanisms such as channel estimation, CSI feedback and synchronization. Here, we briefly explain these facets to help the readers to understand the following sections more easily.

(1) *Theoretical limits.* For any communication system, it is always important to determine the capacity of the underlying channel. This theoretical limit is the maximum rate at which data can be transmitted with asymptotically low probability of error. The introduction of relay stations makes it much more difficult to obtain the maximum transmission rate because the relay(s) can operate in many different ways. As a result, most theoretical limits are derived by first assuming a certain relaying strategy. Other than channel capacity, one may also be interested in asymptotic measures, such the diversity order, or capacity scaling law when the number of relays becomes very large.

(2) *Transceiver design.* In MIMO relay systems, after choosing specific signal constellations and coding schemes for the multiple symbol streams, it is then necessary to optimize the transceiver architecture to improve the link quality for each stream. A linear transceiver for MIMO relay systems includes precoders, relaying matrices and equalizers. Nonlinear transceiver architectures, such as decision feedback equalizer (DFE), Thomlinson-Harashima precoding (THP) and dirty paper coding (DPC), provide additional performance gain in single-hop and relay-assisted MIMO systems

at the price of higher complexity. Transceiver optimization are usually explored from two different approaches. One approach optimizes a certain performance measure under appropriate power constraints imposed on the sources and relays. This objective function can be either an information-theoretic quantity such as the mutual information, or a statistical measure such as the MSE, SINR or BER. The other approach, *power control*, minimizes a power-related function subject to QoS requirements that are usually expressed in terms of SINRs.

(3) *Practical implementations.* Full channel knowledge and centralized processing are the most common assumptions for transceiver design. However, they are never perfect in practice and some works explicitly take these aspects into consideration.

a) *Robust design against channel uncertainty.* CSI knowledge can be obtained only through channel estimation at the receivers, and through channel reciprocity (in a time-duplex division system) or channel feedback at the transmitters. Consequently, estimation errors, feedback quantization, and fast fading often cause model mismatch and performance degradation. Designing practical transceiver architectures that are robust to channel uncertainty is therefore important in fulfilling the theoretical potential of MIMO relaying. Robustness against channel mismatch can be interpreted and pursued from two different perspectives. The Bayesian approach assumes knowledge of the probability distribution function of the CSI errors, and optimizes the mathematical expectation of a proper performance measure. The min-max approach, instead, optimizes the worst-cast performance among all possible channel realizations. This approach is more useful when dependability is of significant importance.

b) *Transceiver design with channel distribution information (CDI).* Under fast fading channels, channel estimation may be very difficult and optimal transceiver design is done based on CDI such as channel mean and covariance.

c) *Distributed implementation.* A centralized transceiver design is carried out at a fusion center which usually leads to some communication overhead because of the resulting information exchange between the source, relay and destination nodes. Distributed implementations emphasize the use of local CSI, and possibly additional shared information (but not full CSI), with the purpose of trading

marginal performance loss for lower implementation complexity.

(4) *Supporting mechanisms.* Channel estimation, limited feedback, codebook design and synchronization are indispensable to a MIMO relay system. These components have also been widely studied in the literature.

Now that necessary background information and guidelines have been provided, we present a survey of the literature on non-regenerative MIMO relaying in the following sections. Specifically, our presentation is organized into three parts that review recent works on point-to-point (1S-1R-1D), combining-type (1S-MR-1D) and multiuser MIMO relaying, respectively.

2.2 Point-to-Point Communications: 1S-1R-1D

The simplest network topology is 1S-1R-1D, in which a point-to-point communication link is assisted by a single relay. Several theoretical upper and lower bounds were derived in [19] on the capacity of a MIMO relay channel with direct link. More often, the direct source-to-destination link is hindered by high level of attenuation and can be neglected. In this case, the overall channel capacity is upper bounded by that of the backward channel and that of the forward channel. The fact that the relay(s) can operate in many different ways makes it extremely difficult to derive the this channel capacity, which is still an open problem. Most existing information-theoretic works assume a certain relay processing scheme. For example in [20], ergodic capacity is studied when the relay uses the simplistic AF strategy which merely amplifies the signals.

Linear transceiver design for 1S-1R-1D systems is highly related to MIMO transceiver design. The major difference lies in the existence of *two* channel matrices and the amplification of the noise term from the relay. Existing works belong to two different categories. The first class of approaches assume special transceiver structures and optimize the underlying parameters. These structures are usually borrowed from the optimal transceiver architectures of point-to-point MIMO systems, such as ZF [21], MMSE [22] and SVD [23–25]. For example, as shown in Fig. 2.1, the SVD-based relaying matrix diagonalizes the middle part of the overall channel. With proper precoder and equalizer, the transceiver design problems are scalarized so that they can be readily solved using the powerful tools of convex

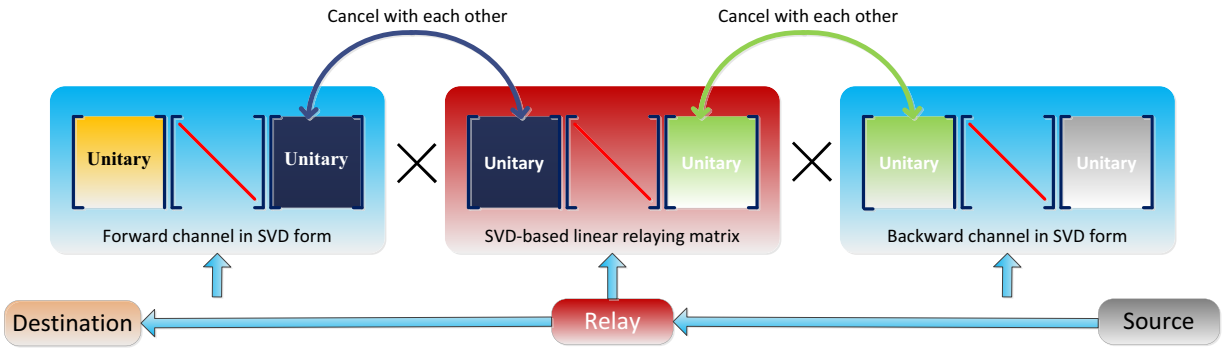

Fig. 2.1 SVD-based relaying structure. From right to left are the SVDs of the backward channel, the relaying matrix and the forward channel.

optimization. This channel diagonalization scheme can be further generalized to multi-hop relaying [26] and broadband multicarrier case [15, 25].

The other category of works do not assume any predefined structure. Instead, they formulate and solve appropriate optimization problems based on matrix theory [27]. Interestingly, for 1S-1R-1D, the previously mentioned SVD relaying structure is optimal in terms of a wide variety of performance measures, including mutual information [28–30], MSE [31] and pairwise error probability [23]. In fact, most of these performance measures are related with each other through an MSE matrix. This matrix is defined as the covariance matrix of the estimation error between the equalizer output and the precoder input [15, 18]. More specifically, a majority of objective functions are essentially additively Schur-concave or Schur-convex functions of the diagonal entries of the MSE matrix, so that they can be considered in a unified framework [15]. For example, the mutual information, the product of the substream MSEs, the sum of the substream SINRs and the product of the SINRs are Schur-concave functions; the sum of the MSEs, the maximum MSE, the harmonic mean of the SINRs and the minimum of the SNRs are Schur-convex. The optimal relaying matrices for both categories are in the SVD form and the difference lies in the optimal source precoder. For Schur-concave objective functions, the optimal precoder is the product of the conjugate transpose of the rightmost unitary matrix in Fig. 2.1 and a diagonal power allocation matrix. This leads to scalarization of the problem and the remaining power allocation problem is discussed in [15, 23–25, 29–31] for various performance criteria. For Schur-convex functions, the optimal precoder has to preprocess the symbol vector using an additional unitary matrix. This matrix is chosen such that the diagonal entries of the MSE

matrix are all equal. When the number of streams is a power of two, the discrete Fourier transform matrix or a Walsh-Hadamard matrix can be used [15]. It is worth mentioning that the optimal transceiver for the power control problem has exactly the same structure. The only difference is that the unitary matrix is chosen such that the diagonal entries of the MSE matrix are equal to the QoS requirements for the substreams [32, 33].

In single-hop MIMO systems, nonlinear architectures such as DFE and THP are known to provide additional performance gain with respect to linear transceivers. Similar conclusions hold for two-hop MIMO relay systems [34, 35]. DFE is based on the multiplicative majorization theory and the equal-diagonal QR decomposition method from [36].

As discussed before, transceiver optimization requires explicit knowledge of the wireless channels which often comes with estimation errors. MMSE-based robust design against CSI mismatch in the statistical Bayesian sense is addressed in [37] for flat-fading channels, and in [38] for an OFDM system with frequency-selective channels. These results are extended to additively Schur-concave and Schur-convex functions in [39]. Simulation results in these works have shown that robust designs provide better performance than non-robust designs. For fast-fading channels, channel estimation may even be impossible and transceiver design must turn to the first-order and second-order channel statistics, namely, the CDI. This aspect was investigated in [40–43] when either the backward or the forward channel is under fast fading.

Channel acquisition at the receivers of the relay and the destination can be done using the same methods as for point-to-point MIMO channels [44]. Algorithms that are tailored for MIMO relay systems are also studied in previous works. For example, both the backward and the forward channels can be estimated at the destination using the pilot-based schemes proposed in [45, 46] and the parallel factor analysis method from [47]. CSI at the transmitters of the source and the relay can be obtained via channel reciprocity for time-division duplex (TDD) systems, and via feedback for both TDD and frequency-division duplex (FDD) systems. The design of appropriate codebooks for limited feedback is considered in [48].

To sum up, 1S-1R-1D systems are well understood from different perspectives. However, as discussed in the following sections, this is not the case with combining-type 1S-MR-1D systems and multiuser relay systems.

2.3 Combining-Type Relaying: 1S-MR-1D

The optimal schemes for 1S-1R-1D can be extended to 1S-MR-1D systems through the use of a selection-type operation, whereby the source signals are forwarded through the single relay that offers the best link quality [49, 50]. A more meaningful approach in the 1S-MR-1D case is to use all the relays (or a subset of them) simultaneously in order to benefit from distributed array gain [13, 51]. This multi-relay configuration leads to a logarithmic capacity scaling with the number of relays [51, 52]. The ergodic capacity with the simplistic amplify-and-forward (SAF) relaying strategy is studied in [20].

Transceiver design is of paramount importance in fulfilling the potential of combining-type 1S-MR-1D systems. In particular, the SVD-based relaying for 1S-1R-1D does not readily extend to this case because, due to the physically separated nature of the multiple relays, their combined transformation matrix inherits a block-diagonal form. The essential feature of an appropriate relaying strategy is that the signals from different relays should be coherently combined at the destination [13, 51]. This coherent combining can easily be guaranteed for single-antenna relays because each relay simply compensates for the phases of the backward and forward *scalar* channels [10, 53, 54]. For multiantenna relays, however, the channels are represented by matrices, which complicates the transceiver design significantly.

For the purpose of coherent combining, some heuristic strategies have been proposed that "borrow" ideas from MIMO transceiver design, including MF [13], ZF [55], linear MMSE [13] and QR decomposition [56–58]. These heuristic methods, although structurally constrained, has been shown to perform much better than SAF which only amplifies the signals.

A more comprehensive approach is to formulate the collaborative design of the relaying matrices as an optimization problem with power constraints [59–63]. The objective can be to maximize the achievable rate [59] or to minimize the MSE [60, 61]. However, most works either assume special structures on the matrices [59], or rely on numerical algorithms such as gradient descent [60], bisection [61] and iterative schemes [60–62] to obtain the optimal solution. For these methods, this lack of closed-form expressions generally leads to high implementation complexity, which in turn limits their potential feasibility.

For completeness, it is worth mentioning that explicit formulas were derived in [64–66] when the power constraints are enforced on the signals *received* at the destination. However,

these results do not carry over to the case when the constraints are imposed on the *transmit* power of the relays [63]. In fact, this thesis was largely motivated by the lack of efficient transceiver design algorithms for combining-type 1S-MR-1D systems.

2.4 Multiuser MIMO Relaying

Most research results on multiuser MIMO relaying have their counterparts in multiuser MIMO. In general, these works either consider information-theoretic aspects such as capacity region, or focus on signal processing problems such as transceiver design, power allocation and robust design. In the following paragraphs, we summarize previous works on the three classes of relay-assisted networks, namely, MAC, BC and IC.

MAC typically models the uplink of a cellular communication system, with a fixed relay between the mobile users and the base station (BS). If the mobile users all have a single antenna, the system is equivalent to a 1S-1R-1D system with a multi-antenna source [31,67]. In this case, the sources simply transmit their corresponding maximum allowable power, and joint design of the relay forwarding matrix and the destination equalizer is done in the same way as in 1S-1R-1D systems [29,31]. If the users are equipped with multiple antennas, finding the source precoders, the relaying matrix and the destination equalizer to maximize a system performance is generally a difficult task. For example, iterative algorithms are proposed in [68] to solve the rate maximization problem with power constraints, and the power minimization problem subject to rate constraints. The MMSE-based design was studied in [69, 70], which is also based on iterative algorithms.

BC is usually found in the downlink of a cellular system. For single-antenna users, several lower bounds on the achievable sum rate are established in [71] by employing ZF DPC and imposing several different structures on the source precoder and the relaying matrix. These performance bounds have motivated the development of implementable transceiver designs. For example, a nonlinear transceiver with THP at the base station, linear processing at the relay and adaptive modulation is proposed in [71]. A linear transceiver architecture is optimized to maximize the achievable sum capacity under fixed transmit power constraints in [72]. This nonconvex optimization problem is solved by being converted to standard convex quadratic programming problems in an iterative manner. Specifically, this is done by approximating the non-convex functions locally by their low-order counterparts with reasonable accuracy, leading to convex sub-problems. An interesting

discovery is that the optimal precoding and relaying matrices almost diagonalize the compound channel at high SNRs, which motivates a simplified transceiver design based on channel diagonalization [72]. An MMSE-based design is considered in [67] based on alternating algorithms. Ref. [73] takes a different *power control* perspective by minimizing the weighted sum-power consumption of the base station and the relay to support a set of minimum-SINR requirements at user terminals. Two practical solutions are proposed: SINR balancing via second-order cone programming (SOCP), and channel-inversion based on the SVD structures and geometric programming (GP).

Multiple antennas at the mobile users bring further gains in system throughput for BC [74], though at the price of more challenging transceiver design. Most existing works either assume special structures for the design matrices [75, 76], or iterate through the precoders, the relaying matrices and the equalizers multiple times [68, 70, 77–81]. An example of the special processing structures is block diagonalization (BD), which is a generalization of ZF. In BD, the signals are transmitted in appropriate subspaces such that the co-channel interferences are zero at the mobile users [82]. In [75], BD is used at the relay without a linear precoder at the source; while in [76], the two hops are designed separately: MMSE transceiver is used for the first hop and BD or geometrical mean decomposition (GMD) are used for the second hop. The iterative (or alternating) methods are essentially block coordinate descent: they optimize each of the design matrices one at a time until convergence. This idea is used to solve the MMSE problems in [70, 77–79], the sum rate problems in [68,80], the power control problems in [68,81] and the limited feedback problems in [83, 84].

In IC, multiple source-destination pairs are assisted by a single or multiple parallel relays. Previous works mostly focus on the power control problems. The optimal relaying matrices are generally obtained by applying semidefinite relaxation and solving SOCP or semidefinite programming (SDP) problems. For single-antenna users, the sum power minimization problem subject to QoS constraints is considered in [85,86] and the corresponding robust design is studied in [87, 88]. A different formulation minimizes the total interference and noise power subject to distortionless constraint on the desired signals [86, 89]. An adaptive and decentralized version based on Kalman filter is proposed in [90]. For multi-antenna users, the power control problem is studied in [91] using a two-tier iterative algorithm which itself solves a sequence of SDP subproblems. Due to the high complexity of IC with relays, it may be more meaningful to explore structured approaches such as

BD [57] and interference alignment [92].

Unlike 1S-1R-1D, transceiver design for multiuser MIMO relay systems usually leads to complex iterative algorithms that may require solving SOCP or SDP subproblems. Fortunately, some of these algorithms may be endowed with good initial search points when the neighbouring transmission blocks are considered together, as we pointed out in Chapter 1. This should be kept in mind in the development and evaluation of new iterative algorithms.

Chapter 3

A Low-Complexity Hybrid Framework for Combining-Type Relaying

Combining-type 1S-MR-1D systems can benefit from distributed array gain if the signals retransmitted from different relays are superimposed coherently at the destination. In this chapter[1], we propose a low-complexity hybrid relaying framework in which the non-regenerative MIMO relaying matrix at each relay is generated by cascading two parametric substructures. This hybrid framework allows for different combinations of substructure candidates and further optimizations of the underlying parameters, thereby achieving a good balance between performance and complexity. The introduction in Sec. 3.1 explains the background and motivation. Sec. 3.2 describes the system model and the underlying assumptions. Sec. 3.3 presents the new hybrid framework along with the proposed non-cooperative and cooperative matrix substructures. Suitable performance criteria and methodology for choosing their design parameters are developed in Sec. 3.4. The numerical results and further discussions are included in Sec. 3.5, followed by a brief summary in Sec. 3.6.

[1]Parts of Chapter 3 have been presented at the 2010 International Conference on Wireless Communications and Signal Processing (WCSP) in Suzhou, Jiangsu, China [93], and published in Wireless Personal Communications [94].

3.1 Introduction

As discussed in Sec. 2.3, for the 1S-1R-1D configuration, the optimal MIMO relaying matrix is well established in terms of various performance criteria [15, 22, 28, 29, 31, 95, 96]. Interestingly, a majority of these criteria lead to a common SVD structure, which can scalarize the problems so that they can be readily solved using convex optimization. These optimal schemes can be extended to 1S-MR-1D systems through the use of a selection-type operation, whereby the source signals are forwarded through the single relay that offers the best link quality [49].

Another approach in the 1S-MR-1D case is to use all the relays (or a subset of them) simultaneously in order to benefit from distributed array gain [13,51]. However, the resulting problem of designing optimal transformation matrices with constraints on the transmit power of the relays remains largely unsolved. In particular, the SVD approach does not readily extend to this case since, due to the physically separated nature of the multiple relays, their combined transformation matrix inherits a block-diagonal form. By imposing the power constraint on the received signals at the destination, instead of the transmitted signals from the relays, one can circumvent this difficulty [64, 65]. However, this cannot guarantee any optimality under the original transmit power constraints. Other existing optimal designs either consider only a total power constraint across the relays [97], or employ iterative approaches with relatively high complexity [60, 98]. As an alternative to the optimization approach, some heuristic strategies have been proposed which "borrow" ideas from MIMO transceiver design, including MF, ZF, linear MMSE [13, 64] and QR decomposition [56]. These methods achieve the distributed array gain and perform well in 1S-MR-1D systems [93].

In theory, the relaying matrices should be chosen so that the retransmitted signals combine coherently at the destination. To this end, we introduce a low-complexity hybrid framework in which the transformation matrix of each relay is obtained by cascading two substructures or factors, akin to an equalizer for the backward channel and a precoder for the forward channel. For each of these two substructures, we propose two different one-dimensional parametric families whose members serve as candiates. The first family, *non-cooperative* by nature, depends only on the backward or forward channel corresponding to the same relay. This family includes ZF, linear MMSE and MF as special cases [13]. The second (*cooperative*) parametric family, inspired by [64, 99], also makes use of infor-

mation derived from the channels of other relays. This hybrid framework allows for the classification and comparison of all possible combinations of these substructure, including several previously investigated methods and their generalizations.

Within this hybrid relaying framework, the design parameters of the matrix factors can be further optimized. This can be done on-line after each update of the channel matrices, or off-line based on *a priori* knowledge of channel statistics. In the latter case, the optimum parameters can be well approximated by linear functions of the SNRs, which reduces the implementation complexity significantly. Through simulations, we show that the capacity of selected hybrid schemes (with optimized parameters) comes within 1bits/s/Hz of the upper bound achieved by the nearly capacity-optimal iterative method in [60]. In the mid-to-high SNR range, the BER performance of one hybrid method even exceeds that of the MSE-optimal iterative method. In summary, the proposed hybrid methods achieve a good balance between performance and complexity: they outperform existing low-complexity strategies by a large margin, and at the same time, are significantly simpler than previous near-optimal iterative algorithms.

3.2 System Model

Fig. 3.1 illustrates a 1S-MR-1D MIMO relaying system in which the source forwards its message to the destination through M parallel relays. The source, destination and individual relays are equipped with N_S, N_D and N_R antennas, respectively, where we assume that $N_S = N_D$.[2] The relays work in a half-duplex mode: their antennas are used for either transmitting or receiving purposes during different time slots. We neglect the presence of the direct source-to-destination link which is typically hindered by high levels of attenuation.

We assume that the wireless channels undergo flat (frequency non-selective) block fading [2]. For now, CSI is assumed to be available globally. After introducing the structures of the relaying matrices, we will be able to discuss in detail how much information is needed at each node. In this work, we assume perfect synchronization between the source, relay and destination nodes. Channel estimation and timing/frequency synchronization are important topics in their own rights, but fall outside the scope of this work. For more details, we refer the reader to [46, 100, 101] and the references therein.

[2]For simplicity, each relay is equipped with the same number of antennas; however, generalization to different numbers of antennas at the relays, i.e. $N_{R,k}$, is straightforward.

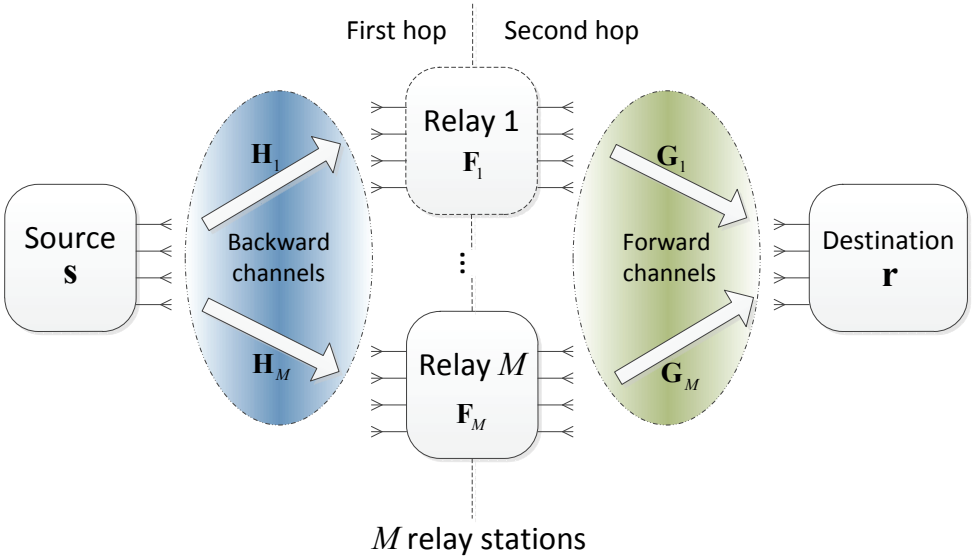

Fig. 3.1 A point-to-point MIMO relaying system

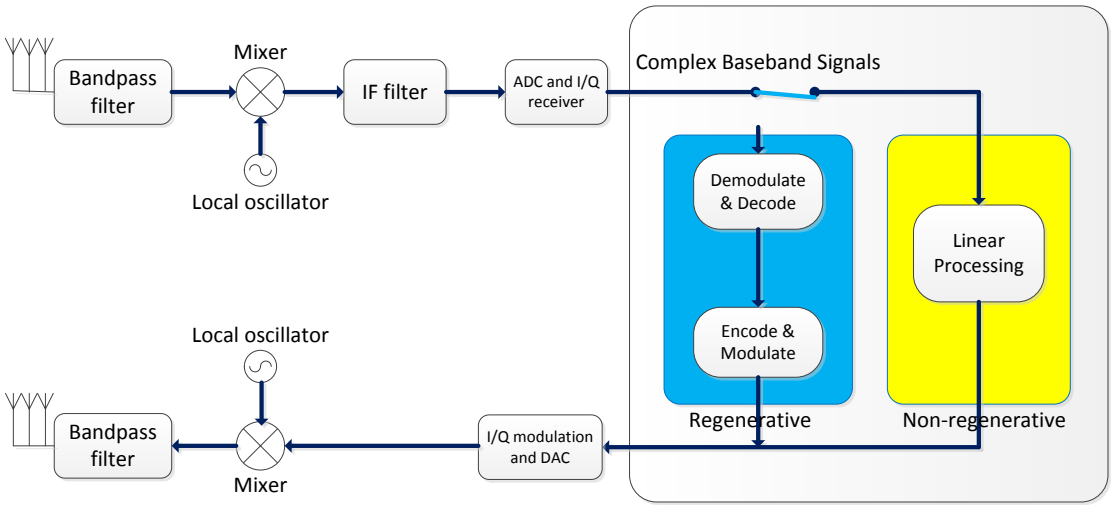

Fig. 3.2 Complex baseband representation of relay processing

The signals, noises and channels are all modeled in terms of their equivalent discrete-time complex baseband representations. That is, standard operations of demodulation, down-conversion, filtering and A/D conversion are assumed at the transmitters of the sources and the relays, with dual operations applied at the receivers of the relays and the

destinations, as shown in Fig. 3.2.

The received signal vector $\mathbf{x}_k \in \mathbb{C}^{N_R \times 1}$ at the k-th relay can be expressed as

$$\mathbf{x}_k = \mathbf{H}_k \mathbf{s} + \mathbf{w}_k, \quad k = 1, \ldots, M \tag{3.1}$$

where $\mathbf{s} \in \mathbb{C}^{N_S \times 1}$ is the source symbol vector comprised of multiple independent streams, $\mathbf{H}_k \in \mathbb{C}^{N_R \times N_S}$ is the backward channel matrix between the source and relay k, and $\mathbf{w}_k \in \mathbb{C}^{N_R \times 1}$ is an additive noise term. The signal and noise terms, $\mathbf{s}$ and $\{\mathbf{w}_k\}$ for $k = 1, \ldots, M$, are modeled as independent, circularly symmetric complex Gaussian random vectors with zero mean and covariance matrices $\mathbf{R}_\mathbf{s} = \mathrm{E}\{\mathbf{s}\mathbf{s}^H\} = \sigma_s^2 \mathbf{I}$ and $\mathbf{R}_{\mathbf{w}_k} = \mathrm{E}\{\mathbf{w}_k \mathbf{w}_k^H\} = \sigma_w^2 \mathbf{I}$, respectively, where σ_s^2 is the average transmit power per antenna at the source and σ_w^2 is the average noise power induced at the individual relay antennas.

The k-th relay multiplies its received noisy signal $\mathbf{x}_k$ by a linear processing matrix $\mathbf{F}_k \in \mathbb{C}^{N_R \times N_R}$ to obtain the retransmitted signal

$$\mathbf{y}_k = \mathbf{F}_k \mathbf{x}_k. \tag{3.2}$$

The received signal vector at the destination, denoted by $\mathbf{r} \in \mathbb{C}^{N_D \times 1}$, takes the form of [3]

$$\mathbf{r} = \sum_{k=1}^{M} \mathbf{G}_k \mathbf{F}_k \mathbf{H}_k \mathbf{s} + \sum_{k=1}^{M} \mathbf{G}_k \mathbf{F}_k \mathbf{w}_k + \mathbf{n}, \tag{3.3}$$

where $\mathbf{G}_k \in \mathbb{C}^{N_D \times N_R}$ is the forward channel matrix from relay k to the destination and $\mathbf{n} \in \mathbb{C}^{N_D \times 1}$ is the noise term induced at the destination receiver. This noise term is assumed independent from $\mathbf{s}$ and $\{\mathbf{w}_k\}$, and modeled as a circularly symmetric complex Gaussian random vector with zero mean and covariance matrix $\mathbf{R}_\mathbf{n} = \mathrm{E}\{\mathbf{n}\mathbf{n}^H\} = \sigma_n^2 \mathbf{I}$, where σ_n^2 is the average noise power received at the individual destination antennas. Eq. (3.3) can also be expressed in a "block-diagonal" form as

$$\mathbf{r} = \mathbf{G}\mathbf{F}\mathbf{H}\mathbf{s} + \mathbf{G}\mathbf{F}\mathbf{w} + \mathbf{n}, \tag{3.4}$$

where we have defined $\mathbf{G} = [\mathbf{G}_1, \ldots, \mathbf{G}_M]$, $\mathbf{H} = [\mathbf{H}_1^T, \ldots, \mathbf{H}_M^T]^T$, $\mathbf{w} = [\mathbf{w}_1^T, \ldots, \mathbf{w}_M^T]^T$ and $\mathbf{F}$ is a block-diagonal matrix with $\mathbf{F}_1, \ldots, \mathbf{F}_M$ as main diagonal blocks. When $M = 1$, this

[3] The reader may also refer to Fig. 4.1.

signal model reduces to the 1S-1R-1D case.

For convenience, we introduce two important SNR parameters. The first SNR ρ_1 describes the link quality of the backward channels and is defined as the ratio of the average transmit power per source antenna to the noise power per relay antenna, i.e., $\rho_1 = {\sigma_s}^2/\sigma_w^2$. The second SNR parameter ρ_2 characterizes the forward channels. Let the total transmit power of the relays be $\mathcal{P}$ and ρ_2 is defined as the ratio of average transmit power per relay antenna to the power of the noise induced at the individual destination antennas, i.e., $\rho_2 = \mathcal{P}/(MN_R\sigma_n^2)$. Note that $\mathcal{P}$ is consumed by the relays to transmit both the desired signal component $\mathbf{s}$ and the additive relay noise terms $\{\mathbf{w}_k\}$. We emphasize that if one of these two SNR parameters is fixed, the system performance is upper bounded due to the corresponding noise term, even if the other SNR goes to infinity.

Note that the above signal model is applicable to a much broader scope than multi-antenna 1S-MR-1D systems. For example, since broadband channels for single-antenna systems can also be represented by matrices, the relaying framework in this chapter applies immediately to broadband single-antenna 1S-MR-1D relaying systems.

3.3 The Unified Hybrid Framework

The focus of this chapter is to design the relay matrices $\{\mathbf{F}_k\}$ for 1S-MR-1D systems, based on the knowledge of the instantaneous channel matrices. One immediate option is to solve for matrices $\mathbf{F}_k$ that collaboratively optimize a suitable performance criterion. However, the block-diagonal matrix $\mathbf{F}$ in (3.4) complicates this problem significantly. Instead, we propose a sub-optimal, yet highly flexible hybrid framework as explained below.

One may contemplate the process of designing the relaying matrices $\mathbf{F}_k$ in (3.2) as that of selecting the equivalent channel

$$\mathbf{GFH} = \sum_{k=1}^{M} \mathbf{G}_k\mathbf{F}_k\mathbf{H}_k, \tag{3.5}$$

with the purpose of maximizing the power of the received signal vector $\mathbf{GFHs}$, without over-amplifying the noise terms $\mathbf{w}_k$ in (3.3). Intuitively, this requires a coherent signal combining of the M parallel transmissions at the destination, i.e., the matrix terms on the right-hand side of (3.5) are superimposed constructively, thereby leading to an M-fold

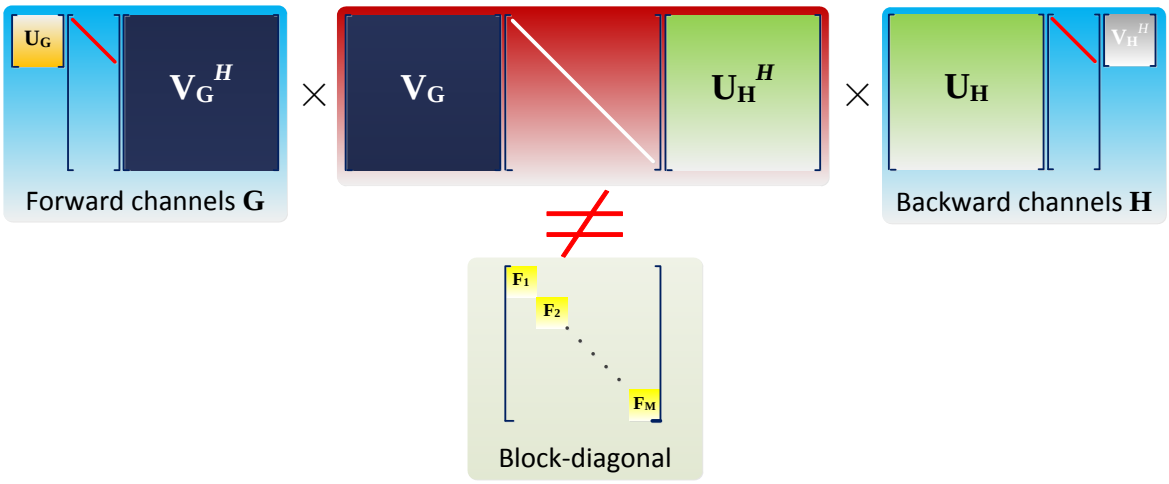

Fig. 3.3 SVD-based relaying does not work for 1S-MR-1D because $\mathbf{F}$ has to be block-diagonal.

distributed array gain.

Although the SVD-based canonical structure is the optimal relaying matrix form for 1S-1R-1D systems under a wide variety of criteria [15,29,31], its optimality cannot be extended to 1S-MR-1D relaying systems. On the one hand, one might compute $\mathbf{F} = \mathbf{V}_G \mathbf{\Lambda} \mathbf{U}_H^H$ from the SVD's of the stacked channel matrices $\mathbf{G} = \mathbf{U}_G \mathbf{\Sigma}_G \mathbf{V}_G^{\mathrm{H}}$ and $\mathbf{H} = \mathbf{U}_H \mathbf{\Sigma}_H \mathbf{V}_H^{\mathrm{H}}$ in (3.4), but this approach would violate the structural constraint on $\mathbf{F}$ as a block-diagonal matrix, as shown in Fig. 3.3. On the other hand, one could form the relaying matrix for the k-th relay, $\mathbf{F}_k = \mathbf{V}_{2k} \mathbf{\Lambda}_k \mathbf{U}_{1k}^H$, from the SVD's of the corresponding backward and forward channel matrices, namely, $\mathbf{H}_k = \mathbf{U}_{1k} \mathbf{\Sigma}_{1k} \mathbf{V}_{1k}^H$ and $\mathbf{G}_k = \mathbf{U}_{2k} \mathbf{\Sigma}_{2k} \mathbf{V}_{2k}^H$, which would result into the following equivalent channel

$$\mathbf{GFH} = \sum_{k=1}^{M} \mathbf{U}_{2k} \mathbf{\Sigma}_{2k} \mathbf{\Lambda}_k \mathbf{\Sigma}_{1k} \mathbf{V}_{1k}^H. \tag{3.6}$$

The matrices $\mathbf{V}_{1k}$ tend to be statistically independent for different k (with a similar argument holding for the $\mathbf{U}_{2k}$), due to the spatially distributed nature of the wireless relays. Consequently, the signal components transmitted from different relays would then add in a random and non-constructive way at the destination.

Motivated by this interpretation, we propose a unified hybrid framework in which the individual relaying matrices $\mathbf{F}_k$ are obtained by cascading two substructures (or factors),

$\mathbf{A}_k \in \mathbb{C}^{N_S \times N_R}$ and $\mathbf{B}_k \in \mathbb{C}^{N_R \times N_S}$, as follows:

$$\mathbf{F}_k = \eta_k \mathbf{B}_k \mathbf{A}_k. \tag{3.7}$$

In light of (3.5), matrix $\mathbf{A}_k$ equalizes the kth backward MIMO channel $\mathbf{H}_k$, generating N_S summary statistics, each of which is a signal stream impaired by noise and also interferences from other streams. Matrix $\mathbf{B}_k$ then serves as the MIMO precoder for the kth forward channel $\mathbf{G}_k$, pre-canceling interstream interferences before transmitting these summary statistics through the forward channels. Finally, η_k is a positive scaling parameter introduced to satisfy the transmit power constraints

$$\mathrm{E}\{\|\mathbf{y}_k\|^2\} = \mathrm{tr}\left(\mathbf{F}_k \mathbf{R}_{\mathbf{x}_k} \mathbf{F}_k^H\right) = P_k, \ \forall 1 \leq k \leq M, \tag{3.8}$$

where $\mathbf{R}_{\mathbf{x}_k} = \mathrm{E}\{\mathbf{x}_k \mathbf{x}_k^H\} = \sigma_s^2 \mathbf{H}_k \mathbf{H}_k^H + \sigma_w^2 \mathbf{I}$. Henceforth, the scaling factor η_k in (3.7) satisfies

$$\eta_k = \sqrt{\frac{P_k}{\mathrm{tr}(\mathbf{B}_k \mathbf{A}_k \mathbf{R}_{\mathbf{x}_k} \mathbf{A}_k^H \mathbf{B}_k^H)}}\ . \tag{3.9}$$

Under the above framework, the relaying strategies can be either non-cooperative or cooperative. For the non-cooperative strategies, the relaying matrix for the kth relay, $\mathbf{F}_k$, only depends on its own backward and forward channel matrices, i.e., $\mathbf{H}_k$ and $\mathbf{G}_k$. For the cooperative strategies, at least one of the substructures also relies on some shared information related to the channels of the other relays. That is, $\mathbf{F}_k$ depends not only on $\mathbf{H}_k$ or $\mathbf{G}_k$, but also on a function of the other channel matrices, as explained below.

3.3.1 Non-Cooperative Approach

Here, each one of the substructures $\mathbf{A}_k$ and $\mathbf{B}_k$ is selected from a corresponding one-dimensional parametric family of matrices. That is, we let

$$\mathbf{A}_k^{\mathrm{NC}} = (\lambda_a \mathbf{I} + \mathbf{H}_k^H \mathbf{H}_k)^{-1} \mathbf{H}_k^H, \tag{3.10a}$$

$$\mathbf{B}_k^{\mathrm{NC}} = \mathbf{G}_k^H (\lambda_b \mathbf{I} + \mathbf{G}_k \mathbf{G}_k^H)^{-1}, \tag{3.10b}$$

where λ_a and λ_b are real, nonnegative design parameters. For instance, by choosing λ_a equal to ∞, 0 or $1/\rho_1$, $\mathbf{A}_k^{\mathrm{NC}}$ is proportional to $\mathbf{H}_k^H, \mathbf{H}_k^\dagger$ or $(\mathbf{I}+\rho_1\mathbf{H}_k^H\mathbf{H}_k)^{-1}\mathbf{H}_k^H$, respectively.[4] In turn, these matrices correspond to the MF, ZF and MMSE substructures which were studied in previous works [13, 49, 51]. A similar argumentation can be made about $\mathbf{B}_k^{\mathrm{NC}}$. The superscript NC in (3.10) means that these substructures are non-cooperative by nature, for they are determined only by the local backward or forward channels, $\mathbf{H}_k$ or $\mathbf{G}_k$.

By cascading $\mathbf{A}_k^{\mathrm{NC}}$ and $\mathbf{B}_k^{\mathrm{NC}}$ as in (3.7), we can obtain a non-cooperative hybrid relaying strategy which includes several previous methods as special cases, as summarized in Table 3.1. By varying each one of the design parameters λ_a and λ_b in (3.10) from zero to infinity, we generalize these previously proposed methods to other intermediate situations of interest.

Table 3.1 Special cases of the non-cooperative hybrid framework

λ_a	λ_b	$\mathbf{A}_k$	$\mathbf{B}_k$	Previous methods
0	0	ZF	ZF	ZF relaying [13]
ρ_1^{-1}	ρ_2^{-1}	MMSE	MMSE	Linear MMSE [13, 49]
∞	∞	MF	MF	MF [13, 49, 51]
ρ_1^{-1}	0	MMSE	ZF	2-step MMSE [64, 99]
$\{0, \rho_1^{-1}, \infty\}$	$\{0, \rho_2^{-1}, \infty\}$	ZF/MMSE/MF		Hybrid [93]

3.3.2 Cooperative Approach

Next, we extend the proposed hybrid framework by considering cooperative strategies where the design of the relaying matrices $\mathbf{F}_k$ explicitly takes into account the combining nature of the signal transmission in 1S-MR-1D systems. This is achieved by exploiting some shared information (but not necessarily all the channel matrices). To this end, we propose

[4] As λ_a goes to infinity, $\mathbf{A}_k^{\mathrm{NC}}$ approaches $\lambda_a^{-1}\mathbf{H}_k^H$ asymptotically; when $\lambda_a = 1/\rho_1$, $\mathbf{A}_k^{\mathrm{NC}} = \rho_1(\mathbf{I} + \rho_1\mathbf{H}_k^H\mathbf{H}_k)^{-1}\mathbf{H}_k^H$. In both cases, the resulting scalar factor can be absorbed by η_k in (3.7).

alternative parametric families of matrix transformations for $\mathbf{A}_k$ and $\mathbf{B}_k$:

$$\mathbf{A}_k^{\mathrm{C}} = \left(\lambda_a \mathbf{I} + \sum_{j=1}^{M} \mathbf{H}_j^H \mathbf{H}_j\right)^{-1} \mathbf{H}_k^H, \tag{3.11a}$$

$$\mathbf{B}_k^{\mathrm{C}} = \mathbf{G}_k^H \left(\lambda_b \mathbf{I} + \sum_{j=1}^{M} \mathbf{G}_j \mathbf{G}_j^H\right)^{-1}, \tag{3.11b}$$

where the superscript C stands for "cooperative". Here, the equalizer (3.11a) is inspired by the works in [64,99]. The two-step MMSE method in [64] first specifies a target signal gain η, and then minimizes the MSE between $\mathbf{r}$ and $\eta\,\mathbf{s}$. This leads to the following expression for the relaying matrix at the kth relay:

$$\mathbf{F}_k = \eta \mathbf{G}_k^{\dagger} \left(\mathbf{I} + \rho_1 \sum_{j=1}^{M} \mathbf{H}_j^H \mathbf{H}_j\right)^{-1} \mathbf{H}_k^H. \tag{3.12}$$

which can be viewed as a hybrid relaying matrix with $\mathbf{A}_k = (\mathbf{I} + \rho_1 \sum_{j=1}^{M} \mathbf{H}_j^H \mathbf{H}_j)^{-1} \mathbf{H}_k^H$, and the precoder $\mathbf{B}_k = \mathbf{G}_k^{\dagger}$. Interestingly, the equalizer $\mathbf{A}_k$ in the above relaying matrix $\mathbf{F}_k$ is using some shared information extracted from all the backward channel matrices, through the sum term $\sum_{j=1}^{M} \mathbf{H}_j^H \mathbf{H}_j$. However, with a limited power budget, the solution $\mathbf{G}_k^{\dagger}$ for the precoder matrix is far from optimal: indeed, when the forward channel matrix $\mathbf{G}_k$ is not well-conditioned and the second-hop SNR ρ_2 is low, this ZF-type pseudo-inverse would waste significant power in compensating the effect of $\mathbf{G}_k$. In this regard, we generalize this equalizer into $\mathbf{A}_k^{\mathrm{C}}$ and extend it to the precoder side as well in (3.11b). The sums in (3.11a) and (3.11b) are the information needed to be shared between relays.

More generally, the hybrid relaying matrix $\mathbf{F}_k$ in (3.7) can be formed by combining factors $\mathbf{A}_k$ and $\mathbf{B}_k$ selected from any of the above proposed non-cooperative and cooperative parametric families of matrices. For notational simplicity, we refer to these hybrid strategies as "A-B(λ_a, λ_b)" where for example, NC-C(0, 0) means that the $\mathbf{A}_k$ factor of the relaying matrix $\mathbf{F}_k$ is the non-cooperative substructure $\mathbf{A}_k^{\mathrm{NC}}$ and the $\mathbf{B}_k$ factor is the cooperative substructure $\mathbf{B}_k^{\mathrm{C}}$, with $\lambda_a = \lambda_b = 0$. In this sense, the proposed hybrid framework enables the formal classification of previously investigated methods as well as their generalization by supplementing them with a rich set of alternatives. We note that for simplicity, the same λ_a and λ_b are used for different relays. Choosing different values for each relay would

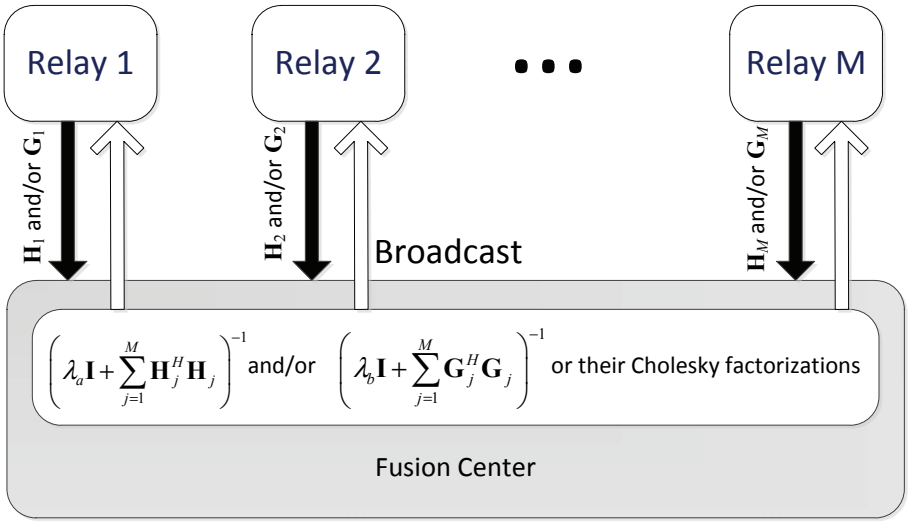

Fig. 3.4 CSI exchange for cooperative relaying strategies The fusion center collects all needed channel information, computes the matrix sum(s) and possibly the Cholesky factorization(s), and then broadcasts the required information.

barely improve system performance but increases complexity significantly when optimizing these parameters.

3.3.3 Implementation Issues

The relaying matrices are computed based on the knowledge of the wireless channels. For the NC-NC strategy, each relay only needs its own backward and forward channel matrices that can be obtained in the same way as in 1S-1R-1D systems [15, 22, 28, 29, 31, 95, 96]. For the cooperative hybrid relaying strategies, it is also essential to share the matrix sums $\sum_{k=1}^{M} \mathbf{H}_k \mathbf{H}_k^H$ and/or $\sum_{k=1}^{M} \mathbf{G}_k^H \mathbf{G}_k$ (but not all the channel matrices) among the relays. These sums can be computed at a fusion center, which may be one of the relays or the destination, and broadcasted to the relays, as shown in Fig. 3.4. In practice, the number of relays M will not be very large, e.g., between 2 and 4. Therefore, compared with non-cooperative NC-NC, the cooperative hybrid strategies can be implemented without much added difficulty, especially when the relays are not far from each other so that dedicated local wireless links or wireline connections are possible.

The procedures for computing $\mathbf{A}_k$, $\mathbf{B}_k$ and η_k are simple and involve only a small number of matrix multiplications and inverses. The resulting complexity is very low, though the

Cholesky/QR factorizations and backward-forward substitution can be used for further simplification [102].

3.4 Optimization of the Parameters

3.4.1 Motivations

λ_a and λ_b can be regarded as "regularization" or "diagonal loading" parameters for the substructures $\mathbf{A}_k$ and $\mathbf{B}_k$: λ_a prevents over-amplification of the noise terms $\mathbf{w}_k$ in (3.1) when equalizing ill-conditioned backward channels; λ_b prevents the transmit power of the relays from being wasted in pre-compensating ill-conditioned forward channels [103, 104]. In the context of a point-to-point MIMO channel, the use of $\lambda_a = \rho_1^{-1}$ and $\lambda_b = \rho_2^{-1}$ in (3.10) leads to the optimal linear MMSE equalizer and precoder, respectively. These are known to offer the best trade-off between noise and interference cancellation, outperforming both MF and ZF over the complete SNR range [2, 105].

Then, for MIMO relaying systems, it is legitimate to ask why it might be more appropriate to choose values other than 0, ρ_i^{-1} $(i = 1, 2)$, or ∞? To begin with, two independent noise sources arise in the signal model. For the first hop, the outputs from the equalizer $\mathbf{A}_k$ are not decoded immediately but need further processing, and thus setting λ_a to $0, \rho_1^{-1}$ or ∞ is not necessarily optimal. For the second hop, the input signals have already been impaired by noise and interferences before being processed by the precoders $\mathbf{B}_k$ and retransmitted, and therefore choosing $\lambda_b = 0, \rho_2^{-1}$ or ∞ is not optimal, either. Furthermore, the combining of signals from multiple relays makes it more complicated to predict the joint effects of λ_a and λ_b on system performance.

Another important concern is that the presence of a linear MIMO equalizer at the destination makes it possible to *exploit* the inter-stream interferences. In contrast to the ZF-type substructures with λ_a or $\lambda_b = 0$, these interferences do not necessarily have to be small or completely eliminated at intermediate steps, such as in the output of the substructure $\mathbf{A}_k$ or in the received signal vector $\mathbf{r}$. Provided that the interfering streams can be efficiently recombined at the destination, the power used by the relays to transmit them can actually contribute to performance improvement.

Therefore, in a relaying scenario, the parameter values 0, ρ_i^{-1} $(i = 1, 2)$, or ∞ are not optimal in general. Our proposed parametric approach provides additional flexibility in

balancing various factors that hinder system performance, and thereby can fully exploit the potential in these seemingly simple substructures.

3.4.2 Performance Measures and Power Constraints

Here, we introduce two classes of performance criteria that can be used to optimize the parameters λ_a and λ_b, as well as to compare the performance of different relaying strategies.

The most fundamental theoretical limit is the channel capacity. In a strict sense, it is the maximum asymptotically achievable rate over *all* possible transceiver schemes and relaying strategies. Here, we abuse this terminology slightly by viewing the AF relaying matrices as parts of the channel. Different relaying schemes result in different equivalent channels between the source and the destination, and we refer to the maximum mutual information between $\mathbf{s}$ and $\mathbf{r}$ as the channel capacity. For deterministic channels, it can be written as

$$C(\mathbf{F}) = \frac{1}{2}\log\det\left(\mathbf{I} + \mathbf{H}_{\text{eq}}\mathbf{R}_s\mathbf{H}_{\text{eq}}^H\right), \tag{3.13}$$

where $\mathbf{H}_{\text{eq}} = (\mathbf{GFR}_w\mathbf{F}^H\mathbf{G}^H + \mathbf{R}_n)^{-1/2}\mathbf{GFH}$ [29]. The factor of $1/2$ in (3.13) is due to the half duplex mode of operation. Under *slow fading*, as assumed throughout this thesis, the system performance is characterized through the outage probability $p_{\text{out}}(R) = \Pr(C(\mathbf{H}_{\text{eq}}, \mathbf{F}) < R)$ and the corresponding outage capacity defined as the supremum

$$C_{\text{out}}(\epsilon) = \sup\left\{R | p_{\text{out}}(R) < \epsilon\right\}. \tag{3.14}$$

Practical systems compromise transmission rate for lower complexity, cost and latency [106]. In this sense, it is also of interest to examine other criteria such as the MSE, the SINR and the average BER. In this chapter, we assume a V-BLAST (Vertical-Bell-Laboratories-Layered-Space-Time) scheme in which the source antennas transmit independent symbol streams with the same average power, and the destination applies a linear MMSE MIMO combiner followed by single-stream decoding [2, p. 333]. Within this framework, the MSE, SINR and theoretical BER of each substream are linked together through the normalized MSE matrix, as defined in [18] and [15] by

$$\mathbf{E} = \left(\mathbf{I} + \rho_1\mathbf{H}^H\mathbf{F}^H\mathbf{G}^H\left(\mathbf{GFF}^H\mathbf{G}^H + \frac{\sigma_n^2}{\sigma_w^2}\mathbf{I}\right)^{-1}\mathbf{GFH}\right)^{-1}. \tag{3.15}$$

Specifically:

(1) The normalized MSE of the kth substream is the kth diagonal entry

$$\mathrm{MSE}_k = \mathbf{E}(k,k) \in (0,1] \; . \tag{3.16}$$

(2) The SINR of the kth substream is a function of its MSE:

$$\mathrm{SINR}_k = \frac{1-\mathrm{MSE}_k}{\mathrm{MSE}_k} \; . \tag{3.17}$$

(3) If the interferences and noise terms are all Gaussian random variables, the symbol-error rate (SER) of the kth substream is upper bounded by a function of SINR_k:

$$p_s(k) = \alpha \; \mathrm{Q}\left(\sqrt{\beta \; \mathrm{SINR}_k}\right) \; , \tag{3.18}$$

where α and β depend on the constellation, and $\mathrm{Q}(x) = (1/\sqrt{2\pi}) \int_x^\infty e^{-y^2/2} \mathrm{d}y$. If the source uses Gray codes in symbol-to-bit mapping, the BER of the k-th substream is $\approx p_s(k)/n$, where 2^n is the constellation size.

Substituting (3.9) and (3.7) into (3.13) or (3.15), the above mentioned performance measures all become continuously differentiable functions of $\boldsymbol{\lambda} = [\lambda_a, \lambda_b]^T$.

3.4.3 Methodology

Assume that we are *minimizing* a general performance measure $f(\boldsymbol{\lambda})$ based on a single instance of the fading channels. In general, setting the gradient to zero, i.e. $\bigtriangledown_{\boldsymbol{\lambda}} f = \mathbf{0}$, does not lead to a closed-form optimal solution. Instead, we can resort to several numerical algorithms that start from an initial point, $\boldsymbol{\lambda}_0$, and search for a locally optimal $\boldsymbol{\lambda}_{\mathrm{opt}} = [\lambda_a^o, \lambda_b^o]^T$ iteratively. In this process, due to the large dynamic range of the SNR parameters, it is more convenient to work with the logarithmic values of $\boldsymbol{\lambda}$. Furthermore, the initial point may be taken as $\boldsymbol{\lambda}_0 = [\rho_1^{-1}, \rho_2^{-1}]^T$.

Gradient-based methods such as gradient descent, Newton and quasi-Newton methods, update $\boldsymbol{\lambda}$ in the following way

$$\log \boldsymbol{\lambda}_{k+1} = \log \boldsymbol{\lambda}_k - \alpha_k \mathbf{P}_k \bigtriangledown_{\log \boldsymbol{\lambda}} f|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_k}, \tag{3.19}$$

where $\mathbf{P}_k$ depends on the specific algorithm and α_k is the step size which satisfies the Wolfe

conditions [107]. If the closed form of $\bigtriangledown_{\log \boldsymbol{\lambda}} f$ is too complicated or unavailable, a finite difference can be used to approximate it [108, Sec. 8.1].

The above approach is applied on-line after each update of the channel matrices. Alternatively, we can optimize $\boldsymbol{\lambda}$ off-line based on *a priori* knowledge of system configurations, fading statistics and SNR values. The above gradient descent method still applies, provided that $f(\boldsymbol{\lambda})$ is replaced by its expectation $\mathrm{E}_{\mathbf{H},\mathbf{G}}\{f(\boldsymbol{\lambda})\}$.[5] The latter is computed by averaging $f(\boldsymbol{\lambda})$ over channel realizations numerically, but this can be done *beforehand* for various possible fading statistics and SNRs. Following this approach, we have found through numerous experiments that the resulting optimal $\boldsymbol{\lambda}_{\text{opt}}$ can be well approximated by linear functions of $\log \rho_1$ and $\log \rho_2$, as in e.g.,

$$\log_{10} \lambda_a^o \approx \mathbf{c}_a^T \boldsymbol{\rho} + d_a \ , \tag{3.20}$$

where $\boldsymbol{\rho} = [\log_{10} \rho_1, \log_{10} \rho_2]^T$, and $\mathbf{c}_a$, d_a are model coefficients. Consequently, we have to minimize $\mathrm{E}_{\mathbf{H},\mathbf{G}}\{f(\boldsymbol{\lambda})\}$ for only a *small* set of representative SNRs, and then use total least-square fitting to get the model coefficients. Then, (3.20) is used in practical system implementation to update $\boldsymbol{\lambda}$ based on the instantaneous SNR measurements.

The complexity of optimizing $\boldsymbol{\lambda}$ depends on which of the above two approaches is taken. If $\boldsymbol{\lambda}$ is optimized for each channel realization, the complexity is relatively high, but still lower than the methods in [60]. The major complexity comes from computing the gradients. For instance, in order to obtain the gradient of the MSE (using finite difference), NC-NC needs approximately $8M$ matrix multiplications and 6 matrix inverses (of size $N_S \times N_S$) per iteration. In contrast, the method in [60] requires $13M$ multiplications and $2M$ inverses (matrix sizes between $N_S \times N_S$ and $N_R \times N_R$) per iteration. In our simulations using finite difference, it usually takes the gradient descent method fewer iterations to converge than the method in [60].

More importantly, if the optimal parameters are designed off-line, the complexity of obtaining the parameters from a table lookup or the linear formula in (3.20) is almost negligible. This simplicity is one of the most attractive aspects of the proposed hybrid framework, especially for systems with fixed relay infrastructures whose channels remain relatively stationary.

[5] For the outage capacity, $f(\boldsymbol{\lambda})$ is replaced by a corresponding implicit function, cf. (3.14), instead of the expectation.

3.5 Numerical Results and Discussion

In this section, the performance criteria introduced in Sec. 3.4.2 are studied numerically to gain a better understanding of the proposed hybrid relaying framework. First, the behaviors of the capacity and sum MSE provide new insights into how the parameters λ_a and λ_b affect system performance, which complements the interpretations in Sec. 3.4.1. Then, numerical comparisons with existing designs illustrate that the hybrid framework achieves a good balance between performance and complexity. Lastly, the linear formula in (3.20) brings further simplifications with minor performance loss.

The following system configurations and parameters are used throughout this section. Unless otherwise stated, the system of interest is a 1S-3R-1D system with $N_S = N_R = N_D = 4$ and $\rho_1 = \rho_2 = 15$dB. The noises induced at the relay and destination have the same power: $\sigma_w^2 = \sigma_n^2$. The M relay stations transmit the same amount of power: $\mathcal{P}/M$, which means that the individual power constraints in (3.8) are uniquely specified by ρ_2. The wireless channels undergo slow fading, and the channel matrices have statistically independent, circularly symmetric complex Gaussian entries with zero mean and unit variance.

3.5.1 Effects of the Parameters on Capacity and MSE Performance

We study the impact of $\boldsymbol{\lambda}$ on system performance by plotting the contours of the capacity and the sum of MSEs in Fig. 3.5. To obtain these contours, each of the backward and forward channel matrices, i.e., $\mathbf{H}_k$ or $\mathbf{G}_k$ for $k = 1, \cdots, M$, is randomly generated but held constant. All the hybrid strategies, NC-NC, C-NC, NC-C and C-C, are considered and for simplicity, the same λ_a and λ_b are used for different relays.[6] In each subplot, the circle represents the optimal operating point, while the square represents the parameter pair $\boldsymbol{\lambda}_\rho \triangleq [\rho_1^{-1}, \rho_2^{-1}]^T$, which is associated to linear MMSE processing (cf. Table 3.1). Several observations and conclusions can be made from these performance contours (and those for other channel realizations not shown here):

(1) Although optimizing $\boldsymbol{\lambda}$ is bound to improve performance, the performance gap can be quite remarkable. The optimal parameter pair $\boldsymbol{\lambda}_{\text{opt}} = [\lambda_1^o, \lambda_2^o]^T$ is also notably larger than $\boldsymbol{\lambda}_\rho$.

[6] In fact, we have verified numerically that choosing different values for each relay brings only marginal performance improvement, but leads to higher complexity because of the multi-dimensional search for the optimal parameters.

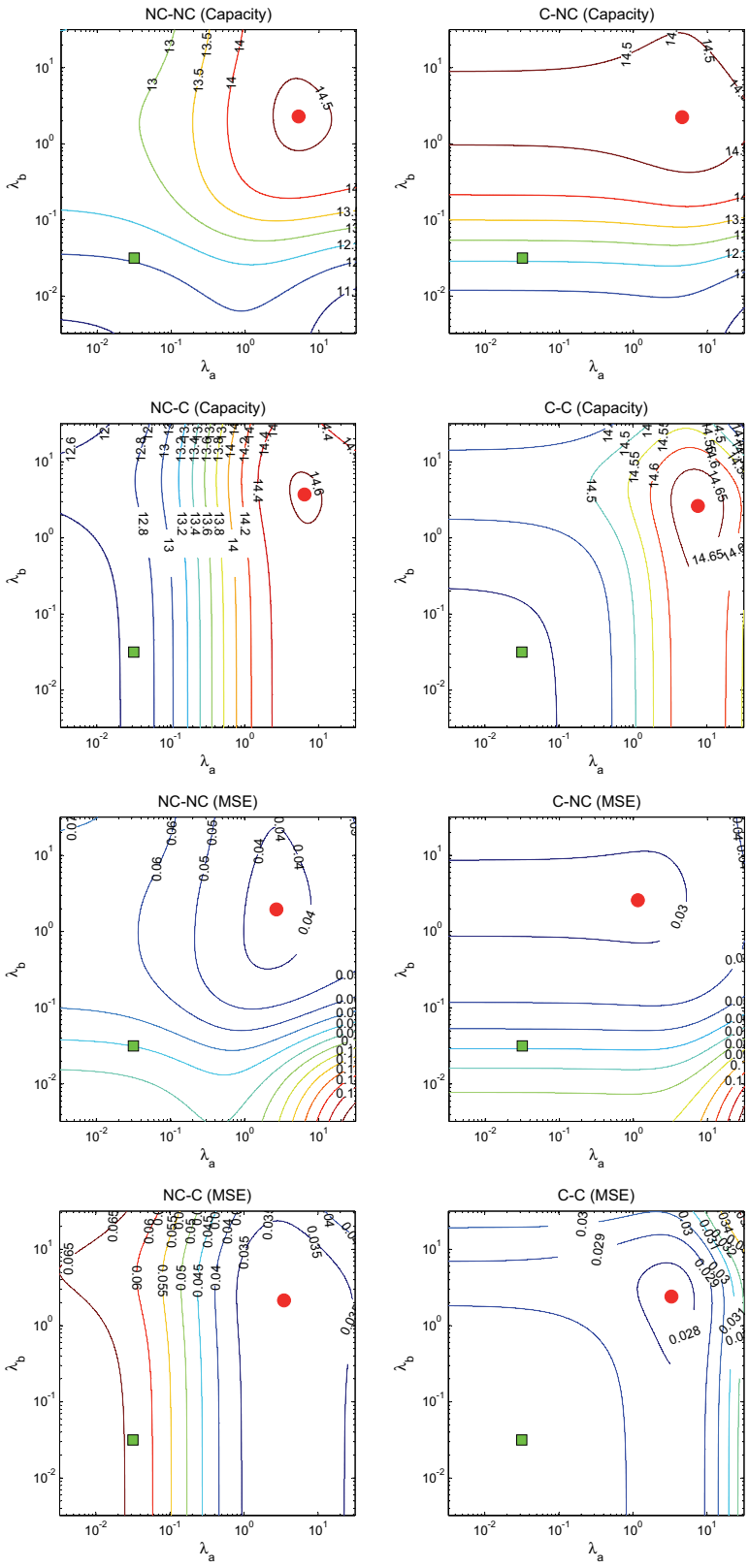

Fig. 3.5 Capacity and MSE contours versus (λ_a, λ_b) for a given realization of the backward and forward channels.

Table 3.2 The relaying strategies under comparison

Strategy	Comments
Simplistic AF (SAF)	$\mathbf{F}_k = \eta_k \mathbf{I}$
SVD with uniform power allocation	$\mathbf{F}_k = \eta_k \mathbf{V}_{2k} \mathbf{U}_{1k}^H$ [15]
MF [13]	NC-NC(∞, ∞)
ZF [13]	NC-NC$(0, 0)$
Linear MMSE [13]	NC-NC$(\rho_1^{-1}, \rho_2^{-2})$
Two-step MMSE [64]	C-NC$(\rho_1^{-1}, 0)$
NC-NC, C-NC, NC-C, C-C	Proposed hybrid methods
Upper bound	Iterative algorithms [60]

(2) The capacity or MSE is not sensitive to small perturbations of $\boldsymbol{\lambda}$. In addition, the system performance is less sensitive to the parameter of the cooperative substructure, than to that of the non-cooperative substructure. This can be explained by the fact that the term $\sum_{j=1}^{M} \mathbf{H}_j^H \mathbf{H}_j$, or $\sum_{j=1}^{M} \mathbf{G}_j \mathbf{G}_j^H$, is the sum of multiple positive semidefinite Wishart-distributed matrices. Since the relays are not close to each other, these matrices are statistically independent because of spatial diversity. If there exists a very small eigenvalue $\epsilon > 0$ whose corresponding eigenvector is $\mathbf{x}$ ($\|\mathbf{x}\|_2 = 1$), each term in the sum

$$\mathbf{x}^H \Big(\sum_{j=1}^{M} \mathbf{H}_j^H \mathbf{H}_j \Big) \mathbf{x} = \sum_{j=1}^{M} \|\mathbf{H}_j \mathbf{x}\|_2^2 = \epsilon$$

cannot be larger than ϵ. Therefore, the probability that the sum matrix has a very small eigenvalue is significantly lower.

(3) If either λ_a or λ_b is fixed and the other parameter increases from zero to infinity, the MSE (or capacity) using first decreases (increases) and then increases (decreases).

From the above observations, the proposed non-cooperative and cooperative relaying substructures, although simple, show strong potential and advantage which were not realized in previous works.

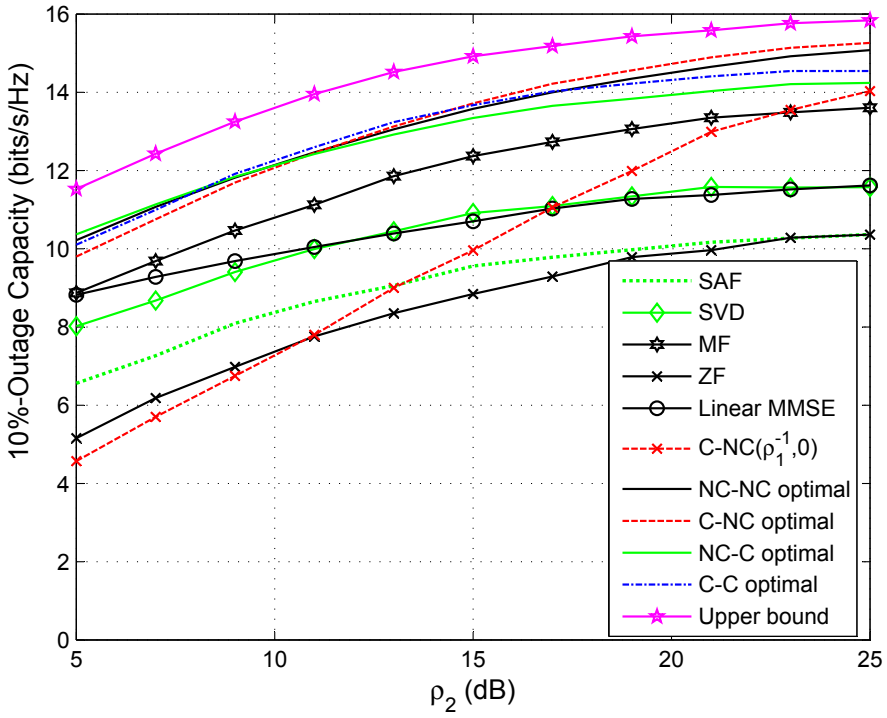

Fig. 3.6 10%-outage capacity for 1S-3R-1D system with $\rho_1 = 15$dB.

3.5.2 Performance Comparison

We next compare the proposed hybrid relaying strategies with other existing approaches in terms of 10%-outage capacity and average BER, based on Monte-Carlo simulations. The non-regenerative MIMO relaying strategies under comparison are listed in Table 3.2. In Sec. 3.5.1, the design parameters $\boldsymbol{\lambda} = [\lambda_a, \lambda_b]^T$ were optimized for a fading channel instance. Here for simplicity, they are optimized using (3.20) based on the *a priori* knowledge of the channel statistics. In the BER simulations, the source antennas transmit independent uncoded 16-QAM modulated streams, and the destination user employs the linear MMSE MIMO combiner described in Section 3.4.2 to decode the information bits. The theoretical SER for each substream is upper bounded by (3.18), in which $\alpha = 3$ and $\beta = 10$ for 16-QAM. The theoretical BER is approximately equal to $P_\epsilon/4$, which is used to search for the BER-optimal $\boldsymbol{\lambda}_{\text{opt}} = [\lambda_a^o, \lambda_b^o]^T$. In the simulations, we set the first SNR $\rho_1 = 15$dB and increase the second SNR ρ_2 from 5 to 25dB.

The 10%-outage capacity of the 1S-3R-1D link for the different relaying strategies is

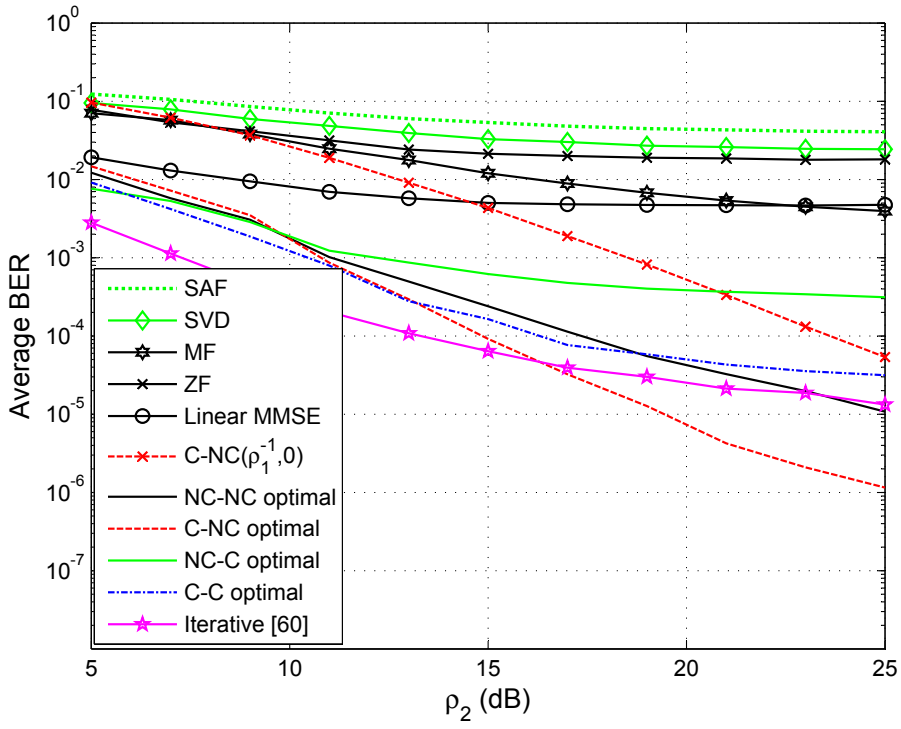

Fig. 3.7 BER performance for 1S-3R-1D system with $\rho_1 = 15$dB.

plotted in Fig. 3.6. As predicted by previous analysis, SAF and SVD perform unsatisfactorily due to their inability to achieve sufficient distributed array gain. Among the four hybrid relaying strategies with special parameter values, i.e., MF, ZF, linear MMSE and two-step MMSE, only MF can outperform SVD over a broad range of SNR values. This is because the good performance of a hybrid relaying strategy is guaranteed not only by coherent superposition of parallel transmissions, but also by efficient exploitation of the interferences without noise over-amplification. As expected, the performance of MF remains inferior to that of the proposed hybrid relaying strategies with optimal parameters. Indeed, the latter can result in significant improvement over MF in spectral efficiency by 1 to 1.5 bits/s/Hz. They come within less than one bit of the upper bound set by the iterative algorithm in [60], but with much lower complexity. Of the four methods, NC-C, C-C and C-NC are, respectively, the best strategy for high, intermediate and low ρ_2 values. The non-cooperative strategy NC-NC remains close to the best performance achieved by the other three cooperative strategies over the range of SNR values considered.

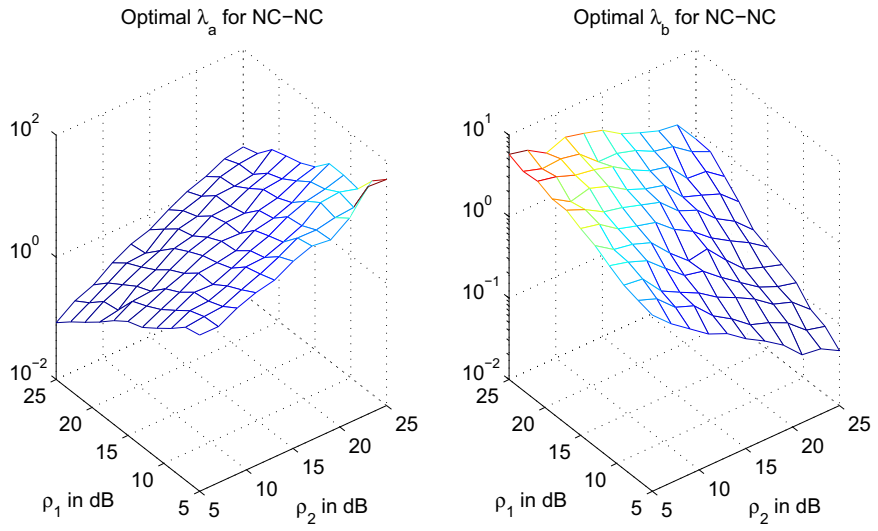

Fig. 3.8 The outage-capacity-optimal λ_a^o and λ_b^o values versus ρ_1 and ρ_2

The BER results are plotted in Fig. 3.7 where again, the performance of SAF and SVD are unsatisfactory and ZF, MF and linear MMSE perform slightly better. The BER of C-NC$(\rho_1^{-1}, 0)$ is much lower for high ρ_2, because this strategy can benefit from the cooperation between relays. However, with the optimal choice of parameters, 1S-MR-1D systems can fully exploit the potential of the proposed hybrid relaying framework: NC-NC, C-C and C-NC all lead to much lower BER values than previously investigated methods. In the mid-to-high SNR range, C-NC even performs better than the iterative MMSE method in [60]. In terms of BER, cooperation between relays brings in significant performance gain.

Finally, simulation results not shown demonstrate that small errors in channel estimation and SNR estimation only lead to small performance degradation. That is, the proposed hybrid strategies are not overly sensitive to such modeling errors.

3.5.3 Further Simplifications

In Sec. 3.4.3, we proposed that the logarithmic values of the optimal parameters, $\boldsymbol{\lambda}_{\text{opt}} = [\lambda_a^o, \lambda_b^o]^T$, can be well approximated by linear functions of $\boldsymbol{\rho} = [\log \rho_1, \log \rho_2]^T$. Considering for example the NC-NC method, the optimal parameters that maximize the outage capacity are plotted against ρ_1 and ρ_2 in Fig. 3.8, where the logarithmic scale is used. It is observed that the relationship between $\log \lambda_a^o$ (or $\log \lambda_b^o$) and $\boldsymbol{\rho}$ is very close to a plane, implying that the expression in (3.20) is sufficiently accurate. Similar relationships can be established for

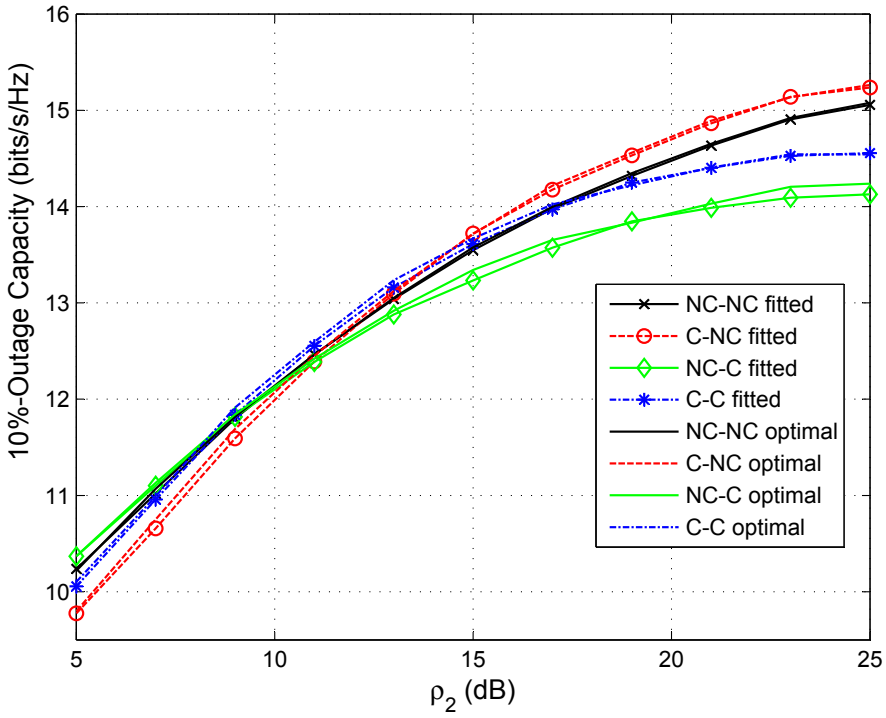

Fig. 3.9 10%-outage capacity: fitted parameters versus optimal parameters

the criterion of average BER, and also for the other three hybrid methods: C-NC, NC-C and C-C.

The outage capacity under the parameters obtained from the empirical linear formulas is now compared with that when the parameters are optimized per channel instance. As seen in Fig. 3.9, the hybrid relaying strategies designed in this way cause negligible loss in performance, but the optimization of the parameters has much lower complexity.

3.6 Summary

In this chapter, in order to achieve a balance between performance and complexity for non-regenerative combining-type 1S-MR-1D relay systems, we proposed a unified hybrid framework in which the relaying matrices are generated by cascading two substructures. For each of these two substructures, we introduced both non-cooperative and cooperative, one-dimensional parametric families of candidate matrix transformations. This unified

framework provides a generalization of several existing approaches and allows for the classification and comparison of all the possible combinations of the proposed substructures.

Within this hybrid framework, the design parameters $\boldsymbol{\lambda}$ can be further optimized, resulting in significant performance improvements. This can be done on-line based on individual channel estimates or off-line based on *a priori* knowledge of the channel statistics. In the latter case, the optimal parameters can be well approximated by linear functions of SNR $[\log \rho_1, \log \rho_2]^T$ with minor performance loss.

The optimal $\boldsymbol{\lambda}_{\text{opt}}$ differs significantly from those corresponding to the ZF, MF and linear MMSE relaying strategies. Through simulations, we showed that the capacity of selected hybrid schemes (with optimized parameters) comes within 1bits/s/Hz of the upper bound achieved by the capacity-optimal iterative method in [60]. In the mid-to-high SNR range, the BER performance of C-NC even exceeds that of the MSE-optimal iterative method. The proposed hybrid methods therefore achieve a good balance between performance and complexity: they outperform existing low-complexity strategies by a large margin in terms of both capacity and BER, and at the same time, are significantly simpler than previous near-optimal iterative algorithms.

Chapter 4

MMSE Transceiver Design for Combining-Type Relaying

In Chapter 3, we proposed a hybrid MIMO relaying framework for combining-type 1S-MR-1D systems.[1] The relaying matrices had special structures and the underlying parameters were further optimized for better performance. In this chapter, instead, we take a more comprehensive approach of formulating the design of these relaying matrices into mathematical optimization problems. This approach leads to new MMSE-based transceivers that perform better than previous ones and the hybrid relaying strategies. The organization of this chapter is as follows: Sec. 4.1 reviews combining-type 1S-MR-1D systems, reiterates our motivation and introduces our main contributions in this chapter. Sec. 4.2 presents the system model and formulates the mathematical problem. Sec. 4.3 derives the closed-form optimal relaying matrices under the weighted sum power constraint. Sec. 4.4 studies the per-relay power constraints. The joint design of the relaying matrices and the MIMO equalizer is discussed in Sec. 4.5. Sec. 4.6 covers the implementation issues and computational complexity. Numerical results are presented in Sec. 4.7 followed by a brief conclusion in Sec. 4.8.

[1]Parts of Chapter 4 have been presented at IEEE 2011 Global Communications Conference (GLOBECOM) in Houston, Texas, USA [97], and published in IEEE Transactions on Signal Processing [109].

4.1 Introduction

Communication between a source and a destination can be assisted by either a single or multiple relays [49]. As before, these configurations are referred to as 1S-1R-1D and 1S-MR-1D. The optimal relaying matrices for 1S-1R-1D configurations are well established for a wide variety of criteria when the transmit power of the relay is constrained [15, 28, 29, 31]. These matrices share a common SVD structure which diagonalizes the backward and forward channels. This framework, however, cannot be extended to the joint design of multiple relaying matrices for 1S-MR-1D systems. Indeed, since the relays can only process their own signals, the compound AF matrix has to be *block-diagonal*.

The essential feature of an appropriate relaying strategy is that the signals from different relays should be coherently combined at the destination, thereby leading to a *distributed array gain* [13, 51]. In this regard, some strategies have been proposed that "borrow" ideas from MIMO transceiver design, including MF, ZF, linear MMSE [13], QR decomposition [56] and the hybrid relaying framework in Chapter 3. These heuristic methods, although structurally constrained, were shown to perform much better than simplistic AF which only amplifies the signals. A more comprehensive approach is to formulate the collaborative design of the relaying matrices as optimization problems with power constraints [59–63, 97]. The objective can be to maximize the achievable rate [59] or to minimize the MSE [60, 61]. However, most works rely on numerical algorithms such as gradient descent [60], bisection [61] and iterative schemes [60–62] to obtain the optimal solution. These methods have high implementation complexity and lack closed-from expressions that provide insights, which in turn limits their potential feasibility. For completeness, it is worth mentioning that explicit formulas were derived in [64–66] when the power constraints are enforced on the signals *received* at the destination. However, these results do not carry over to the case when the constraints are imposed on the *transmit* power of the relays [63].

In this chapter, we concentrate on the similar problems of designing the multiple relaying matrices, with the purpose of minimizing the MSE between the input and output signals. Two types of constraints on the transmit power of the relays are considered separately: 1) a weighted sum power constraint which was not investigated before, and 2) per-relay power constraints. The problems are first recast as standard quadratically constrained quadratic programs (QCQPs) through vectorization. As opposed to using general-purpose interior-point methods [108], we exploit the inherent structures of the problems to develop

more efficient algorithms. Under the weighted sum power constraint, the optimal solution is expressed as an explicit function of a Lagrangian parameter. By introducing a complex scaling factor at the destination, we derive a closed-form expression for this parameter, thereby overcoming the hurdle of solving an implicit nonlinear equation. Under the per-relay power constraints, the optimal solution is the same as that under the weighted sum power constraint if a particular set of weights is chosen. We then propose a simple iterative power balancing algorithm to compute these weights efficiently. In addition, under both types of constraints, we investigate the joint design of a MIMO equalizer at the destination and the relaying matrices, using block coordinate descent or steepest descent. The BER simulation results demonstrate better performance for the proposed MMSE-based relaying strategies, under either type of constraints, with or without the equalizer, than previous methods.

Our work provides new insights into the design of non-regenerative 1S-MR-1D systems. Firstly, we point out the possible non-uniqueness of the solution to the first-order necessary condition, which was overlooked in [61,62,97]. Moreover, it is not legitimate to simply choose the minimum-norm solution, unless the vectorization is done on specific transformations of the relaying matrices instead of these matrices themselves. Secondly, the optimal design does not require global CSI availability: each relay only needs to know its own backward and forward channel, together with a little additional information. Thirdly, under the weighted sum power constraint, the optimal strategy tends to allocate more power to those relays with better source-relay links or worse relay-destination links. Lastingly, under the per-relay power constraints, the optimal strategy sometimes does not use the maximum power at some relays. Forcing equality in the per-relay power constraints as in [60] and Chapter 3 would result in loss of optimality. Another interesting point is that, no matter how low the SNR is at a particular relay, this relay does not have to be turned off completely.

4.2 System Model and Problem Formulation

4.2.1 System Model

In the 1S-MR-1D system model depicted by Fig. 4.1, a multi-antenna source is sending symbols to a multi-antenna destination with the aid of multiple multi-antenna relays. The

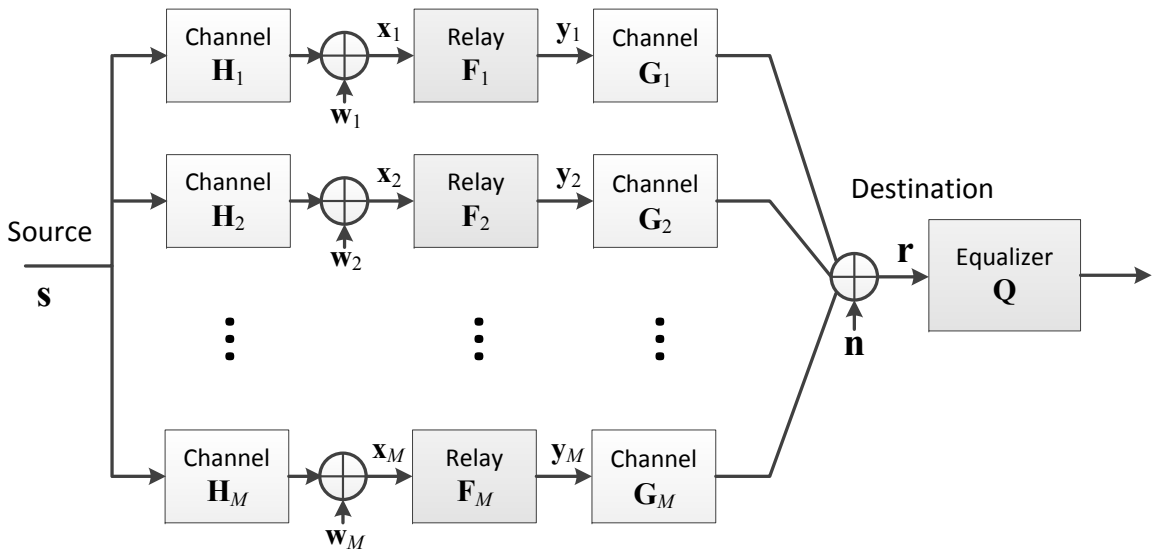

Fig. 4.1 System model of 1S-MR-1D

transmitted signals propagate through the backward channels between the source and the relays. These signals are processed at the individual relays and propagate through the forward channels to the destination. The relays work in a half-duplex mode: their antennas are used for either transmitting or receiving during different time slots. As in Chapter 3, we neglect the presence of a possible direct source-to-destination link, which is typically hindered by high levels of attenuation.

We assume that the channels undergo frequency-flat block-fading [2]. The source does not have access to the CSI; each relay knows its own backward and forward channels and a little additional shared information; the destination may need an equalizer matrix. Once knowing the structures of the optimal solution, we shall be able to discuss this topic in detail (*cf.* Section 4.6). The channel matrices have to be estimated timely and accurately, which is an important topic in its own right. For more details, we refer the reader to [46, 47, 101] and the references therein.

The bandpass signals and channels are modeled in terms of their discrete-time complex baseband counterparts. The numbers of antennas at the source, relays and destination are respectively denoted by N_S, N_R and N_D.[2] The source signal $\mathbf{s} \in \mathbb{C}^{N_S \times 1}$ consists of N_S statistically independent symbol streams. It is assumed to have zero mean and a full-rank

[2]For notational simplicity, each relay is equipped with the same number of antennas; however, generalization to different numbers of antennas at the relays, i.e. $N_{R,k}$, is straightforward.

covariance matrix $\mathbf{R}_s \triangleq \mathrm{E}\{\mathbf{s}\mathbf{s}^H\}$. The received signal at the kth relay, $\mathbf{x}_k \in \mathbb{C}^{N_R \times 1}$, can be expressed as

$$\mathbf{x}_k = \mathbf{H}_k \mathbf{s} + \mathbf{w}_k, \quad \forall \ 1 \leq k \leq M, \tag{4.1}$$

where $\mathbf{H}_k \in \mathbb{C}^{N_R \times N_S}$ is the backward channel matrix from the source to the kth relay, and $\mathbf{w}_k \in \mathbb{C}^{N_R \times 1}$ is an additive noise term modeled as a circularly symmetric complex Gaussian random vector with zero mean and full-rank covariance matrix $\mathbf{R}_{w_k} \triangleq \mathrm{E}\{\mathbf{w}_k \mathbf{w}_k^H\}$. The random vectors $\mathbf{w}_1, \ldots, \mathbf{w}_M$ and $\mathbf{s}$ are statistically independent.

The kth relay retransmits its noisy signal $\mathbf{x}_k$ as

$$\mathbf{y}_k = \mathbf{F}_k \mathbf{x}_k, \quad \forall \ 1 \leq k \leq M, \tag{4.2}$$

where $\mathbf{F}_k \in \mathbb{C}^{N_R \times N_R}$ is the corresponding non-regenerative MIMO relaying matrix. The signal received at the destination, denoted by $\mathbf{r} \in \mathbb{C}^{N_D \times 1}$, takes the form of

$$\mathbf{r} = \sum_{k=1}^{M} \mathbf{G}_k \mathbf{y}_k + \mathbf{n} = \sum_{k=1}^{M} \mathbf{G}_k \mathbf{F}_k \mathbf{H}_k \mathbf{s} + \sum_{k=1}^{M} \mathbf{G}_k \mathbf{F}_k \mathbf{w}_k + \mathbf{n}, \tag{4.3}$$

where $\mathbf{G}_k \in \mathbb{C}^{N_D \times N_R}$ is the forward channel matrix from the kth relay to the destination and $\mathbf{n} \in \mathbb{C}^{N_D \times 1}$ is the noise induced at the destination receiver. This term is independent from $\mathbf{s}$ and $\{\mathbf{w}_k\}$ and also modeled as a circularly symmetric complex Gaussian random vector, with zero mean and covariance $\mathbf{R}_n \triangleq \mathrm{E}\{\mathbf{n}\mathbf{n}^H\}$. The destination may apply a linear MIMO equalizer (combiner) $\mathbf{Q} \in \mathbb{C}^{N_S \times N_D}$, resulting in

$$\hat{\mathbf{r}} = \mathbf{Q}\mathbf{r}. \tag{4.4}$$

The above signal model can also be expressed in a compact block-diagonal form, *viz.*,

$$\hat{\mathbf{r}} = \mathbf{QGFHs} + \mathbf{QGFw} + \mathbf{Qn}, \tag{4.5}$$

where we define $\mathbf{G} \triangleq [\mathbf{G}_1, \ldots, \mathbf{G}_M]$, $\mathbf{H} \triangleq \left[\mathbf{H}_1^H, \ldots, \mathbf{H}_M^H\right]^H$, $\mathbf{F} \triangleq \mathrm{diag}(\mathbf{F}_1, \ldots, \mathbf{F}_M)$ and $\mathbf{w} \triangleq \mathrm{col}(\mathbf{w}_1, \ldots, \mathbf{w}_M)$ with $\mathbf{R}_w \triangleq \mathrm{E}\{\mathbf{w}\mathbf{w}^H\} = \mathrm{diag}\{\mathbf{R}_{w_1}, \cdots, \mathbf{R}_{w_M}\}$. If $M = 1$, this signal model reduces to the 1S-1R-1D case.

Compared with Chapter 3, the model here is more general because we do not assume specific forms for the covariance matrices $\mathbf{R}_w$ and $\mathbf{R}_n$. Therefore, the noise vectors can

also include interferences from other wireless systems. In addition, with proper refinement, the mathematical model in this chapter is applicable to a much broader scope such as 1S-MR-1D systems with broadband transmission [110], distributed relaying systems, and multiuser multi-relay systems.

4.2.2 Problem Formulation

The major goal is to design the relay processing matrices $\{\mathbf{F}_k\}$, so that the distortion between the output $\hat{\mathbf{r}}$ and the input $\mathbf{s}$ is minimized. Our choice of the objective function, for practical reasons, is the MSE:

$$\begin{aligned}\mathrm{MSE}(\mathbf{F},\mathbf{Q}) &\triangleq \mathrm{E}_{\mathbf{s},\mathbf{w},\mathbf{n}}\left\{ \|\hat{\mathbf{r}}-\mathbf{s}\|^2 \right\} \\ &= \mathrm{tr}\left((\mathbf{QGFH}-\mathbf{I})\mathbf{R}_s(\mathbf{QGFH}-\mathbf{I})^H\right) + \\ &\quad + \mathrm{tr}(\mathbf{QGFR}_w\mathbf{F}^H\mathbf{G}^H\mathbf{Q}^H) + \mathrm{tr}(\mathbf{QR}_n\mathbf{Q}^H). \end{aligned} \tag{4.6}$$

Although BER performance also depends on nonlinear components such as channel coding, space-time coding, interleaving and constellation mapping, the MSE serves as a good performance indicator and is more mathematically tractable [111].

Two types of power constraints are separately imposed on the relays. The first is the weighted sum power constraint

$$\sum_{k=1}^{M} w_k \mathrm{tr}(\mathbf{F}_k\mathbf{R}_{x_k}\mathbf{F}_k^H) \le P_r, \tag{4.7}$$

where $\mathbf{R}_{x_k} \triangleq \mathrm{E}\left\{\mathbf{x}_k\mathbf{x}_k^H\right\} = \mathbf{H}_k\mathbf{R}_s\mathbf{H}_k^H + \mathbf{R}_{w_k}$, and $w_k \ge 0$ for $1 \le k \le M$ are the weights assigned to different relays. The other type is the per-relay power constraints, *i.e.*, each relay has its own power budget, expressed as

$$\mathrm{tr}(\mathbf{R}_{y_k}) = \mathrm{tr}(\mathbf{F}_k\mathbf{R}_{x_k}\mathbf{F}_k^H) \le P_k, \qquad \forall\ 1 \le k \le M, \tag{4.8}$$

where $\mathbf{R}_{y_k} \triangleq \mathrm{E}\left\{\mathbf{y}_k\mathbf{y}_k^H\right\} = \mathbf{F}_k\mathbf{R}_{x_k}\mathbf{F}_k^H$. Unlike the per-relay power constraints, the physical meaning may not seem very straightforward for the weighted sum power constraint. Since

$\mathrm{tr}(\mathbf{F}_k\mathbf{R}_{x_k}\mathbf{F}_k^H)$ is always nonnegative, we have

$$0 \le w_k \mathrm{tr}(\mathbf{F}_k\mathbf{R}_{x_k}\mathbf{F}_k^H) \le \sum_{j=1}^{M} w_j \mathrm{tr}(\mathbf{F}_j\mathbf{R}_{x_j}\mathbf{F}_j^H) \le P_r. \tag{4.9}$$

Therefore, the weighed sum power constraint implicitly includes a set of per-relay power constraints

$$\mathrm{tr}(\mathbf{F}_k\mathbf{R}_{x_k}\mathbf{F}_k^H) \le \frac{P_r}{w_k}, \quad 1 \le k \le M. \tag{4.10}$$

The weighted sum power constraint itself refers to a half space on one side of a hyperplane. For example, when there are $M = 2$ relays, the per-relay power constraints refers to a rectangular region in the two-dimensional plane, whereas the weighted sum power constraint denotes a triangular region formed by the x-axis, y-axis and a straight line.

To simplify the mathematical development, it is convenient to vectorize the relaying matrices. To this end, we define

$$\mathbf{f}_k \triangleq \mathrm{vec}(\mathbf{F}_k\mathbf{R}_{x_k}^{1/2}), \quad 1 \le k \le M, \tag{4.11}$$

and $\mathbf{f} \triangleq \mathrm{col}(\mathbf{f}_1, \cdots, \mathbf{f}_M)$. [3] The reason for this definition, instead of $\mathrm{vec}(\mathbf{F}_k)$, is that the square of the 2-norm of $\mathbf{f}_k$ is equal to the transmit power of the kth relay, *viz.*,

$$\|\mathbf{f}_k\|_2^2 = \mathbf{f}_k^H\mathbf{f}_k = \mathrm{tr}(\mathbf{F}_k\mathbf{R}_{x_k}\mathbf{F}_k^H). \tag{4.12}$$

As shown later, this will bring much convenience. It is straightforward to invert (4.11) as $\mathbf{F}_k = \mathrm{unvec}(\mathbf{f}_k)\mathbf{R}_{x_k}^{-1/2}$. For notational simplicity, we also define the matrices

$$\mathbf{T}_k \triangleq (\mathbf{H}_k^H\mathbf{R}_{x_k}^{-1/2})^T \otimes \mathbf{G}_k^H\mathbf{Q}^H, \tag{4.13a}$$

$$\mathbf{S}_k \triangleq (\mathbf{R}_{x_k}^{-1/2}\mathbf{R}_{w_k}\mathbf{R}_{x_k}^{-1/2})^T \otimes (\mathbf{G}_k^H\mathbf{Q}^H\mathbf{Q}\mathbf{G}_k), \tag{4.13b}$$

for $1 \le k \le M$. These matrices serve as the building blocks for the following matrices and

[3] The square root of an Hermitian positive semidefinite matrix $\mathbf{R}$ is defined as another Hermitian positive semidefinite matrix $\mathbf{R}^{1/2}$, satisfying $\mathbf{R}^{1/2}\mathbf{R}^{1/2} = \mathbf{R}$. If the eigenvalue decomposition (EVD) of $\mathbf{R}$ is $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$, its unique square root is equal to $\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^H$.

vectors:

$$\mathbf{T} \triangleq [\mathbf{T}_1^T, \ldots, \mathbf{T}_M^T]^T, \tag{4.14a}$$

$$\mathbf{S} \triangleq \mathrm{diag}(\mathbf{S}_1, \ldots, \mathbf{S}_M), \tag{4.14b}$$

$$\mathbf{\Phi} \triangleq \mathbf{T}(\mathbf{R}_s^T \otimes \mathbf{I}_{N_S})\mathbf{T}^H + \mathbf{S}, \tag{4.14c}$$

$$\mathbf{b} \triangleq \mathbf{T}(\mathbf{R}_s^T \otimes \mathbf{I}_{N_S})\mathrm{vec}(\mathbf{I}_{N_S}). \tag{4.14d}$$

With the above notations, the objective function in (4.6) becomes a quadratic function of the vector $\mathbf{f}$:

$$\mathrm{MSE}(\mathbf{f}, \mathbf{Q}) = \mathbf{f}^H\mathbf{\Phi}\mathbf{f} - \mathbf{f}^H\mathbf{b} - \mathbf{b}^H\mathbf{f} + \mathrm{tr}(\mathbf{R}_s) + \mathrm{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H), \tag{4.15}$$

where we have used the following properties [112]

$$\mathrm{vec}(\mathbf{ABC}) = \left(\mathbf{C}^T \otimes \mathbf{A}\right)\mathrm{vec}(\mathbf{B}), \tag{4.16a}$$

$$(\mathbf{AB}) \otimes (\mathbf{CD}) = (\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}), \tag{4.16b}$$

$$\mathrm{tr}(\mathbf{A}^T\mathbf{Y}^T\mathbf{BX}) = [\mathrm{vec}(\mathbf{Y})]^T(\mathbf{A} \otimes \mathbf{B})\mathrm{vec}(\mathbf{X}). \tag{4.16c}$$

Hereafter, we may denote the arguments of the function $\mathrm{MSE}(\mathbf{f}, \mathbf{Q})$ in (4.15) differently, to emphasize its dependence on certain variables, vectors or matrices.

The power constraints are also represented in terms of $\mathbf{f}$. For convenience, define $\mathbf{I}_{(k)}$ as

$$\mathbf{I}_{(k)} \triangleq \mathrm{diag}(\mathbf{0}_{N_R^2}, \ldots, \mathbf{I}_{N_R^2}, \ldots, \mathbf{0}_{N_R^2}), \tag{4.17}$$

where $\mathbf{I}_{N_R^2}$ is in the kth diagonal sub-block. A weighted sum of such matrices is also defined: $\mathbf{I}_{\mathrm{sum}} \triangleq \sum_{k=1}^{M} w_k\mathbf{I}_{(k)}$. Then, the weighted sum power constraint in (4.7) becomes

$$\mathbf{f}^H\mathbf{I}_{\mathrm{sum}}\mathbf{f} \leq P_r, \tag{4.18}$$

and the per-relay power constraints in (4.8) would be

$$\mathbf{f}^H\mathbf{I}_{(k)}\mathbf{f} \leq P_k, \qquad \forall\ 1 \leq k \leq M. \tag{4.19}$$

To help the readers understand the above redefinitions of matrix blocks, we provide some

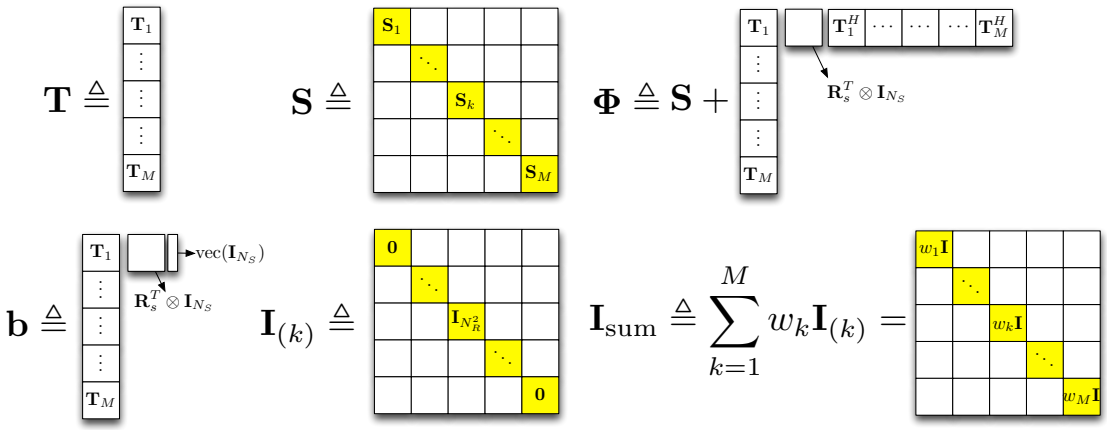

Fig. 4.2 Graphical demonstrations of the redefined matrix blocks

graphical demonstrations in Fig. 4.2.

The above formulated optimization problems are flexible with respect to the equalizer $\mathbf{Q}$ — it can be either pre-determined, or jointly designed with $\{\mathbf{F}_k\}$ (or equivalently $\mathbf{f}$). If $\mathbf{Q}$ is fixed, the problems under both types of constraints are standard QCQPs that can be solved by general-purpose interior-point methods [108, 113]. In Section 4.3 and 4.4, however, the proper exploitation of the sparse structures of $\mathbf{\Phi}$, $\mathbf{b}$ and $\mathbf{I}_{(k)}$ leads to more efficient algorithms and in some cases closed-form expressions for the optimal solution. For the joint design of $\mathbf{Q}$ and $\mathbf{f}$, we shall propose two algorithms in Section 4.5, both of which rely upon the results from Section 4.3 and 4.4.

4.3 The Weighted Sum Power Constraint

In this section, we assume $\mathbf{Q}$ is fixed and derive the optimal $\mathbf{f}^\star$ under the weighted sum power constraint. The first step is to establish the optimality conditions through the framework of Lagrangian duality. Then, the optimal solution $\mathbf{f}^\star$ is expressed as an explicit function of a Lagrangian parameter. By introducing a complex scaling factor in $\mathbf{Q}$, we derive a closed-form expression for this parameter.

4.3.1 Optimality Conditions

Most constrained optimization problems are solved through Lagrangian duality [108, 113, 114]. The starting step is to define the Lagrangian function as

$$L(\mathbf{f}, \lambda) \triangleq \mathrm{MSE}(\mathbf{f}) + \lambda(\mathbf{f}^H \mathbf{I}_{\mathrm{sum}} \mathbf{f} - P_r), \tag{4.20}$$

where the dual variable satisfies $\lambda \geq 0$. The infimum of (4.20) over $\mathbf{f} \in \mathbb{C}^{MN_R^2 \times 1}$ is defined as

$$D(\lambda) \triangleq \inf_{\mathbf{f}} L(\mathbf{f}, \lambda). \tag{4.21}$$

The dual problem is defined as:

$$\begin{aligned} &\text{maximize} && D(\lambda) \\ &\text{subject to} && \lambda \geq 0. \end{aligned} \tag{4.22}$$

Let the solution of the primal problem be $\mathbf{f}^\star$ and $p^\star \triangleq \mathrm{MSE}(\mathbf{f}^\star)$; let the solution of the dual problem be $\lambda^\star$ and $d^\star \triangleq D(\lambda^\star)$. For a convex primal problem, strong duality holds (*i.e.*, the duality gap $p^\star - d^\star$ is zero) if the Slater's condition is satisfied [113, p.226]. Here, the primal problem is convex ($\mathbf{\Phi}$ and $\mathbf{I}_{\mathrm{sum}}$ are both positive semidefinite) and the Slater's condition is always satisfied ($\mathbf{f} = \mathbf{0}$ is strictly feasible: $0 < P_r$). Henceforth, the Karush-Kuhn-Tucker (KKT) conditions are *necessary and sufficient* for the optimal primal-dual pair $(\mathbf{f}^\star, \lambda^\star)$, *viz.*,

$$\nabla_{\bar{\mathbf{f}}} L(\mathbf{f}, \lambda^\star)|_{\mathbf{f}=\mathbf{f}^\star} = \mathbf{0}, \tag{4.23a}$$

$$\mathbf{f}^{\star H} \mathbf{I}_{\mathrm{sum}} \mathbf{f}^\star - P_r \leq 0, \tag{4.23b}$$

$$\lambda^\star \geq 0, \tag{4.23c}$$

$$\lambda^\star \big(\mathbf{f}^{\star H} \mathbf{I}_{\mathrm{sum}} \mathbf{f}^\star - P_r\big) = 0. \tag{4.23d}$$

Among these four conditions, the first-order necessary condition (4.23a), which we referred to as the *stationarity condition* from now on, determines the analytical form for the optimal solution; the complementary slackness condition (4.23d) serves as the key to computing the value of $\lambda^\star$.

4.3.2 The Solution Space of the Stationarity Condition

The stationarity condition (4.23a) can be rewritten as the following linear equation

$$\mathbf{\Psi f}^\star = \mathbf{b}, \tag{4.24}$$

in which $\mathbf{\Psi} \triangleq \mathbf{\Phi} + \lambda^\star \mathbf{I}_{\text{sum}}$ is a function of $\lambda^\star$. Three questions can be asked about (4.24): Does the solution always exist? If yes, is it unique? If non-unique, what is the general solution form? The following theorem answers the first question:

Theorem 4.3.1. *The stationarity equation* (4.24) *has at least one solution for any* $\lambda^\star \geq 0$ *and* $w_k \geq 0$*, that is:* $\mathbf{b} \in \mathcal{R}(\mathbf{\Psi})$*.*

Proof. We proceed by contradiction. If no solution exists, $\mathbf{b} \notin \mathcal{R}(\mathbf{\Psi})$. Let $\mathbf{b}_0 = \mathbf{\Psi f}_0$ be the orthogonal projection of $\mathbf{b}$ onto $\mathcal{R}(\mathbf{\Psi})$ and then the non-zero residual $\mathbf{b}_\perp$ would satisfy

$$\mathbf{b}_\perp \triangleq \mathbf{b} - \mathbf{b}_0 = \mathbf{b} - \mathbf{\Psi f}_0 \in \mathcal{N}(\mathbf{\Psi}^H) = \mathcal{N}(\mathbf{\Psi}).$$

The value of the Lagrangian function (4.20) evaluated at $\mathbf{f} = \mathbf{f}_0 + \alpha \mathbf{b}_\perp$ would be

$$L(\mathbf{f}_0 + \alpha \mathbf{b}_\perp, \lambda^\star) = -\mathbf{f}_0^H \mathbf{\Psi f}_0 - \mathbf{b}_\perp^H \mathbf{f}_0 - \mathbf{f}_0^H \mathbf{b}_\perp - 2\alpha \|\mathbf{b}_\perp\|^2 + \text{tr}(\mathbf{R}_s) + \text{tr}(\mathbf{Q R}_n \mathbf{Q}^H) - \lambda^\star P_r.$$

As $\alpha \to +\infty$, $L(\mathbf{f}_0 + \alpha \mathbf{b}_\perp, \lambda^\star) \to -\infty$ because $\|\mathbf{b}_\perp\|^2 > 0$. However, all terms in the definition of $L(\mathbf{f}, \lambda^\star)$ in (4.20) are nonnegative except for the constant $-\lambda^\star P_r$. This means $L(\mathbf{f}_0 + \alpha \mathbf{b}_\perp, \lambda^\star) \geq -\lambda^\star P_r$, which leads to a contradiction. □

The second question is whether the solution is unique. This is true, if and only if, $\mathbf{\Psi f} = \mathbf{0}$ does not have a non-zero solution, or equivalently $\mathcal{N}(\mathbf{\Psi}) = \{\mathbf{0}\}$. The following propositions establish some facts on $\mathcal{N}(\mathbf{\Phi})$ and $\mathcal{N}(\mathbf{\Psi})$, respectively:

Proposition 4.3.2.

(1) If $\mathbf{f}_k^\perp \in \mathbb{C}^{N_R^2 \times 1}$ *is in the null space of* $\mathbf{I}_{N_R} \otimes \mathbf{Q G}_k$*, the vector* $\mathbf{f}_{(k)}^\perp \triangleq \text{col}(\mathbf{0}, \ldots, \mathbf{f}_k^\perp, \ldots, \mathbf{0}) \in \mathbb{C}^{MN_R^2 \times 1}$ *must be in the null space of* $\mathbf{\Phi}$*.*

(2) For any k*, all such* $\mathbf{f}_{(k)}^\perp$ *together span a subspace* $\mathcal{F}_k$ *with dimension* $N_R^2 - N_R \,\text{rank}(\mathbf{Q G}_k)$*.*

(3) For $k \neq l$*,* $\mathcal{F}_k$ *and* $\mathcal{F}_l$ *are orthogonal to each other.*

(4) The vector $\mathbf{b}$ *is orthogonal to all* $\mathcal{F}_k$ *for* $1 \le k \le M$.

Proof. 1) Define $\mathbf{\Phi}_{lk}$ as the (l,k)th sub-block of $\mathbf{\Phi}$. The lth sub-block of $\mathbf{\Phi}\mathbf{f}_{(k)}^{\perp}$ is

$$[\mathbf{\Phi}_{l1}, \cdots, \mathbf{\Phi}_{lM}]\mathbf{f}_{(k)}^{\perp} = \mathbf{\Phi}_{lk}\mathbf{f}_{k}^{\perp} = \mathbf{X}_k\big(\mathbf{I}_{N_R} \otimes \mathbf{Q}\mathbf{G}_k\big)\mathbf{f}_{k}^{\perp} = \mathbf{0},$$

where $\mathbf{X}_k = \big(\mathbf{R}_{x_k}^{-1/2}\mathbf{H}_k\mathbf{R}_{\mathbf{s}}\mathbf{H}_l^H\mathbf{R}_{x_l}^{-1/2}\big)^T \otimes \mathbf{G}_l^H\mathbf{Q}^H$ if $l \neq k$ and $\mathbf{X}_k = \mathbf{I} \otimes \mathbf{G}_l^H\mathbf{Q}^H$ if $l = k$. 2) Because a one-to-one correspondence exists between $\mathbf{f}_k^{\perp}$ and $\mathbf{f}_{(k)}^{\perp}$ according to the definition, all $\mathbf{f}_{(k)}^{\perp}$ together also span a subspace $\mathcal{F}_k$ isomorphic to $\mathcal{N}(\mathbf{I}_{N_R} \otimes \mathbf{Q}\mathbf{G}_k)$. Since the rank of $\mathbf{I}_{N_R} \otimes \mathbf{Q}\mathbf{G}_k$ (with dimension $N_R N_S \times N_R^2$) is $N_R \operatorname{rank}(\mathbf{Q}\mathbf{G}_k)$, the dimension of its null space is $N_R^2 - N_R \operatorname{rank}(\mathbf{Q}\mathbf{G}_k)$, which is also that of $\mathcal{F}_k$. 3) From the definition, $\mathbf{f}_{(k)}^{\perp H}\mathbf{f}_{(l)}^{\perp} = 0$ always holds for $k \neq l$, and so $\mathcal{F}_k$ and $\mathcal{F}_l$ are orthogonal to each other. 4) According to Theorem 4.3.1, $\lambda^\star = 0$ leads to $\mathbf{b} \in \mathcal{R}(\mathbf{\Phi})$. Therefore, $\mathbf{b}$ is orthogonal to $\mathcal{N}(\mathbf{\Phi}^H) = \mathcal{N}(\mathbf{\Phi})$ which includes any $\mathcal{F}_k$ as a subset. □

Proposition 4.3.3. *Let* $\mathcal{M} \triangleq \{1, \ldots, M\}$ *be the set of relay indexes, and* $\mathcal{K} \triangleq \{m_1, \ldots, m_K\} \subseteq \mathcal{M}$ *includes all the indexes* m_k *satisfying* $\lambda^\star w_{m_k} = 0$. *The null space of* $\mathbf{\Psi}$ *is the direct sum of* $\{\mathcal{F}_{m_k}\}$, *that is,* $\mathcal{N}(\mathbf{\Psi}) = \mathcal{F}_{m_1} \oplus \cdots \oplus \mathcal{F}_{m_K}$.

Proof. We have $\mathcal{F}_{m_1} \oplus \cdots \oplus \mathcal{F}_{m_K} \subseteq \mathcal{N}(\mathbf{\Psi})$ because any $\mathbf{f}_{(m_k)}^{\perp} \in \mathcal{F}_{m_k}$ as defined in Proposition 4.3.2 would satisfy

$$\mathbf{\Psi}\mathbf{f}_{(m_k)}^{\perp} = \mathbf{\Phi}\mathbf{f}_{(m_k)}^{\perp} + \lambda^\star w_{m_k}\mathbf{I}_{(m_k)}\mathbf{f}_{(m_k)}^{\perp} = \mathbf{0}.$$

To prove equality, we only need to prove that $\mathcal{F}_{m_1} \oplus \cdots \oplus \mathcal{F}_{m_K}$ and $\mathcal{N}(\mathbf{\Psi})$ have the same dimension. On the one hand, the subset relation leads to

$$\dim(\mathcal{N}(\mathbf{\Psi})) \ge \dim(\mathcal{F}_{\pi_1} \oplus \cdots \oplus \mathcal{F}_{\pi_M}).$$

On the other hand, we have

$$\begin{aligned}\dim(\mathcal{N}(\mathbf{\Psi})) &= MN_R^2 - \operatorname{rank}(\mathbf{\Psi}) \\ &\leq MN_R^2 - \operatorname{rank}(\mathbf{S} + \lambda^\star \mathbf{I}_{\text{sum}}) \\ &= MN_R^2 - \sum\nolimits_{k=1}^{M} \operatorname{rank}(\mathbf{S}_k + \lambda^\star w_k \mathbf{I}_{N_R^2}) \\ &= KN_R^2 - \sum\nolimits_{k=1}^{K} N_R \times \operatorname{rank}(\mathbf{Q}\mathbf{G}_{m_k}) \\ &= \dim(\mathcal{F}_{m_1} \oplus \cdots \oplus \mathcal{F}_{m_M}),\end{aligned}$$

where the inequality comes from the fact that $\mathbf{\Psi}$ is a sum of positive semidefinite matrices and hence its rank cannot increase by removing $\mathbf{T}(\mathbf{R}_s^T \otimes \mathbf{I})\mathbf{T}^H$. □

Proposition 4.3.3 indicates that a positive $\lambda^\star w_k$ causes the null space of $\mathbf{\Psi}$ to shrink and its column space to expand, by the "amount" of $\mathcal{F}_k$. When $\lambda^\star w_k$ is positive, any vector $\mathbf{f}_{(k)}^\perp \in \mathcal{F}_k$ would satisfy

$$\mathbf{\Psi}\mathbf{f}_{(k)}^\perp = \mathbf{\Phi}\mathbf{f}_{(k)}^\perp + \sum_{l=1}^{M} \lambda^\star w_l \mathbf{I}_{(l)} \mathbf{f}_{(k)}^\perp = \lambda^\star w_k \mathbf{f}_{(k)}^\perp. \tag{4.25}$$

This means that $\mathbf{f}_{(k)}^\perp$ is always an eigenvector of $\mathbf{\Psi}$ with eigenvalue $\lambda^\star w_k$.

Since the linear equation in (4.24) is consistent and the null space $\mathcal{N}(\mathbf{\Psi})$ has been established in Proposition 4.3.2 and 4.3.3, we are ready to answer the third question:

Theorem 4.3.4. *The general solution form of* (4.24) *is*

$$\mathbf{f}^\star = \mathbf{\Psi}^\dagger \mathbf{b} + \sum_{k=1}^{K} \mathbf{f}_{(m_k)}^\perp, \tag{4.26}$$

in which $\mathbf{f}_{(m_k)}^\perp = \operatorname{col}(\mathbf{0}, \ldots, \mathbf{f}_{m_k}^\perp, \ldots, \mathbf{0})$ *satisfies* $(\mathbf{I}_{N_R} \otimes \mathbf{Q}\mathbf{G}_{m_k})\mathbf{f}_{m_k}^\perp = \mathbf{0}$ *and* $\{m_1, \ldots, m_K\}$ *is the set of all indexes such that* $\lambda^\star w_{m_k} = 0$. *The* $K+1$ *terms in* (4.26) *are orthogonal to each other. The first term,* $\mathbf{\Psi}^\dagger \mathbf{b}$, *is the solution that minimizes the transmit power of each relay simultaneously.*

Proof. Since $\mathbf{b} \in \mathcal{R}(\mathbf{\Psi})$ and $\mathbf{\Psi}\mathbf{\Psi}^\dagger$ is a projection matrix onto $\mathcal{R}(\mathbf{\Psi})$, we have $\mathbf{\Psi}(\mathbf{\Psi}^\dagger \mathbf{b}) = \mathbf{b}$. The general form in (4.26) follows immediately because $\mathcal{N}(\mathbf{\Psi}) = \mathcal{F}_{m_1} \oplus \cdots \oplus \mathcal{F}_{m_K}$ (*cf.* Proposition 4.3.3). Since $\mathcal{R}(\mathbf{\Psi})$, $\mathcal{F}_{m_1}, \ldots, \mathcal{F}_{m_K}$ are mutually orthogonal, the $K+1$ terms

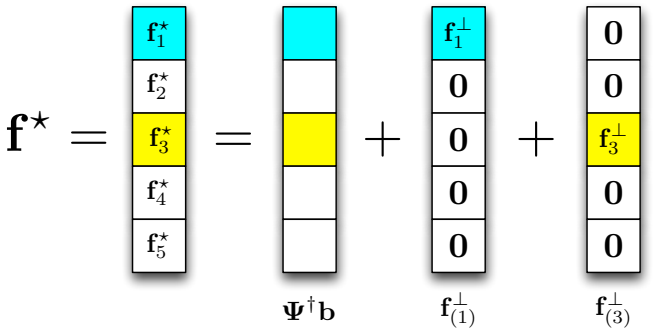

Fig. 4.3 A graphical example of the general solution in (4.26). There are five relays ($M = 5$) and $w_1 = w_3 = 0$.

in (4.26) are also orthogonal to each other. Let $\mathbf{f}^\star = \mathrm{col}(\mathbf{f}_1^\star, \ldots, \mathbf{f}_M^\star)$. For those k satisfying $\lambda^\star w_k > 0$, $\mathbf{f}_k^\star$ is unique; for those $k \in \{m_1, \ldots, m_K\}$, the transmit power of the kth relay is

$$\|\mathbf{I}_{(k)}\mathbf{f}^\star\|_2^2 = \|\mathbf{I}_{(k)}\mathbf{\Psi}^\dagger\mathbf{b}\|_2^2 + \|\mathbf{f}_{(k)}^\perp\|_2^2,$$

which can only be minimized by setting $\mathbf{f}_{(k)}^\perp = \mathbf{0}$, as shown in Fig. 4.3. Therefore among all the solutions, $\mathbf{\Psi}^\dagger\mathbf{b}$ minimizes the transmit power of each relay simultaneously. □

It appears that immediately after Theorem 4.3.1, we could have applied the pseudo-inverse to obtain the solution. This approach, however, would only guarantee that among all the solutions of (4.24), $\mathbf{\Psi}^\dagger\mathbf{b}$ minimizes the *sum* power of the relays, $\|\mathbf{f}^\star\|_2^2$. With the help of Proposition 4.3.2 and 4.3.3, we were able to prove a stronger conclusion: $\mathbf{\Psi}^\dagger\mathbf{b}$ minimizes the power of *each* relay *simultaneously*. Moreover, if $\mathbf{f}_k$ were defined as $\mathrm{vec}(\mathbf{F}_k)$ instead of $\mathrm{vec}(\mathbf{F}_k\mathbf{R}_k^{1/2})$, taking the pseudo inverse directly would not even minimize the sum power. This is because for Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$ (with $\mathbf{B}$ nonsingular), $(\mathbf{BAB})^\dagger \neq \mathbf{B}^{-1}\mathbf{A}^\dagger\mathbf{B}^{-1}$ except under special situations.

The main drawback of computing $\mathbf{f}^\star$ from (4.26) is that the dimension of $\mathbf{\Psi}$, $MN_R^2 \times MN_R^2$, is larger than those of the original matrices. We can use the relationships in (4.14) to simplify (4.26) as in the following corollary:

Corollary 4.3.5. *The minimum-norm solution of* (4.24) *can be expressed in an alternative form:*

$$\mathbf{f}_k^\star = (\mathbf{S}_k + \lambda^\star w_k \mathbf{I}_{N_R^2})^\dagger \mathbf{T}_k \mathbf{\Sigma}^{-1} \mathrm{vec}(\mathbf{I}_{N_S}), \tag{4.27}$$

where $\mathbf{\Sigma} \in \mathbb{C}^{N_S^2 \times N_S^2}$ *is defined as*

$$\mathbf{\Sigma} \triangleq \mathbf{R}_s^{-T} \otimes \mathbf{I}_{N_S} + \sum_{l=1}^{M} \mathbf{T}_l^H (\mathbf{S}_l + \lambda^\star w_l \mathbf{I}_{N_R^2})^\dagger \mathbf{T}_l. \tag{4.28}$$

Proof. See Appendix A.1. □

4.3.3 Optimal Solution

We now return to the KKT conditions in (4.23). The complementary slackness in (4.23d) indicates that either $\lambda^\star = 0$ and the constraint (4.23b) is inactive, or $\lambda^\star > 0$ and the constraint is tightly satisfied. Define the weighted sum power as a function of λ, *viz.*,

$$g(\lambda) \triangleq \mathbf{b}^H \mathbf{\Psi}(\lambda)^\dagger \mathbf{I}_{\text{sum}} \mathbf{\Psi}(\lambda)^\dagger \mathbf{b}, \tag{4.29}$$

where the explicit argument for $\mathbf{\Psi}$ is used to emphasize its dependence on λ. If $0 \leq g(0) \leq P_r$, the unconstrained solution satisfies the weighted sum power constraint (4.23d) and $\lambda^\star = 0$. Otherwise, $g(0) > P_r$ and $\lambda^\star > 0$ should be the implicit solution to the nonlinear equation

$$g(\lambda) = P_r. \tag{4.30}$$

The following proposition justifies the uniqueness of $\lambda^\star$:

Proposition 4.3.6. *If* $g(0) > 0$, $g(\lambda)$ *is a monotonically decreasing function of* $\lambda > 0$ *with* $\lim_{\lambda \to \infty} g(\lambda) = 0$.

Proof. See Appendix A.2. □

Up to now, the optimal solution has been expressed in closed forms (4.26) or (4.27), but the dual variable $\lambda^\star$ does not have an explicit formula. Numerical methods such as bisection or Newton's method [108] are necessary to solve the nonlinear equation (4.30). This, in fact, can be improved by allowing a complex scaling in the equalizer $\mathbf{Q}$. That is, we consider the set $\{\eta^{-1} e^{-j\phi} \mathbf{Q} | \eta > 0,\ 0 \leq \phi < 2\pi\}$, in which each member is a complex-scaled version of $\mathbf{Q}$. For different (η, ϕ), the optimal $\lambda^\star$ and $\mathbf{f}^\star$, and the corresponding minimum value of the MSE in (4.15) are also different. We are interested in a single (η_o, ϕ_o) leading to the smallest minimum MSE, so that any other member in the set can be replaced by

$\eta_o^{-1}e^{-j\phi_o}\mathbf{Q}$. Interestingly, for this special (η_o, ϕ_o), the optimal $\lambda^\star$, $\mathbf{f}^\star$ and the minimum MSE always have *explicit* formulas, as shown in the following theorem:

Theorem 4.3.7. *Any equalizer* $\mathbf{Q}$ *can be replaced by a complex-scaled version* $\eta_o^{-1}e^{-j\phi}\mathbf{Q}$ *so that:*

(1) The optimal solution is

$$\mathbf{f}^\star = \eta_o e^{j\phi}(\mathbf{\Phi} + \theta\mathbf{I}_{\text{sum}})^\dagger\mathbf{b}, \tag{4.31}$$

where $\theta \triangleq \text{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H)/P_r$, ϕ *can be an arbitrary number in* $[0, 2\pi)$, *and* $\eta_o > 0$ *is the unique number satisfying* $\mathbf{f}^{\star H}\mathbf{I}_{\text{sum}}\mathbf{f}^\star = P_r$, *that is,*

$$\eta_o = \sqrt{P_r/\mathbf{b}^H(\mathbf{\Phi} + \theta\mathbf{I}_{\text{sum}})^\dagger\mathbf{I}_{\text{sum}}(\mathbf{\Phi} + \theta\mathbf{I}_{\text{sum}})^\dagger\mathbf{b}}. \tag{4.32}$$

The optimal duality parameter is $\lambda_o^\star = \theta\eta_o^{-2}$.

(2) The minimum MSE with the equalizer $\eta_o^{-1}e^{-j\phi}\mathbf{Q}$, *i.e.,*

$$\text{MSE}_{\min} = \text{tr}(\mathbf{R}_s) - \mathbf{b}^H\big(\mathbf{\Phi} + \theta\mathbf{I}_{\text{sum}}\big)^\dagger\mathbf{b} \tag{4.33a}$$

$$= \text{vec}(\mathbf{I})^H\mathbf{\Sigma}^{-1}\text{vec}(\mathbf{I}), \tag{4.33b}$$

is always smaller than or equal to that with any other scaled equalizer $\eta^{-1}e^{-j\phi}\mathbf{Q}$ $(\eta > 0,\ 0 \le \phi < 2\pi)$, *including* $\mathbf{Q}$ *itself.*

Proof. With the equalizer $\eta^{-1}e^{-j\phi}\mathbf{Q}$, we rewrite the MSE function in (4.15) as

$$\begin{aligned}\text{MSE}(\mathbf{f}, \mathbf{Q}, \eta, \phi) = \eta^{-2}\mathbf{f}^{\star H}\mathbf{\Phi}\mathbf{f}^\star - \eta^{-1}e^{j\phi}\mathbf{f}^{\star H}\mathbf{b} - \eta^{-1}e^{-j\phi}\mathbf{b}^H\mathbf{f}^\star \\ + \text{tr}(\mathbf{R}_s) + \eta^{-2}\text{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H).\end{aligned} \tag{4.34}$$

The minimum-norm solution is obtained by replacing $\mathbf{Q}$ in (4.26) with $\eta^{-1}e^{-j\phi}\mathbf{Q}$:

$$\mathbf{f}^\star = \eta e^{j\phi}(\mathbf{\Phi} + \lambda^\star\eta^2\mathbf{I}_{\text{sum}})^\dagger\mathbf{b}. \tag{4.35}$$

The duality parameter $\lambda^\star$ should satisfy the KKT conditions in (4.23b), (4.23c) and (4.23d). If the unconstrained solution satisfies

$$\mathbf{f}^{\star H}\mathbf{I}_{\text{sum}}\mathbf{f}^\star\big|_{\lambda^\star=0} = \eta^2\mathbf{b}^H\mathbf{\Phi}^\dagger\mathbf{I}_{\text{sum}}\mathbf{\Phi}^\dagger\mathbf{b} \le P_r, \tag{4.36}$$

or equivalently, $\eta \leq \eta_c \triangleq \sqrt{P_r/(\mathbf{b}^H\mathbf{\Phi}^\dagger\mathbf{I}_{\text{sum}}\mathbf{\Phi}^\dagger\mathbf{b})}$, the constraint (4.23b) is inactive, *i.e.*, $\lambda^\star = 0$. Substituting (4.35) into the MSE expression in (4.34), we have

$$\text{MSE}_{\min}(\eta) = \text{tr}(\mathbf{R}_s) - \mathbf{b}^H\mathbf{\Phi}^\dagger\mathbf{b} + \eta^{-2}\text{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H). \tag{4.37}$$

If $\eta \geq \eta_c$, the constraint (4.23b) is tightly satisfied:

$$\eta^2\mathbf{b}^H\big(\mathbf{\Phi} + \lambda^\star\eta^2\mathbf{I}_{\text{sum}}\big)^\dagger\mathbf{I}_{\text{sum}}\big(\mathbf{\Phi} + \lambda^\star\eta^2\mathbf{I}_{\text{sum}}\big)^\dagger\mathbf{b} = P_r, \tag{4.38}$$

through which $\lambda^\star$ is an implicit function of η. Substituting the optimal solution in (4.35) into the MSE expression in (4.34) and using (4.38) to replace η^{-2}, we have

$$\begin{aligned}\text{MSE}_{\min}(\eta) = \text{tr}(\mathbf{R}_s) - \mathbf{b}^H&\big(\mathbf{\Phi} + \lambda^\star\eta^2\mathbf{I}_{\text{sum}}\big)^\dagger \\ &\times \big(\mathbf{\Phi} + (2\lambda^\star\eta^2 - \theta)\mathbf{I}_{\text{sum}}\big)\big(\mathbf{\Phi} + \lambda^\star\eta^2\mathbf{I}_{\text{sum}}\big)^\dagger\mathbf{b}.\end{aligned} \tag{4.39}$$

Up to now, the minimum MSE, which does not depend on ϕ, has been expressed as a function of η. From (4.37), $\text{MSE}_{\min}(\eta)$ is a monotonically decreasing function of η in the interval $[0, \eta_c]$ and therefore can only be minimized when $\eta \geq \eta_c$. Although $\lambda^\star$ does not have an explicit formula, the minimum MSE in (4.39) depends only on the product $\lambda^\star\eta^2 \triangleq \gamma$, which can take all nonnegative values. Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ be a set of orthonormal basis vectors for $\mathcal{N}(\mathbf{\Phi})$ and define $\mathbf{\Psi}_e \triangleq \mathbf{\Phi} + \gamma\mathbf{I}_{\text{sum}} + \sum_{k=1}^{p}\mathbf{u}_k\mathbf{u}_k^H$. Using the same tactics as in the proof of Proposition 4.3.6, we get the derivative of $\text{MSE}_{\min}(\gamma)$ as

$$\frac{\text{dMSE}_{\min}(\gamma)}{\text{d}\gamma} = 2(\gamma - \theta)\mathbf{b}^H\mathbf{\Psi}_e^{-1}I_{\text{sum}}\mathbf{\Psi}_e^{-1}\mathbf{I}_{\text{sum}}\mathbf{\Psi}_e^{-1}\mathbf{b}. \tag{4.40}$$

Obviously, $\text{MSE}_{\min}(\gamma)$ is monotonically decreasing if $0 \leq \gamma < \theta$, and monotonically increasing if $\gamma > \theta$. Therefore, $\gamma_o = \theta = \text{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H)/P_r$ is the unique solution to minimize (4.39) and it is straightforward to derive (4.31) and (4.33). □

We may visualize η^2 as a target signal power level at the destination and η^{-1} as an automatic gain control factor. For $\eta \leq \eta_c$, the power budget at the relays is sufficient to support the unconstrained optimal solution ($\lambda^\star = 0$). As seen in (4.37), the first part, $\text{tr}(\mathbf{R}_s) - \mathbf{b}^H\mathbf{\Phi}^\dagger\mathbf{b}$, does not depend on η. The second part, $\eta^{-2}\text{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H)$, decreases monotonically as a function of $\eta \leq \eta_c$, indicating weaker effects of the noise term $\mathbf{n}$ in

(4.3) on the decoding process. Once η exceeds the threshold η_c, the power budget becomes insufficient and therefore the power regularization term $\lambda^\star\eta^2\mathbf{I}_{\text{sum}}$ is introduced. This slightly increases the first part of (4.34) (the first four terms), but the overall MSE still decreases because the second part $\eta^{-2}\text{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H)$ is reduced by more. Nonetheless, there is a critical and therefore optimal η_o above which the latter cannot completely compensate for the former any more.

An alternative formulation is to introduce η as early as in the definition of the objective function in (4.6), which was used before for other relaying systems [22,67,97]. In this case, the objective function would be a convex function of $\mathbf{f}$, but not of both η and $\mathbf{f}$. Therefore, it does not formally guarantee optimality to set to zero the partial derivatives with respect to both η and $\mathbf{f}$. We also note that if $M = 1$ (a single relay), the optimal relaying matrix in (4.31) would be in agreement with the result in [22].

4.4 Per-Relay Power Constraints

Due to practical reasons such as the dynamic range of power amplifiers, it may sometimes be more appropriate to consider the per-relay power constraints. In this section, we study the optimality conditions and propose a power balancing algorithm to compute the optimal solution. Our analysis provides some insights into the power usage at the relays.

4.4.1 KKT Conditions and the Optimal Solution

The Lagrangian function for the relay optimization problem with the per-relay power constraints in (4.18) is given by

$$L(\mathbf{f},\boldsymbol{\lambda}) = \text{MSE}(\mathbf{f}) + \sum_{k=1}^{M}\lambda_k(\mathbf{f}^H\mathbf{I}_{(k)}\mathbf{f} - P_k), \tag{4.41}$$

where $\boldsymbol{\lambda} \triangleq \text{col}(\lambda_1,\cdots,\lambda_M)$. By comparing with (4.20), we note that many results for the weighted sum power constraint extend to the per-relay power constraints, simply by replacing λw_k with λ_k. Subsequently, we skip the details to focus on presenting the main results. Redefine $\boldsymbol{\Psi} \triangleq \boldsymbol{\Phi} + \sum_{k=1}^{M}\lambda_k^\star\mathbf{I}_{(k)}$. The optimal $\mathbf{f}^\star$ and its dual-optimal variables λ_k^*,

$1 \leq k \leq M$, satisfy the following KKT conditions:

$$\nabla_{\bar{\mathbf{f}}} L(\mathbf{f}, \boldsymbol{\lambda}^\star)|_{\mathbf{f}=\mathbf{f}^\star} = \boldsymbol{\Psi}\mathbf{f}^\star - \mathbf{b} = \mathbf{0}, \tag{4.42a}$$

$$\mathbf{f}^{\star H}\mathbf{I}_{(k)}\mathbf{f}^\star \leq P_k, \tag{4.42b}$$

$$\lambda_k^\star \geq 0, \tag{4.42c}$$

$$\lambda_k^\star(\mathbf{f}^{\star H}\mathbf{I}_{(k)}\mathbf{f}^\star - P_k) = 0, \tag{4.42d}$$

for all $1 \leq k \leq M$. Akin to Theorem 4.3.4 and Corollary 4.3.5, the minimum-norm solution to (4.42a) is $\mathbf{f}^\star = \boldsymbol{\Psi}^\dagger\mathbf{b}$ with an alternative form $\mathbf{f}_k^\star = (\mathbf{S}_k + \lambda_k^\star\mathbf{I}_{N_R^2})^\dagger\mathbf{T}_k\boldsymbol{\Sigma}^{-1}\mathrm{vec}(\mathbf{I}_{N_S})$, where $\boldsymbol{\Sigma} \triangleq \mathbf{R}_\mathbf{s}^{-T} \otimes \mathbf{I}_{N_S} + \sum_{l=1}^M \mathbf{T}_l^H(\mathbf{S}_l + \lambda_l^\star\mathbf{I}_{N_R^2})^\dagger\mathbf{T}_l$.

The only difference from the weighted sum power case is the existence of *multiple* dual variables and complementary slackness conditions. This requires algorithms that are more sophisticated than bisection or Newton's method, *e.g.*, interior-point methods such as the path-following and the primal-dual methods [108, 113]. Software packages for these algorithms, *e.g.*, Gurobi, CPLEX and SeDuMi, are available and can be used in Matlab via YALMIP [115] or CVX [116].

Here, it is reasonable to consider the same complex-scaled equalizer $\eta^{-1}e^{-j\phi}\mathbf{Q}$ as in Section 4.3.3, which results in the following theorem:

Theorem 4.4.1. *Any equalizer* $\mathbf{Q}$ *can be replaced by a complex-scaled version* $\eta_o^{-1}e^{-j\phi}\mathbf{Q}$ *so that:*

(1) The optimal solution is

$$\mathbf{f}^\star = \eta_o\Big(\boldsymbol{\Phi} + \sum_{k=1}^M \gamma_k^o\mathbf{I}_{(k)}\Big)^\dagger\mathbf{b}, \tag{4.43}$$

where $\phi \in [0, 2\pi)$, $\gamma_k^o \triangleq \lambda_k^\star\eta_o^2$ $(1 \leq k \leq M)$ *satisfies*

$$\sum_{k=1}^M \gamma_k^o P_k = \mathrm{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H) \tag{4.44}$$

and $\eta_o > 0$ *is the unique positive number such that* $\gamma_k^o(\mathbf{f}^{\star H}\mathbf{I}_{(k)}\mathbf{f}^\star - P_k) = 0$ *for* $1 \leq k \leq M$.

(2) The minimum MSE with the equalizer $\eta_o^{-1}e^{-j\phi}\mathbf{Q}$ *is*

$$\mathrm{MSE}_{\min}(\eta_o) = \mathrm{tr}(\mathbf{R}_s) - \mathbf{b}^H\Big(\mathbf{\Phi} + \sum_{k=1}^{M}\gamma_k^o\mathbf{I}_{(k)}\Big)^{\dagger}\mathbf{b}. \tag{4.45}$$

Proof. For any $\eta > 0$ and $0 \le \phi < 2\pi$, the optimal solution is $\mathbf{f}^\star = \eta e^{j\phi}\mathbf{\Psi}^{\dagger}\mathbf{b}$, where $\mathbf{\Psi} \triangleq \mathbf{\Phi} + \sum_{k=1}^{M}\lambda_k^\star\eta^2\mathbf{I}_{(k)}$, and the dual-optimal variables $\lambda_1^\star, \ldots, \lambda_M^\star$ are implicit functions of η through (4.42b)-(4.42d). The minimum MSE is also a function of $\eta > 0$, *viz.*,

$$\mathrm{MSE}_{\min}(\eta) = \mathrm{tr}(\mathbf{R}_s) - \mathbf{b}^H\mathbf{\Psi}^{\dagger}\mathbf{b} + \eta^{-2}\mathrm{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H) - \sum\nolimits_{k=1}^{M}\lambda_k^\star P_k. \tag{4.46}$$

Next, we minimize (4.46) over $\eta > 0$. With similar argument to that in Section 4.3.3, there exists an η_c such that for all $\eta \le \eta_c$, $\lambda_1^\star = \cdots = \lambda_M^\star = 0$. As a result, $\mathrm{MSE}_{\min}(\eta)$ is a monotonically decreasing function in $(0, \eta_c]$ according to (4.46). This means that (4.46) is minimized only when $\eta > \eta_c$. In this situation, at least one of the dual variables is nonzero and hence $\sum_{k=1}^{M}\lambda_k^\star P_k \neq 0$. For convenience, rewrite the complementary slackness from (4.42d) as

$$\lambda_k^\star P_k = \lambda_k^\star\eta^2\mathbf{b}^H\mathbf{\Psi}^{-1}\mathbf{I}_{(k)}\mathbf{\Psi}^{-1}\mathbf{I}. \tag{4.47}$$

Adding from $k = 1$ to M, we have

$$\eta^{-2} = \frac{\sum_{k=1}^{M}\lambda_k^\star\mathbf{b}^H\mathbf{\Psi}^{-1}\mathbf{I}_{(k)}\mathbf{\Psi}^{-1}\mathbf{b}}{\sum_{k=1}^{M}\lambda_k^\star P_k}. \tag{4.48}$$

Substituting (4.47) and (4.48) into (4.46) and using similar techniques to those in the proof of Theorem 4.3.7, we can prove that the optimal η_o, the corresponding $\lambda_k^\star$ and $\gamma_k^o = \lambda_k^\star\eta_o^2$ satisfy (4.44), and the minimum MSE takes the form of (4.46). □

Eq. (4.44) provides an elegant relationship between different γ_k^o. These parameters serve as regularization terms that control the transmit power of the relays. When the power budget P_k is higher, the required regularization tends to be lower and the value of γ_k^o tends to be smaller. Although complex scaling does not lead to closed-form expressions for the *individual* parameters $\{\gamma_k^o\}$, the results in (4.44), (4.43) and (4.45) lay the foundation for the power balancing algorithm proposed in the next subsection.

4.4.2 Power Balancing

The optimal solutions, (4.31) under the weighted sum power constraint and (4.43) under the per-relay power constraints, share some common structure. In particular, they are identical if the weights $\{w_k\}$ and P_r are chosen to satisfy

$$w_k\theta = w_k\frac{\mathrm{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H)}{P_r} = \gamma_k^o \quad \text{and} \quad P_r = \sum_{k=1}^{M} w_k P_k. \tag{4.49}$$

The minimum MSE, (4.33) and (4.45), would also be equal. In other words, if we know this *equivalent* weighted sum power constraint, the optimal solution is immediately available from Theorem 4.3.7. Of course, scaling all w_k simultaneously by a common positive factor does not alter the optimal solution.

With this in mind, we propose a power balancing algorithm which finds these weights iteratively. The initial weights are all set to 1. In each iteration, the algorithm computes the optimal relaying matrices with the previous weights, compares the actual power of the relays with the per-relay power constraints, and adjusts the weights accordingly. If the actual power of the kth relay is higher than P_k, the weight w_k is increased and vice versa. The algorithm stops when all the constraints are satisfied. These steps are summarized in Algorithm 4.1.

Algorithm 4.1: Power Balancing

Initiate the counter $m = -1$;

Initiate the weights: $w_1(0) = \cdots = w_M(0) = 1$;

repeat

Add the counter $m \leftarrow m + 1$;

Compute the weighted sum power:

$$P_r(m) = \sum\nolimits_{k=1}^{M} w_k(m)P_k; \tag{4.50}$$

Compute the optimal $\mathbf{f}^\star(m)$ from (4.31) or (4.27);

Compute $P_k(m) = \mathbf{f}^\star(m)^H\mathbf{I}_{(k)}\mathbf{f}^\star(m)$, $1 \le k \le M$;

Update the weights for $1 \le k \le M$:

$$w_k(m+1) \leftarrow w_k(m)\frac{P_k(m)}{P_k}; \tag{4.51}$$

until $\max\left(P_1(m)/P_1, \cdots, P_M(m)/P_M\right) \le 1$.

4.4.3 Remarks on Power Usage

With the per-relay power constraints, the optimal relaying strategy may not use the maximum power at some relays. We show this for a simplified single-antenna case but our analysis also captures the essence of multi-antenna systems. The matrices/vectors in the signal model in Section 4.2 become scalars, represented by the corresponding lowercase italic letters. For convenience, we assume g_k, f_k and h_k $(1 \le k \le M)$ are all real positive numbers and extension to the complex scenarios is straightforward.

Define $\alpha \triangleq \sum_{l=1,l\neq k}^{M} g_l f_l h_l$ and $w \triangleq \sum_{l=1,l\neq k}^{M} g_l f_l w_l$ with $\mathrm{var}(w) = \sigma_w^2$. Without the kth relay, the signal received by the destination would be

$$r = \alpha s + w + n \tag{4.52}$$

and the SNR at the destination would be $\mathrm{SNR}_d \triangleq \alpha^2\sigma_s^2/(\sigma_w^2 + \sigma_n^2)$. With the kth relay, the signal would be

$$r' = (\alpha + g_k f_k h_k)s + g_k f_k w_k + w + n \tag{4.53}$$

and the SNR would be a function of f_k

$$\mathrm{SNR}'_d = \mathrm{SNR}_d \, \frac{(1 + g_k f_k h_k/\alpha)^2}{1 + g_k^2 f_k^2 \sigma_k^2/(\sigma_w^2 + \sigma_n^2)}, \tag{4.54}$$

where the second operand ($\triangleq \kappa(f_k)$) is the *gain or penalty* due to the kth relay, depending on whether it is larger than or smaller than 1.

By taking the first-order derivative, we know that $\kappa(f_k)$ is strictly monotonically increasing when

$$0 \le f_k < f_o \triangleq \frac{h_k(\sigma_w^2 + \sigma_n^2)}{\alpha g_k \sigma_k^2} \tag{4.55}$$

and strictly monotonically decreasing when $f_k > f_o$. As $f_k \to \infty$, the limit would be

$$\kappa(\infty) \triangleq \lim_{f_k \to \infty} \kappa(f_k) = \frac{h_k^2 \sigma_s^2 / \sigma_k^2}{\alpha^2 \sigma_s^2 / (\sigma_w^2 + \sigma_n^2)} = \frac{\mathrm{SNR}_r}{\mathrm{SNR}_d}, \tag{4.56}$$

where $\mathrm{SNR}_r \triangleq h_k^2 \sigma_s^2 / \sigma_k^2$ is the SNR at the kth relay. Since the power transmitted by the relay is proportional to f_k^2, the above properties of $\kappa(f_k)$ lead to several interesting conclusions:

(1) If $\mathrm{SNR}_r \geq \mathrm{SNR}_d$, $\kappa(\infty) \geq 1$ according to (4.56). Since $\kappa(0) = 1$, and $\kappa(f_k)$ increases in $[0, f_o]$ and decreases in $[f_o, \infty]$, $\kappa(f_k) > 1$ always holds in $(0, \infty)$, implying that the system always benefits from the use of the kth relay, no matter how much power the relay transmits. However, it is *not necessarily* better to use more power. Any $f_k > f_o$ would not be as good as f_o.

(2) If $\mathrm{SNR}_r < \mathrm{SNR}_d$, $\kappa(\infty) < 1$ and there exists an f_c such that $\kappa(f_c) = 1$. Hence, $\kappa(f_k) > 1$ in the interval $(0, f_c)$, which means that the kth relay can still contribute to the SNR at the destination as long as it reduces its transmit power to a level low enough. The reason for this is that the signal components are added coherently, whereas the noise components are not.

(3) If the value of f_k corresponding to the power constraint P_k falls in the interval $[0, f_o]$, the relay would use the maximum amount of power; otherwise, it would use only a portion ($f_k = f_o$). It is not justified to turn off a relay completely.

In practice, it is a waste of resources if a relay transmits only a small amount of power. Thanks to the randomness of the channels and users, this problem with the ideal narrowband configuration is probably not as important in practice. Firstly, most modern communication systems are based on a multicarrier scheme such as OFDM. A relay station may transmit less power on one subcarrier but more on another, and so variations of transmit power between different relays is small. Secondly, the multiple relays are simultaneously serving several randomly located users (with different subcarriers or time intervals). This will further reduce the disparity between the transmit power of different relays. Lastly, if the expected transmit power of a particular relay is abnormally small, the problem likely comes from inappropriate network layout and the relays should be relocated instead.

4.5 The Optimal Equalizer

In this section, we consider the joint design of the MIMO equalizer $\mathbf{Q}$ and the relaying matrices (or equivalently $\mathbf{f}$), under the weighted sum power constraint. For any $\mathbf{Q}$, the optimal $\mathbf{f}^\star$ is in the form of (4.31); for any $\mathbf{f}$ (or $\mathbf{F}$), the optimal equalizer $\mathbf{Q}^\star$ is the MMSE equalizer

$$\mathbf{Q}^\star = \mathbf{R}_{rs}^H \mathbf{R}_r^{-1}, \tag{4.57}$$

where $\mathbf{R}_{rs} \triangleq \mathrm{E}\{\mathbf{r}\mathbf{s}^H\} = \mathbf{G}\mathbf{F}\mathbf{H}\mathbf{R}_s$ and $\mathbf{R}_r \triangleq \mathrm{E}\{\mathbf{r}\mathbf{r}^H\}$.

This observation suggests a block coordinate descent method. The algorithm starts from an initial $\mathbf{Q}_0$ and repeats the following steps: it first updates $\mathbf{f}$ using (4.31) while fixing $\mathbf{Q}$, and then calculates $\mathbf{Q}$ as in (4.57) while holding $\mathbf{f}$ constant. Thanks to the optimality in each step, the (bounded) sequence of MSE values is monotonically non-increasing, which must converge. As a result, the block coordinate descent algorithm is guaranteed to converge to a local optima. This idea is widely used in the literature such as [65, 117]. We also note that the design of a precoder $\mathbf{B}$ is also possible through this framework. This is done by replacing all the $\mathbf{H}_k$ with $\mathbf{H}_k\mathbf{B}$ in the previous sections.

The other approach is to consider the joint design as a two-step process. The first step is to design $\mathbf{f}$ as a function of $\mathbf{Q}$, which is what we have done in Section 4.3. After this, the second step is to optimize $\mathbf{Q}$ to further minimize the MSE in (4.33). This approach handles the constraints in the convex problem (the first step), so that the remaining problem, though still non-convex, is an unconstrained one.[4] The line search algorithms are readily applicable to find a local minima. Beginning with $\mathbf{Q}_0$, these methods generate a sequence of iterates $\{\mathbf{Q}_n\}_{n=0}^{\infty}$ until a solution has been approximated with sufficient accuracy. Specifically, these algorithms choose a direction $\Delta\mathbf{Q}_n$ and search along this direction from the current $\mathbf{Q}_n$ for a new iterate $\mathbf{Q}_{n+1}$ with a lower MSE value. The distance to move along $\mathbf{Q}_n$ should satisfy criteria such as Wolfe's conditions [108]. In particular, the steepest descent method uses the opposite direction of the gradient (see Appendix A.3 for derivation), *viz.*,

$$\begin{aligned}\Delta\mathbf{Q}_n &= -\nabla_{\bar{\mathbf{Q}}}\mathrm{MSE}_{\min}|_{\mathbf{Q}=\mathbf{Q}_n} \\ &= -\eta_n^{-2}\mathbf{Q}_n(\mathbf{G}\mathbf{F}_{(n)}\mathbf{R}_w\mathbf{F}_{(n)}^H\mathbf{G}^H + \mathbf{R}_n) + \eta_n^{-1}\mathrm{unvec}\big(\mathbf{\Sigma}_n^{-1}\mathrm{vec}(\mathbf{I})\big)\mathbf{H}^H\mathbf{F}_{(n)}^H\mathbf{G}^H, \end{aligned} \tag{4.58}$$

[4] An alternative, usually more popular, approach is to first set the equalizer $\mathbf{Q}$ as the MMSE equalizer. After substituting this optimal $\mathbf{Q}$ into (4.6), the MSE becomes a function of the matrices $\{\mathbf{F}_k\}$. The resulting problem is, however, not only non-convex but also with constraints.

where $\mathbf{F}_{(n)}$ is the optimal (block-diagonal) relaying matrix under $\mathbf{Q}_n$, and $\mathbf{\Sigma}_n$ and η_n are the corresponding intermediate matrices/variables when computing $\mathbf{F}_{(n)}$, *cf.* (4.5), (4.11) and (4.31). The steepest descent method is summarized in Algorithm 4.2. Numerical results in Sec. 4.7 will show that this method converges much faster than block coordinate descent.

Algorithm 4.2: Steepest Descent

Choose the weights w_k, $1 \le k \le M$, P_r and $\epsilon > 0$;
Choose $\bar{\alpha} > 0$, $\rho, c \in (0,1)$; {Line search parameters.}
Initiate the counter $n = -1$ and the equalizer $\mathbf{Q}_0$;
repeat
  Increment counter $n \leftarrow n + 1$;
  Compute $\Delta\mathbf{Q}_n$ from (4.58);
  Set $\alpha \leftarrow \bar{\alpha}$;
  repeat
    $\alpha \leftarrow \rho\alpha$; {Backtracking line search.}
  until $\mathrm{MSE}_{\min}(\mathbf{Q}_n + \alpha\Delta\mathbf{Q}_n) \le \mathrm{MSE}_{\min}(\mathbf{Q}_n) - c\alpha\|\Delta\mathbf{Q}_n\|_{\mathrm{F}}^2$;
  Update $\mathbf{Q}_{n+1} \leftarrow \mathbf{Q}_n + \alpha\Delta\mathbf{Q}_n$;
until $\|\Delta\mathbf{Q}_n\|_{\mathrm{F}}^2 < \epsilon$.

As for the per-relay power constraints, Algorithm 4.1 is still applicable except that in each iteration, $\mathbf{Q}$ is updated along with $\mathbf{f}$ using the above methods.

4.6 Implementation Issues and Complexity

We discuss implementation issues for the proposed algorithms, including the requirements on communication and computing resources. Firstly, an important feature of the optimal methods is that they only require local CSI knowledge and a little additional shared information. As seen from (4.27), all the kth relay needs to know, in addition to its own backward channel $\mathbf{H}_k$ and forward channel $\mathbf{G}_k$, is the vector $\mathbf{\Sigma}^{-1}\mathrm{vec}(\mathbf{I})$ (of size $N_S^2 \times 1$). Thanks to this attractive feature, these methods naturally lend themselves to distributed implementations. One possible implementation scheme is that a fusion center collects the channel matrices, computes this vector and feeds it back to the relays via broadcasting. An alternative way is to compute the relaying matrices also at the fusion center. The fusion

center can be the destination or one of the relays. Some dedicated resources are required but for a small number of relays (say 2 or 3), the overall complexity will be manageable.

Secondly, the existence of closed-form expressions such as (4.27) results in relatively low computational complexity. The major computing task consists of two parts: evaluating $\mathbf{\Sigma}^{-1}\text{vec}(\mathbf{I})$ and $\{\mathbf{f}_k\}$. The power balancing algorithm in Algorithm 4.1 for the per-relay constraints and the algorithms for the equalizer $\mathbf{Q}$ (*cf.* Section 4.5) are primarily composed of repetitions of these operations. Although vectorization increases the dimensions of the matrices and vectors in the system model (*cf.* (4.13)), the complexity is not notably higher thanks to the properties of the SVDs and EVDs for Kronecker products. For example, we have

$$\left.\begin{aligned}\mathbf{X} &= \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^H\\ \mathbf{Y} &= \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^H\end{aligned}\right\} \Rightarrow \mathbf{X}\otimes\mathbf{Y} = (\mathbf{U}_1\otimes\mathbf{U}_2)(\mathbf{S}_1\otimes\mathbf{S}_2)(\mathbf{V}_1\otimes\mathbf{V}_2)^H,$$

which is essentially the SVD of $\mathbf{X}\otimes\mathbf{Y}$ except that the singular values are not sorted in descending order. A similar property holds for the EVD. Consequently, the EVD of the pseudo inverse $(\mathbf{S}_k+\lambda w_k\mathbf{I})^\dagger = \mathbf{U}_k\mathbf{\Lambda}_k\mathbf{U}_k^H$ can be obtained based on those of $(\mathbf{R}_{\mathbf{x}_k}^{-1/2}\mathbf{R}_{\mathbf{w}_k}\mathbf{R}_{\mathbf{x}_k}^{-1/2})^T$ and $\mathbf{G}_k^H\mathbf{Q}^H\mathbf{Q}\mathbf{G}_k$. Then, two matrix multiplications (not including those involving diagonal matrices) are needed to compute $\mathbf{T}_k^H(\mathbf{S}_k+\lambda w_k\mathbf{I})^\dagger\mathbf{T}_k = (\mathbf{T}_k^H\mathbf{U}_k)\mathbf{\Lambda}_k(\mathbf{T}_k^H\mathbf{U}_k)^H$. One additional matrix multiplication is necessary to compute $(\mathbf{S}_k+\lambda w_k\mathbf{I})^\dagger\mathbf{T}_k = \mathbf{U}_k\mathbf{\Lambda}_k(\mathbf{T}_k^H\mathbf{U}_k)^H$. Subsequently, we compute the sum matrix $\mathbf{\Sigma}$ and solve the linear equation $\mathbf{\Sigma}\mathbf{x} = \text{vec}(\mathbf{I})$ to get $\mathbf{\Sigma}^{-1}\text{vec}(\mathbf{I})$. In the end, $\mathbf{f}_k$ is obtained from (4.27), which requires only matrix-vector multiplications. In summary, the major operations include

- $2M$ EVDs of matrices with dimension $N_R\times N_R$;
- $3M$ matrix multiplications involving matrices of dimension $N_S^2\times N_R^2$, $N_R^2\times N_R^2$ or $N_R^2\times N_S^2$.
- solving one linear equation of size $N_S^2\times N_S^2$.

Thirdly, the proposed algorithms in Section 4.3 and 4.4 minimize the MSE over not only the relaying vector $\mathbf{f}$, but also the scaling factor $\eta e^{j\phi}$. In contrast, the interior-point methods can merely optimize $\mathbf{f}$ for a single η because the problem is not convex if $\mathbf{f}$ and η are simultaneously considered. To obtain the same result as our algorithms do, the interior-point methods have to run for different $\eta > 0$, which further increases their complexity.

In summary, the system complexity is well manageable for multi-antenna systems if the number of relays is small. The benefits brought by a three-relay configuration can be remarkable as shown by the simulation results in Sec. 4.7 and previous publications such as [13, 59].

4.7 Numerical Results

In this section, we first investigate the effects of channel gains on the power allocation among relays under the weighted sum power constraint. Next, we verify the convergence behaviors of the proposed iterative algorithms, including power balancing for the per-relay power constraints, block coordinate descent and steepest descent for the equalizer. In the end, we compare the BER results of the proposed designs and previous strategies.

The following assumptions are made. The variances of $\mathbf{s}$, $\mathbf{w}_k$ and $\mathbf{n}$ are respectively $\mathbf{R}_s = \sigma_s^2\mathbf{I}$, $\mathbf{R}_{w_k} = \sigma_w^2\mathbf{I}_{N_R}$ and $\mathbf{R}_n = \sigma_n^2\mathbf{I}_{N_D}$ and the covariance matrices of the noise terms have been normalized: $\sigma_w^2 = \sigma_n^2 = 1$. It is convenient to introduce two SNR parameters as follows. The first SNR is defined as $\rho_1 \triangleq \sigma_s^2/\sigma_w^2$, i.e. the ratio of transmitted signal power per antenna to the received noise power per antenna. The second SNR, defined in terms of the sum power $P_R \triangleq \sum_{k=1}^{M} P_k$ as $\rho_2 \triangleq P_R/(MN_R\sigma_n^2)$, gives the ratio of average transmitted power per relay antenna to the power of the noise induced at the individual destination antennas. In our simulations, the channel matrices have independent and identically distributed entries. Each entry is a zero mean circular symmetric complex Gaussian variable with unit variance.

4.7.1 Weighted Sum Power Constraint: Power Allocation

To study the effects of channel gains on the power allocation among the relays, we consider a 1S-2R-1D system with $N_S = N_R = N_D = 2$, $\rho_1 = 20$dB and $\mathbf{Q} = \mathbf{I}$. In particular, we multiply $\mathbf{H}_1$ by α which represents a relative channel gain, and multiply $\mathbf{G}_1$ by β which is independent from α. Then, the optimal relaying matrices and the transmit power of the relays (P_1', P_2') are all functions of these relative gains. A randomly generated set of channel matrices are used and the other parameters are chosen as $w_1 = w_2 = 1$ and $P_r = 400$. As shown by the contours for P_1' in Fig. 4.4, if β is fixed, the larger α becomes, the more power is allocated to the first relay; if α is held constant, the larger β comes, the less power to

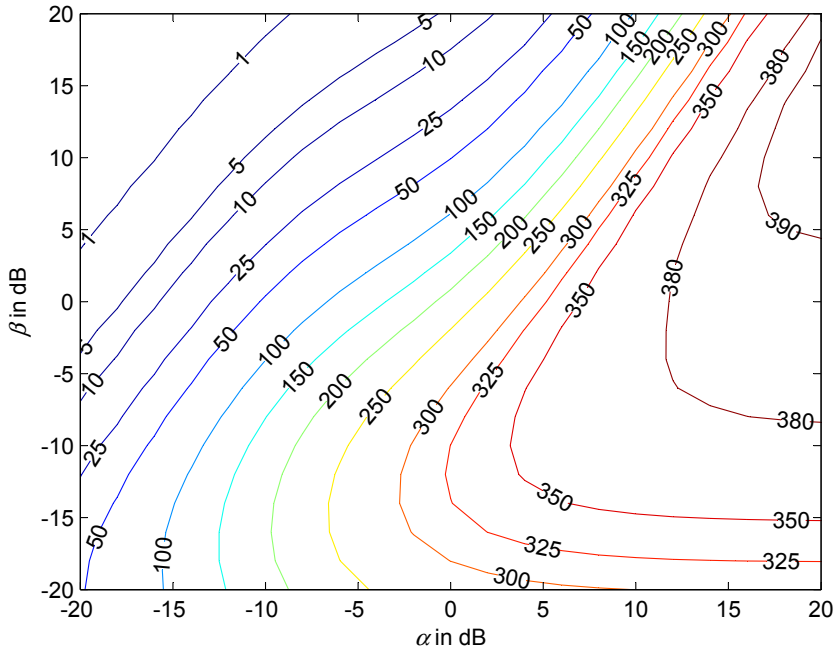

Fig. 4.4 Contours of the power of the first relay P_1' versus relative channel gains α and β

the first relay. In other words, the optimal relaying strategy tends to allocate more power to the relay with better backward channel and/or worse forward channel.

To observe the effects of the weights, we adjust the value of $0 \leq w_1 \leq 2$, let $w_2 = 2 - w_1$, and compute the optimal relaying matrices and the powers, P_1' and P_2'. The same set of channel matrices are used in the simulations, possibly with $\alpha = 6$dB or $\beta = 6$dB. Fig. 4.5 illustrates the relationship between the power ratio, P_1'/P_2', and the weight ratio w_1/w_2. As one would expect, as w_1/w_2 increases, the power ratio decreases.

Channel fading also plays an important role in power allocation. We generate 10^5 channel instances, obtain the corresponding optimal solutions and then compute the ratio between the power of the first relay and the total power, $P_1'/(P_1' + P_2')$. The other settings are: $w_1 = w_2 = 1$ and $P_r = 400$. Fig. 4.6 shows the histograms of this ratio, with the x axis representing the ratio and the y axis denoting the number of occurrences. Any value between 0 and 1 is possible and the histogram is close to a parabola if $\alpha = \beta = 0$dB. If $\mathbf{H}_1$ is of better quality than $\mathbf{H}_2$ ($\alpha = 6$dB), the histogram shifts rightward, implying that the optimal solution tends to allocate more power to the first relay; if $\mathbf{G}_1$ is of better

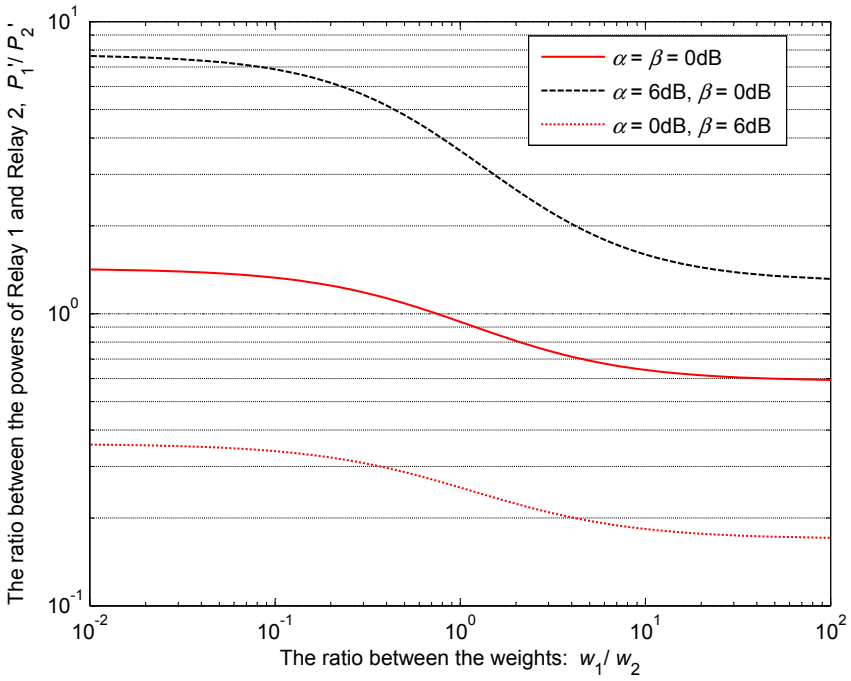

Fig. 4.5 Effects of w_1 and w_2 on the ratio between the powers of two relays

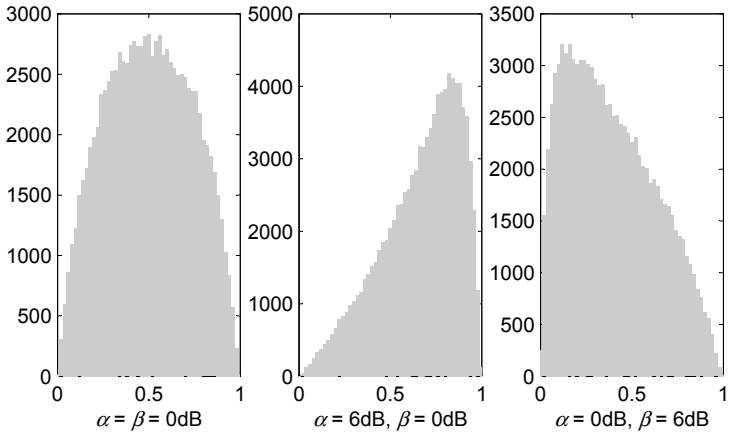

Fig. 4.6 Histograms of the ratio between the power of the first relay P_1' and the total power $P_1' + P_2'$. The x-axis represents the power ratio and the y-axis denotes the number of occurrences.

quality than $\mathbf{G}_2$ ($\beta = 6$dB), the histogram shifts rightward. This is in accordance with our previous conclusion drawn from Fig. 4.4.

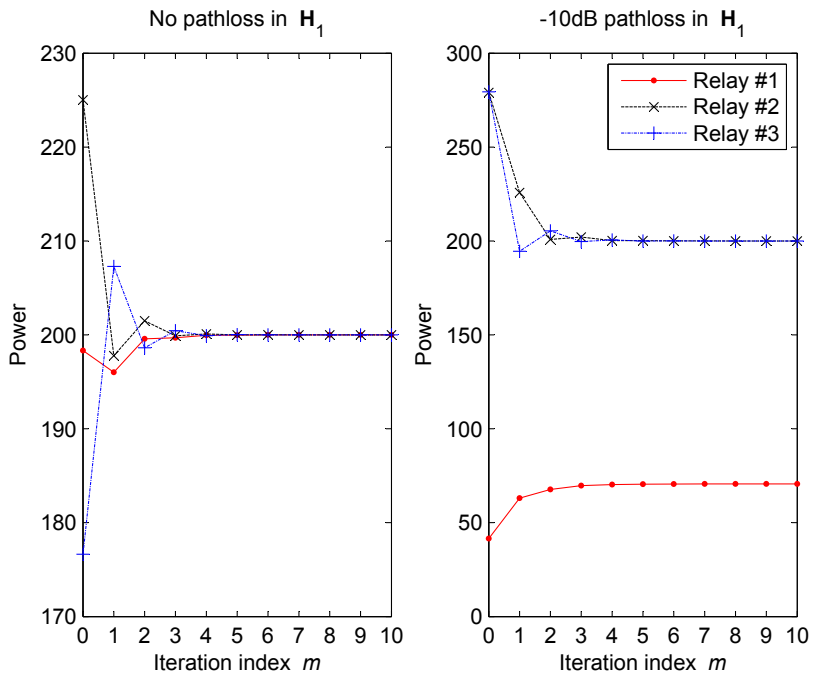

Fig. 4.7 Convergence behaviours of Algorithm 4.1: transmit power of the relays

4.7.2 Convergence of Iterative Algorithms

Power balancing for the per-relay power constraints. We study the convergence behaviors of Algorithm 4.1 base on a 1S-3R-1D system with $N_S = N_R = N_D = 4$, $\rho_1 = 20$dB and the power constraints $P_1 = P_2 = P_3 = 200$. For a particular representative channel instance (randomly generated), Fig. 4.7 plots the power of the individual relays versus the iteration index m. Algorithm 4.1 usually converges within about 5 steps.

Our analysis in Section 4.4.3 leads to the conclusion that the optimal relaying strategy does not necessarily use the maximum amount of power at some relays. This is verified by the right subfigure in Fig. 4.7, where the channel matrices used are the same as those in the left, except that $\mathbf{H}_1$ has a relative path loss of $\alpha = -10$dB. We note that with optimal relaying strategy, the first relay is neither using the maximum power, 200, nor being turned off completely.

Equalizer design. We study the convergence behaviors of the block coordinate descent method and the steepest descent method proposed in Section 4.5, based on the same settings as above. The weights are $w_1 = w_2 = w_3 = 1$ and $P_r = 600$. The convergence behaviors

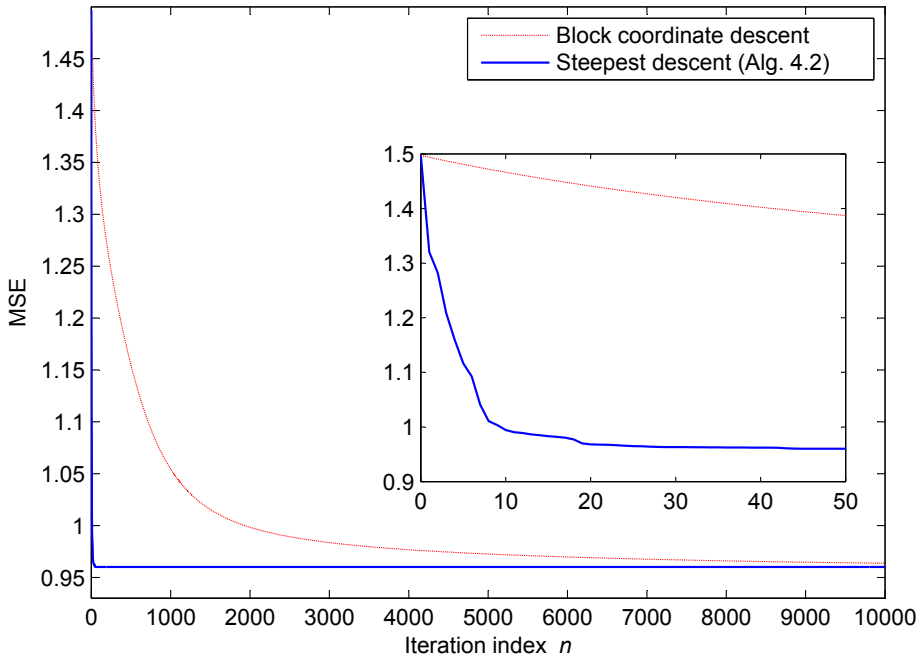

Fig. 4.8 Speed of convergence for the joint design: block coordinate descent and steepest descent (Algorithm 4.2)

are shown in Fig. 4.8 for a representative channel instance. The steepest descent method converges significantly faster than the block coordinate descent method. For the former, the MSE comes close to the locally optimal value after only 10 line-searches, whereas for the latter, it takes several thousand iterations. One interesting observation is that as far as we could verify, the optimal equalizer $\mathbf{Q}^\star$ does not depend on the algorithm used or the initial $\mathbf{Q}_0$ (except for a linear scaling factor). This suggests that the solution so obtained may be *globally* optimal, though it seems very difficult to prove and remains a conjecture for now because the expression of the Hessian is rather involved.

4.7.3 BER Performance

In this subsection, we compare the BER performance of the following relaying strategies:

(1) Simplistic AF, $\mathbf{F}_k \propto \mathbf{I}$, [13];

(2) MMSE-MMSE [13, 93];

(3) CMMSE-MMSE [93];

(4) Gradient-based MMSE [60];

(5) Proposed method (sum power constraint, $\mathbf{Q} = \mathbf{I}$);

(6) Proposed method (per-relay power constraints, $\mathbf{Q} = \mathbf{I}$);

(7) Proposed joint design of the relaying matrices and the equalizer (sum power constraint);

(8) Relay selection based on the JMMSE strategy [31].

For methods (1), (2), (3), (4) and (6), the total power is evenly split between different relays. For the selection-based strategy, the total power is allocated only to the single relay that would result in the minimum MSE based on the JMMSE relaying strategy [31]. We consider a 1S-3R-1D system with $N_S = N_R = N_D = 4$. In the simulations, each source antenna transmits independent uncoded 16-PSK symbol streams. The relay stations apply one of the above relaying schemes to their input signals and retransmit them. The destination applies a linear MMSE equalizer and then employs single-stream maximum likelihood decoding. The BER values are averaged over channel realizations.

First, we set ρ_2 to 20dB and vary ρ_1 between 5dB and 25dB. Then, we set $\rho_1 = 20$dB and vary ρ_2. The BER values are plotted in Fig. 4.9 and 4.10. As explained earlier, SAF cannot achieve distributed array gain and accordingly has the worst performance. The heuristic strategies including MMSE-MMSE and CMMSE-MMSE perform better. The gradient-based MMSE method provides further gain especially under low-to-mid SNR levels. The proposed MMSE-based strategies, (5), (6) and (7), outperform the above ones by large gaps. The choice depends on the compromise between performance and complexity: the joint design leads to lower BER but comes with higher computational complexity. It is worth mentioning that the proposed strategies are much superior to the selection-based one, which justifies the use of multiple relays.

4.8 Summary

In this chapter, we have considered the MMSE-based joint design of the multiple relaying matrices. Under the weighted sum power constraint, we derived closed-form expressions for the optimal relaying matrices. The optimal strategy tends to allocate more power

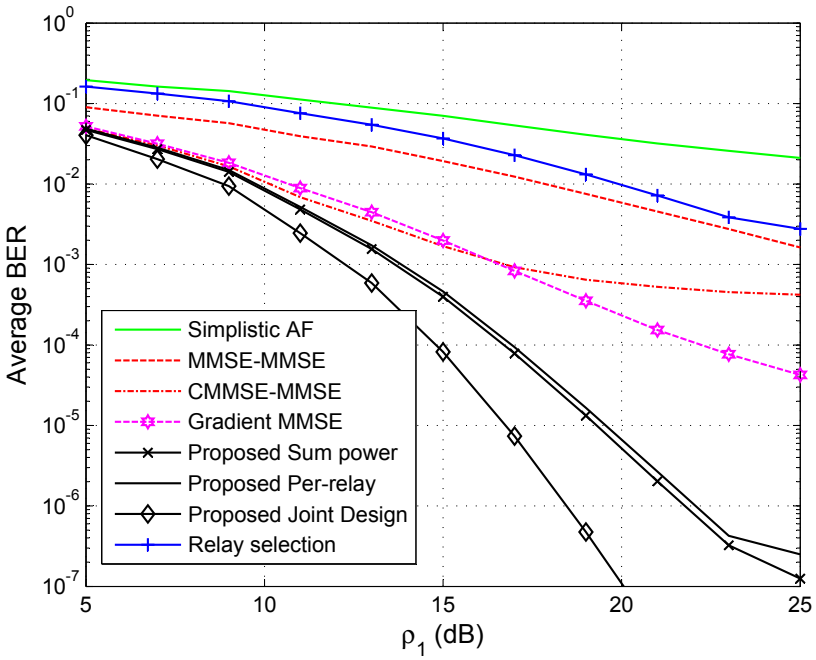

Fig. 4.9 Comparison of uncoded 16-PSK BER versus ρ_1 for different relaying strategies when $\rho_2 = 20$dB

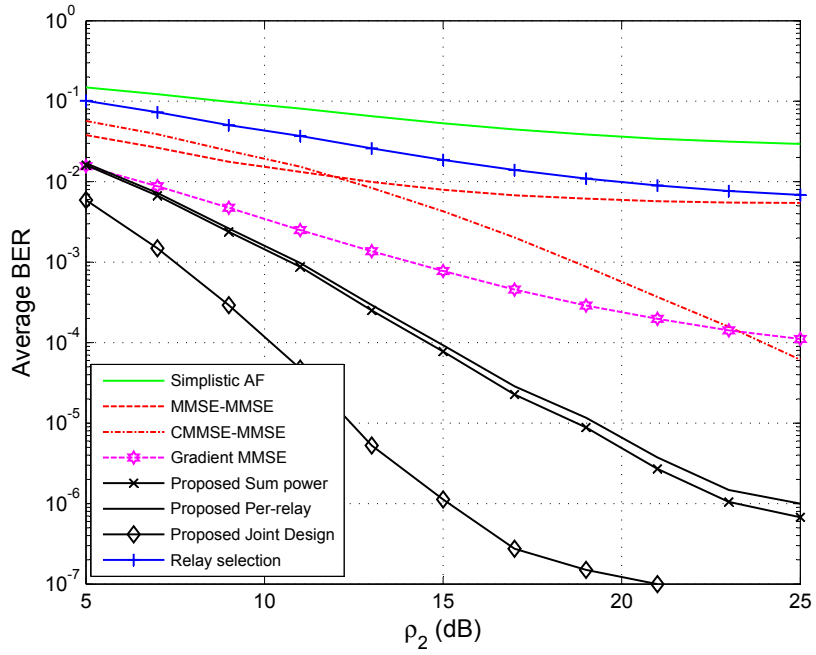

Fig. 4.10 Comparison of uncoded 16-PSK BER versus ρ_2 for different relaying strategies when $\rho_1 = 20$dB

to those relays with better backward channels and/or worse forward channels, and to those with smaller weights. Under the per-relay power constraints, we proposed the power balancing algorithm (Algorithm 4.1) which is more efficient than general-purpose interior-point methods. The optimal strategy may not use the maximum amount of power at some relays, but does not turn off a relay either, no matter how low the SNR is at that relay. Additionally, under both types of constraints, a MIMO equalizer at the destination may be designed together with the relaying matrices. The steepest descent method (Algorithm 4.2) converges much faster than the block coordinate descent method. The BER simulations show that all the proposed designs, under either type of constraints, with or without the equalizer, outperform previous ones by large margins. These simulations also illustrate significant performance advantage of multi-relay systems over single-relay ones.

Chapter 5

A Unified Framework for Adaptive Transceiver Design

In this chapter[1], we propose a transceiver optimization framework which allows us to study not only 1S-1R-1D and 1S-MR-1D, but also other multiuser networks in a unified way. This framework leads to new transceiver design algorithms for certain system configurations, and also makes it convenient to exploit the slow variations of the wireless channels between successive transmission blocks as mentioned in Sec. 1.3.1. Sec. 5.1 presents a brief review of relevant literature and reiterates our motivations. In Sec. 5.2, we formulate a unified system model which accommodates various network topologies by imposing appropriate structural constraints on the source precoder, the relaying matrix and the destination equalizer. In Sec. 5.3, we derive the optimal relaying matrix as a function of the other two matrices, thereby removing it from the optimization problem. Then in Sec. 5.4, we study in detail how to optimize either the precoder or equalizer with various structural constraints. Subsequently, the precoder and equalizer are jointly designed based on an alternating algorithm. Inter-block adaptation is used in this algorithm to exploit the above mentioned inter-block relationship. The proposed framework is further validated numerically within the context of different system configurations in Sec. 5.6.

[1]Parts of Chapter 5 have been presented at the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP) [118].

5.1 Introduction

Optimization of the relaying matrices (possibly together with source precoders and destination equalizers) under power constraints has been studied within the context of different system configurations. For 1S-1R-1D, the optimal relaying matrix takes an SVD-based form under a wide variety of criteria [15, 28, 29, 31]. The major difficulty for other system configurations comes from structural constraints on the underlying mathematical models. Taking the relay-assisted BC (1S-1R-MD or 1S-MR-MD) for example, the equalizer has to be diagonal or blockdiagonal: indeed, the users are not collocated and hence cannot jointly process their received signals. Therefore, the SVD-based relaying framework cannot be readily generalized and most existing works either 1) assume special structures for the design matrices [13, 56, 71, 75, 76, 94, 119]; or 2) iterate through the precoders, the relaying matrices and the equalizers multiple times [59–63, 67, 69, 70, 77–81, 83, 84, 118, 120]; or 3) turn to general optimization techniques such as geometrical programming (GP), semidefinite programming (SDP) or second-order cone programming (SOCP) [72,73,85,87–89,121].

Optimal transceiver design generally requires knowledge of the underlying wireless channels. In practice, the entire period is divided into transmission blocks and in each block, the channels are estimated and then used for transceiver optimization, followed by the actual data transmission. Transceiver design based on CSI implicitly assumes that the channels stay almost constant within each block. Otherwise, model mismatch would deteriorate performance significantly. This implies that both the channels and the corresponding optimal transceivers evolve gradually across successive blocks. This important aspect of the wireless channels and its implications on the design of relay transceiver has been overlooked in the development and evaluation of the above cited algorithms. In fact, by properly exploiting this inter-block correlation in the radio environment, it should be possible to reduce the complexity of iterative algorithms. For instance, a numerical algorithm that once seemed complicated may converge much faster and become competitive with those based on well-defined structures such as SVD.

In this regard, we propose a transceiver optimization framework which allows us to study various non-regenerative MIMO relay networks in a unified way. It leads to new and more efficient algorithms with better performance for certain network topologies. This framework also makes it convenient to exploit the above mentioned relationship between successive transmission blocks via inter-block adaptation. We formulate a unified system model which

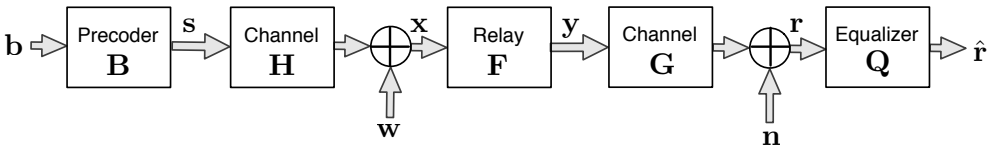

Fig. 5.1 The unified system model

includes a source precoder, a relaying matrix and a destination equalizer as design variables (matrices). This model is more general than those in Chapter 3 and 4 because it can accommodate other system configurations if we impose appropriate structural constraints on the precoder, the relaying matrix and the equalizer. For example, the equalizer needs to be diagonal or block-diagonal for broadcast networks. We then design those matrices in order to minimize the MSE between the source input and the destination output, subject to power constraints. More specifically, we first derive the optimal relaying matrix as a function of the other two matrices, thereby removing this matrix and its corresponding power constraint from the optimization problem. This is the common step for point-to-point and multiuser systems. Subsequently, we study optimization of either the precoder or the equalizer under different structural constraints and propose an alternating algorithm for the joint design of these two matrices. To exploit the previously mentioned inter-block relationship, the optimal equalizer from the *previous* block is chosen as the initial search point for the current block. This inter-block adaptation speeds up convergence and henceforth reduces computational complexity significantly.

The proposed framework is further explained and validated numerically within the context of different system configurations. For example, for relay-assisted BC with single-antenna users, the proposed framework leads to a new diagonal scaling scheme which provides more flexibility and better BER performance by allowing different users to apply their own amplitude scaling and phase rotation before decoding, in contrast to [67] which assumes the same scaling for these users. As a special case, our approach provides a closed form for the optimal solution when the users apply the same scaling, whereas only an iterative approach was used in [67].

5.2 System Model

Fig. 5.1 depicts a unified system model in which a multi-antenna source is sending multiple symbol streams simultaneously to a destination, with the assistance of a multi-antenna relay. This generic model can represent different types of MIMO relaying systems, including the 1S-MR-1D systems we studied in Chaper 3 and 4. We first explain its operations in terms of 1S-1R-1D, and then explain how it can be modified to accommodate more complex topologies by imposing structural constraints on its constituent building blocks. This system operates in a two-hop half-duplex mode: in the first time slot, the source transmits signals to the relay through the backward channel; in the second slot, the relay forwards these signals to the user via the forward channel. The relay applies a linear transformation matrix to its received signals before retransmitting them. The direct source-destination link is neglected due to the assumed high level of attenuation. The numbers of antennas at the source, relay and destination are respectively N_S, N_R and N_D.

Assuming flat fading for the wireless channels, we propose the following discrete-time complex baseband-equivalent signal model. The matrices and vectors in the model change over time, but we suppress the discrete-time block index temporarily to keep the notation simple. Later on, the block index will be introduced when we discuss inter-block adaptation. The input symbol vector $\mathbf{b} \in \mathbb{C}^{N_B \times 1}$, with zero mean and covariance $\mathbf{R}_b = \mathbf{I}_{N_B}$, consists of N_B statistically independent symbols. This vector is preprocessed by a linear *precoder* matrix $\mathbf{B} \in \mathbb{C}^{N_S \times N_B}$ to generate the transmitted signal vector

$$\mathbf{s} = \mathbf{B}\mathbf{b}. \tag{5.1}$$

The backward channel between the source and the relay is represented by matrix $\mathbf{H} \in \mathbb{C}^{N_R \times N_S}$. The signal vector $\mathbf{x} \in \mathbb{C}^{N_R \times 1}$ received at the relay is therefore

$$\mathbf{x} = \mathbf{H}\mathbf{s} + \mathbf{w}, \tag{5.2}$$

where $\mathbf{w} \in \mathbb{C}^{N_R \times 1}$ is an additive, zero-mean, circularly symmetric complex Gaussian noise with covariance $\mathbf{R}_w$.

In this baseband-equivalent model, the linear processing at the relay is represented by a matrix $\mathbf{F} \in \mathbb{C}^{N_R \times N_R}$. That is, the relay retransmits its received noisy signal $\mathbf{x}$ as in

$$\mathbf{y} = \mathbf{F}\mathbf{x}. \tag{5.3}$$

The signal received by the destination user is

$$\mathbf{r} = \mathbf{G}\mathbf{x} + \mathbf{n} = \mathbf{G}\mathbf{F}\mathbf{H}\mathbf{s} + \mathbf{G}\mathbf{F}\mathbf{w} + \mathbf{n}, \tag{5.4}$$

in which $\mathbf{G} \in \mathbb{C}^{N_D \times N_R}$ denotes the forward channel matrix from the relay to the destination. The noise term $\mathbf{n}$ is independent from $\mathbf{b}$ and $\mathbf{w}$, and modeled as a circularly symmetric complex Gaussian random vector with zero mean and covariance $\mathbf{R}_n$. The destination applies a linear equalizer $\mathbf{Q} \in \mathbb{C}^{N_B \times N_D}$ whose output is

$$\hat{\mathbf{r}} = \mathbf{Q}\mathbf{r} = \mathbf{Q}\mathbf{G}\mathbf{F}\mathbf{H}\mathbf{B}\mathbf{b} + \mathbf{Q}\mathbf{G}\mathbf{F}\mathbf{w} + \mathbf{Q}\mathbf{n}. \tag{5.5}$$

It is straightforward to extend the above signal model to other types of relay system configurations. If there are multiple sources, relays or destinations, with their numbers denoted by L, M and N, respectively, the channel matrix $\mathbf{H}$ now consists of $M \times L$ blocks, that is,

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{1,1} & \cdots & \mathbf{H}_{1,L} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{M,1} & \cdots & \mathbf{H}_{M,L} \end{bmatrix} \tag{5.6}$$

where $\mathbf{H}_{j,i}$, the (j,i)th block, corresponds to the MIMO channel from source i to relay j. Similary, matrix $\mathbf{G}$ has $N \times M$ blocks, that is

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{1,1} & \cdots & \mathbf{G}_{1,M} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{N,1} & \cdots & \mathbf{G}_{N,M} \end{bmatrix} \tag{5.7}$$

where $\mathbf{G}_{k,j}$ is the channel between relay j and destination k. The processing matrices $\mathbf{B}$, $\mathbf{F}$ or $\mathbf{Q}$ are block-diagonal:

$$\mathbf{B} \triangleq \mathrm{diag}(\mathbf{B}_1, \ldots, \mathbf{B}_L), \tag{5.8a}$$

$$\mathbf{F} \triangleq \mathrm{diag}(\mathbf{F}_1, \ldots, \mathbf{F}_M), \tag{5.8b}$$

$$\mathbf{Q} \triangleq \mathrm{diag}(\mathbf{Q}_1, \ldots, \mathbf{Q}_N). \tag{5.8c}$$

If all the destination users have a single antenna, $\mathbf{Q}$ is diagonal. It is similar with $\mathbf{B}$ and

F. The noise vectors are the stacked versions of the individual ones, that is,

$$\mathbf{w} \triangleq \mathrm{col}(\mathbf{w}_1, \ldots, \mathbf{w}_M), \tag{5.9a}$$

$$\mathbf{n} \triangleq \mathrm{col}(\mathbf{n}_1, \ldots, \mathbf{n}_N), \tag{5.9b}$$

where $\mathbf{w}_k$ is the additive noise induced at the kth relay and $\mathbf{n}_k$ is the noise at the kth destination. The relationship between structural constraints and system configurations is summarized in Table 5.1.

Table 5.1 Typical system configurations

	Q *is arbitrary*	**Q** *is diagonal/block-diagonal*
B *is arbitrary*	Point-to-point channel	Broadcast channel (BC)
B *is diagonal/block-diagonal*	Multiple access channel (MAC)	Interference channel (IC)

F is arbitrary for one-relay systems, and diagonal/block-diagonal for multi-relay systems.

5.3 The Unified Framework for Transceiver Design

5.3.1 Problem Formulation

We consider the general problem of optimizing the relaying matrix **F**, the source precoder **B** and the equalizer **Q**, in order to minimize the MSE between the precoder input and the equalizer output

$$\begin{aligned}\mathrm{MSE}(\mathbf{F},\mathbf{B},\mathbf{Q}) &\triangleq \mathrm{E}\big\{\|\hat{\mathbf{r}}-\mathbf{b}\|^2\big\} \\ &= \mathrm{tr}\big(\mathrm{E}\big\{(\hat{\mathbf{r}}-\mathbf{b})(\hat{\mathbf{r}}-\mathbf{b})^H\big\}\big) \\ &= \mathrm{tr}\left((\mathbf{QGFHB}-\mathbf{I})(\mathbf{QGFHB}-\mathbf{I})^H\right) \\ &\qquad + \mathrm{tr}(\mathbf{QGFR}_w\mathbf{F}^H\mathbf{G}^H\mathbf{Q}^H) + \mathrm{tr}(\mathbf{QR}_n\mathbf{Q}^H). \end{aligned} \tag{5.10}$$

Two power constraints are imposed simultaneously. The first is the expected transmit power of the source

$$\mathrm{E}\big\{\|\mathbf{s}\|_2^2\big\} = \mathrm{tr}(\mathbf{R}_s) = \mathrm{tr}(\mathbf{BB}^H) \le P_s. \tag{5.11}$$

If there are multiple sources, this is the sum power of these users and we shall discuss per-user constraints later. The other constraint is on the expected transmit power of the

relay

$$\mathrm{E}\big\{\|\mathbf{y}\|_2^2\big\} = \mathrm{tr}(\mathbf{R}_y) = \mathrm{tr}(\mathbf{F}\mathbf{R}_x\mathbf{F}^H) \leq P_r, \tag{5.12}$$

where $\mathbf{R}_x \triangleq \mathrm{E}\{\mathbf{x}\mathbf{x}^H\} = \mathbf{H}\mathbf{B}\mathbf{B}^H\mathbf{H}^H + \mathbf{R}_w$. In the case of multiple relays, this corresponds to the sum power of these relays.

A popular approach to solving similar problems starts from the equalizer $\mathbf{Q}$. For any choice of $\mathbf{B}$ and $\mathbf{F}$, the optimal equalizer $\mathbf{Q}^\star$ is the MMSE equalizer [18]

$$\mathbf{Q}^\star = \mathbf{R}_{rb}^H\mathbf{R}_r^{-1}, \tag{5.13}$$

where the cross-correlation and correlation matrices are defined as[2]

$$\mathbf{R}_{rb} \triangleq \mathrm{E}\{\mathbf{r}\mathbf{b}^H\} = \mathbf{G}\mathbf{F}\mathbf{H}\mathbf{B}, \tag{5.14}$$

$$\mathbf{R}_r \triangleq \mathrm{E}\{\mathbf{r}\mathbf{r}^H\} = \mathbf{G}\mathbf{F}\mathbf{H}\mathbf{B}\mathbf{B}^H\mathbf{H}^H\mathbf{F}^H\mathbf{G}^H + \mathbf{G}\mathbf{F}\mathbf{R}_w\mathbf{F}^H\mathbf{G}^H + \mathbf{R}_n. \tag{5.15}$$

However, after substituting $\mathbf{Q}^\star$ into the MSE expression in (5.10), not only the remaining problem is non-convex but also the number of constraints remains the same as in the original problem. Moreover, this approach does not work at all in the presence of multiple destinations, *i.e.*, $N > 1$, because $\mathbf{Q}$ has to be diagonal or block-diagonal.

In this thesis, we take a different approach which begins with the relaying matrix $\mathbf{F}$. The first step is to derive the optimal $\mathbf{F}^\star$ as a closed-form function of $\mathbf{B}$ and $\mathbf{Q}$, thereby removing the power constraint (5.12) from the problem. This alternative approach offers two important advantages:

(1) Various network topologies can now be treated in a unified manner because we do not need to consider their individual structural constraints until the next step.

(2) After substituting the optimal $\mathbf{F}^\star$ into (5.10), the optimization problem becomes unconstrained with respect to $\mathbf{Q}$. This fosters the use of numerical algorithms such as gradient descent, which in turn makes it easier to handle structural constraints of $\mathbf{Q}$. In addition, $\mathbf{Q}$ is not involved in any power constraint, which makes it convenient to exploit the slow variations between successive blocks.

[2]These matrices are the same as the corresponding cross-covariance and covariance matrices because the vectors have zero mean.

5.3.2 Optimal Design of the Relaying Matrix

In this subsection, we derive the optimal relaying matrix $\mathbf{F}$ as a function of the precoder $\mathbf{B}$ and the equalizer $\mathbf{Q}$. This step is the same for different network topologies. The constraint in (5.11) does not depend on $\mathbf{F}$ and henceforth does not need to be considered for now. We first consider single-relay systems. The convex optimization problem can be solved by defining the Lagrangian multiplier

$$\mathcal{L}(\mathbf{F}, \lambda) = \mathrm{MSE}(\mathbf{F}, \mathbf{B}, \mathbf{Q}) + \lambda \left(\mathrm{tr}(\mathbf{F}\mathbf{R}_x\mathbf{F}^H) - P_r\right) \tag{5.16}$$

where $\lambda > 0$. The dual problem is to

$$\begin{aligned} &\text{maximize} && D(\lambda) \triangleq \inf_{\mathbf{F}} L(\mathbf{F}, \lambda) \\ &\text{subject to} && \lambda \geq 0. \end{aligned} \tag{5.17}$$

Let the solution of the primal problem be $\mathbf{F}^\star$ and define $p^\star \triangleq \mathrm{MSE}(\mathbf{F}^\star)$; let the solution of the dual problem be $\lambda^\star$ and $d^\star \triangleq D(\lambda^\star)$. For a convex primal problem, strong duality holds (*i.e.*, the duality gap $p^\star - d^\star$ is zero) if the Slater's condition is satisfied [113, p.226]. Here, the primal problem is convex and the Slater's condition is always satisfied ($\mathbf{F} = \mathbf{0}$ is strictly feasible: $0 < P_r$). Henceforth, the KKT conditions are *necessary and sufficient* for the optimal primal-dual pair $(\mathbf{F}^\star, \lambda^\star)$, *viz.*,

$$\frac{\partial \mathcal{L}}{\partial \bar{\mathbf{F}}}\Big|_{\mathbf{F}=\mathbf{F}^\star} = \mathbf{0}, \tag{5.18}$$

$$\mathrm{tr}(\mathbf{F}^\star\mathbf{R}_x\mathbf{F}^{\star H}) - P_r \leq 0, \tag{5.19}$$

$$\lambda^\star \geq 0, \tag{5.20}$$

$$\lambda^\star \left(\mathrm{tr}(\mathbf{F}^\star\mathbf{R}_x\mathbf{F}^{\star H}) - P_r\right) = 0. \tag{5.21}$$

The first-order necessary condition in (5.18) is expressed as

$$(\mathbf{G}^H\mathbf{Q}^H\mathbf{Q}\mathbf{G} + \lambda^\star\mathbf{I})\mathbf{F}^\star\mathbf{R}_x = \mathbf{G}^H\mathbf{Q}^H\mathbf{B}^H\mathbf{H}^H. \tag{5.22}$$

If $\lambda^\star > 0$, there is a unique solution; if $\lambda^\star = 0$, the solution is not unique when $\mathbf{G}^H\mathbf{Q}^H\mathbf{Q}\mathbf{G}$ is not of full rank. In the latter case, we are interested in the particular solution (among all solutions) that leads to the smallest value of $\mathrm{tr}(\mathbf{F}\mathbf{R}_x\mathbf{F}^H) = \|\mathbf{F}\mathbf{R}_x^{1/2}\|_F^2$ and hence would most

likely satisfy (5.19). This solution is obtained by taking the pseudo-inverse. Therefore, in both cases, the optimal solution has the same form

$$\mathbf{F}^{\star}(\lambda^{\star}) = (\mathbf{G}^H\mathbf{Q}^H\mathbf{Q}\mathbf{G} + \lambda^{\star}\mathbf{I})^{\dagger}\mathbf{G}^H\mathbf{Q}^H\mathbf{B}^H\mathbf{H}^H\mathbf{R}_x^{-1}, \tag{5.23}$$

which is an explicit function of $\lambda^{\star}$. Two scenarios may apply for this duality parameter:

(1) If $\mathrm{tr}(\mathbf{F}^{\star}\mathbf{R}_x\mathbf{F}^{\star H})|_{\lambda^{\star}=0} \leq P_r$, $\lambda^{\star} = 0$ satisfies (5.19), (5.20) and (5.21). That is, the power constraint at the relay is inactive and the solution is equal to that to the unconstrained problem.

(2) If $\mathrm{tr}(\mathbf{F}^{\star}\mathbf{R}_x\mathbf{F}^{\star H})|_{\lambda^{\star}=0} > P_r$, $\lambda^{\star} = 0$ is not a solution and according to (5.21), $\lambda^{\star} > 0$ has to satisfy $\mathrm{tr}(\mathbf{F}^{\star}\mathbf{R}_x\mathbf{F}^{\star H}) = P_r$. It is straightforward to prove that the left hand side (LHS) is a monotonically decreasing function of $\lambda^{\star}$. The value is larger than P_r at $\lambda^{\star} = 0$ and converges to zero as $\lambda^{\star} \to \infty$. Therefore, a unique solution exists and can be obtained via bisection or Newton's method.

Based on the strong duality, the minimum MSE can be obtained by substituting (5.23) into (5.16), *viz.*,

$$\begin{aligned}\mathrm{MSE}_{\min}(\mathbf{B},\mathbf{Q}) = & -\mathrm{tr}\big(\mathbf{Q}\mathbf{G}(\mathbf{G}^H\mathbf{Q}^H\mathbf{Q}\mathbf{G} + \lambda^{\star}\mathbf{I})^{\dagger}\mathbf{G}^H\mathbf{Q}^H\mathbf{B}^H\mathbf{H}^H\mathbf{R}_x^{-1}\mathbf{H}\mathbf{B}\big) \\ & + N_B + \mathrm{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H) - \lambda^{\star}P_r.\end{aligned} \tag{5.24}$$

The above expression depends on the parameter $\lambda^{\star}$ which is in turn an *implicit* function of $\mathbf{Q}$. This lack of an explicit formula would complicate later design, and our approach to overcome this problem is to introduce a linear scaling $\eta > 0$ in the equalizer. If we replace any given $\mathbf{Q}$ with a scaled version $\eta^{-1}\mathbf{Q}$, the duality parameter $\lambda^{\star}$, the optimal relaying matrix $\mathbf{F}^{\star}$ and the corresponding minimum MSE are all functions of η. It turns out that for the optimal $\eta^{\star}$ leading to the smallest MSE, these quantities can all be expressed in closed forms, as presented in the following theorem:

Theorem 5.3.1. *For any* $\mathbf{Q}$*, there exists* $\eta^{\star-1}\mathbf{Q}$ *(*$\eta^{\star} > 0$*) so that:*

(a) The optimal relaying matrix is in the closed form

$$\mathbf{F}^{\star} = \eta^{\star}(\mathbf{G}^H\mathbf{Q}^H\mathbf{Q}\mathbf{G} + \theta\mathbf{I})^{-1}\mathbf{G}^H\mathbf{Q}^H\mathbf{B}^H\mathbf{H}^H\mathbf{R}_x^{-1}, \tag{5.25}$$

where $\theta \triangleq \mathrm{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H)/P_r$.

(b) $\eta^\star$ *is the unique number satisfying* $\mathrm{tr}(\mathbf{F}^\star\mathbf{R}_x\mathbf{F}^{\star H}) = P_r$.

(c) The minimum MSE with $\mathbf{F} = \mathbf{F}^\star$ *takes the form*

$$\mathrm{MSE}_{\min}(\mathbf{B},\mathbf{Q}) = \mathrm{tr}(\mathbf{I}_{N_B}) - \mathrm{tr}\big(\mathbf{B}^H\mathbf{H}^H\mathbf{R}_x^{-1}\mathbf{H}\mathbf{B} \mathbf{Q}\mathbf{G}\ (\mathbf{G}^H\mathbf{Q}^H\mathbf{Q}\mathbf{G} + \theta\mathbf{I})^{-1}\mathbf{G}^H\mathbf{Q}^H\big). \tag{5.26}$$

(d) Any other choice of $\eta^{-1}\mathbf{Q}$ *together with the corresponding* $\mathbf{F}^\star$ *in* (5.23) *would lead to an MSE no smaller than* (5.26).

Proof. See Appendix A.4. □

An alternative formulation is to introduce η as early as in the definition of the objective function in (5.10), which was used before for different network topologies [22, 67, 97]. In this case, the objective function would be a convex function of $\mathbf{F}$, but not of both η and $\mathbf{F}$. Therefore, it does not formally guarantee optimality to set to zero the partial derivatives with respect to both η and $\bar{\mathbf{F}}$.

For multi-relay system configurations, the optimal design of the block-diagonal relaying matrix has already been studied in Chapter 4, which can be logically regarded as part of this section. In fact, the principal methodologies in the above development are similar to those in the previous chapter. The main differences are that the precoder $\mathbf{B}$ was not considered in Chapter 4 and vectorization is not needed here. In the following sections, we concentrate on the single-relay scenario but the principles are equally applicable to multi-relay systems.

From now on, we suppose that for any given $\mathbf{Q}$, the corresponding $\eta^{\star-1}\mathbf{Q}$ and $\mathbf{F}^\star$ in Theorem 5.3.1 are automatically selected. Hence, the optimization problem is reduced to that of designing $\mathbf{B}$ and $\mathbf{Q}$ in order to minimize (5.26) subject to (5.11). For convenience, we define

$$\mathbf{E}_h \triangleq \big(\mathbf{I} + \mathbf{B}^H\mathbf{H}^H\mathbf{R}_w^{-1}\mathbf{H}\mathbf{B}\big)^{-1}, \tag{5.27}$$

$$\mathbf{E}_g \triangleq (\mathbf{I} + \mathbf{Q}\mathbf{G}\mathbf{G}^H\mathbf{Q}^H/\theta)^{-1} \tag{5.28}$$

and derive the following matrix equations

$$\mathbf{B}^H\mathbf{H}^H\mathbf{R}_x^{-1}\mathbf{H}\mathbf{B} = \mathbf{I} - \mathbf{E}_h \succeq \mathbf{0}, \tag{5.29}$$

$$\mathbf{Q}\mathbf{G}\ (\mathbf{G}^H\mathbf{Q}^H\mathbf{Q}\mathbf{G} + \theta\mathbf{I})^{-1}\mathbf{G}^H\mathbf{Q}^H = \mathbf{I} - \mathbf{E}_g \succeq \mathbf{0}. \tag{5.30}$$

Substituting the above equations into (5.26), the objective function is rewritten as

$$\mathrm{MSE}_{\min}(\mathbf{B}, \mathbf{Q}) = \mathrm{tr}(\mathbf{E}_h) + \mathrm{tr}(\mathbf{E}_g) - \mathrm{tr}(\mathbf{E}_h\mathbf{E}_g). \tag{5.31}$$

In fact, the optimal relaying matrix in (5.25) can be viewed as a cascade of an MMSE equalizer ($\mathbf{B}^H\mathbf{H}^H\mathbf{R}_x^{-1}$) for the backward channel, and an MMSE precoder ($\eta^\star(\mathbf{G}^H\mathbf{Q}^H\mathbf{Q}\mathbf{G}+\theta\mathbf{I})^{-1}\mathbf{G}^H\mathbf{Q}^H$) for the forward channel [105]. Correspondingly in (5.31), the first term $\mathrm{tr}(\mathbf{E}_h)$ is the minimum MSE achieved by the first-hop transmission alone [18], and $\mathrm{tr}(\mathbf{E}_g) - \mathrm{tr}(\mathbf{E}_h\mathbf{E}_g)$ is the penalty due to the second-hop transmission. The last term in (5.31), $-\mathrm{tr}(\mathbf{E}_h\mathbf{E}_g)$, is determined by both $\mathbf{B}$ and $\mathbf{G}$, implying that the joint design cannot easily be decoupled. In this regard, we first consider the optimal design for either of the two matrices while the other is fixed. The structural constraints imposed by different system configurations must be handled appropriately. Then, we propose an alternating algorithm for the joint design.

5.4 Optimization of Precoder and Equalizer with Structural Constraints

5.4.1 Optimization of the Precoder B with Fixed Equalizer Q

From (5.31), the objective is to minimize

$$g(\mathbf{B}) \triangleq \mathrm{tr}\big(\mathbf{E}_h(\mathbf{I} - \mathbf{E}_g)\big), \tag{5.32}$$

subject to (5.11). We consider two scenarios when $\mathbf{B}$ is either structurally unconstrained or diagonal. For the latter case, the power constraint is slightly different from (5.11).

Precoder with arbitrary structure. Define the EVD

$$\mathbf{I} - \mathbf{E}_g = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H, \tag{5.33}$$

where $\mathbf{U}$ is unitary and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{N_D}) \succeq \mathbf{0}$ with its diagonal entries sorted in descending order. Similarly, define the EVD

$$\mathbf{H}^H \mathbf{R}_w^{-1} \mathbf{H} = \mathbf{V} \mathbf{\Lambda}' \mathbf{V}^H, \tag{5.34}$$

where $\mathbf{V}$ is unitary and $\mathbf{\Lambda}' = \mathrm{diag}(\lambda'_1, \cdots, \lambda'_{N_S}) \succeq \mathbf{0}$. The problem of minimizing (5.26) subject to the power constraint (5.11) is equivalent to that of minimizing

$$f(\mathbf{B}) \triangleq \mathrm{tr}\big(\mathbf{\Lambda}(\mathbf{I} + \mathbf{U}^H \mathbf{B}^H \mathbf{V} \mathbf{\Lambda}' \mathbf{V}^H \mathbf{B} \mathbf{U})^{-1}\big) \tag{5.35}$$

subject to (5.11). Next, we show that the optimal precoder $\mathbf{B}^\star$ diagonalizes the matrix in (5.35).

Lemma 5.4.1. *For any* $\mathbf{B}$ *satisfying* (5.11)*, there exists another* $\tilde{\mathbf{B}}$ *such that:*

(a) $\tilde{\mathbf{B}}$ *also satisfies the power constraint* (5.11)*;*

(b) $\mathbf{U}^H \tilde{\mathbf{B}}^H \mathbf{V} \mathbf{\Lambda}' \mathbf{V}^H \tilde{\mathbf{B}} \mathbf{U}$ *is diagonal with non-increasing main diagonal entries;*

(c) $f(\tilde{\mathbf{B}}) \leq f(\mathbf{B})$.

Proof. See Appendix A.5. □

Up to now, we can assume that $\mathbf{B}$ diagonalizes $\mathbf{U}^H \mathbf{B}^H \mathbf{V} \mathbf{\Lambda}' \mathbf{V}^H \mathbf{B} \mathbf{U}$ without loss of generality. Furthermore, for any such $\mathbf{B}$, it can be shown as in [18] that there exists a $\hat{\mathbf{B}}$ satisfying $\mathbf{U}^H \hat{\mathbf{B}}^H \mathbf{V} \mathbf{\Lambda}' \mathbf{V}^H \hat{\mathbf{B}} \mathbf{U} = \mathbf{U}^H \mathbf{B}^H \mathbf{V} \mathbf{\Lambda}' \mathbf{V}^H \mathbf{B} \mathbf{U}$ and $\mathrm{tr}(\hat{\mathbf{B}} \hat{\mathbf{B}}^H) \leq \mathrm{tr}(\mathbf{B} \mathbf{B}^H)$. Specifically, $\hat{\mathbf{B}} = \mathbf{V} \mathbf{\Sigma}_B \mathbf{U}^H$, where $\mathbf{\Sigma}_B \triangleq [\mathrm{diag}(b_1, \cdots, b_{N_D}), \mathbf{0}]^T \in \mathbb{C}^{N_D \times N_S}$. The weighted sum MSE in (5.35) is now reduced to

$$f(\mathbf{B}) = \sum_{k=1}^{N_D} \lambda_k \lambda'_k / (1 + \lambda'_k b_k^2), \tag{5.36}$$

subject to $\mathrm{tr}(\mathbf{\Sigma}_B \mathbf{\Sigma}_B^H) = \sum_{k=1}^{N_D} b_k^2 \leq P_S$. This convex problem can easily be solved. We skip the details to present the following theorem directly:

Theorem 5.4.2. *For a fixed equalizer* $\mathbf{Q}$*, with the optimal* $\mathbf{F}^\star$ *in* (5.25)*, the optimal precoder* $\mathbf{B}$ *is of the SVD form*

$$\mathbf{B}^\star = \mathbf{V} \mathbf{\Sigma}_B \mathbf{U}^H. \tag{5.37}$$

The diagonal entries b_k *of* $\mathbf{\Sigma_B}$ *satisfy*

$$b_k^2 = \left(\sqrt{\frac{\lambda_k}{\lambda_k'}}\frac{1}{\sqrt{\gamma}} - \frac{1}{\lambda_k'}\right)^+, \tag{5.38}$$

where $x^+ \triangleq \max(x, 0)$ *and* $\gamma > 0$ *is the unique solution of* $\sum_{k=1}^{N_D} b_k^2 = P_S$.

Diagonal precoder. This corresponds to system configurations with multiple single-antenna source users. It is impossible for the users to cooperate in terms of signaling because their antennas are spatially distributed. Henceforth, the precoder $\mathbf{B}$ has to be diagonal and each user has its own power constraint, which can be expressed by the positive semidefinite ordering

$$\mathbf{B}\mathbf{B}^H \preceq \mathbf{P}_s = \mathrm{diag}(P_{s,1}, \ldots, P_{s,N_s}). \tag{5.39}$$

It is straightforward that each user should transmit its maximum allowable power, that is,

$$\mathbf{B} = \mathbf{P}_s^{1/2}. \tag{5.40}$$

5.4.2 Optimization of the Equalizer Q with Fixed Precoder B

Since the power constraint in (5.11) and $\mathbf{E}_h$ do not depend on $\mathbf{Q}$, the problem is equivalent to that of minimizing

$$f(\mathbf{Q}) \triangleq \mathrm{tr}\big((\mathbf{I} - \mathbf{E}_h)\mathbf{E}_g\big). \tag{5.41}$$

This *unconstrained* optimization problem can be solved by *gradient-based line search*. Starting from an initial matrix $\mathbf{Q}_0$, this method generates a sequence of iterates $\{\mathbf{Q}_n\}_{n=0}^{\infty}$ until a solution has been approximated with sufficient accuracy. Specifically, it chooses and searches along a direction $\Delta\mathbf{Q}_n$ from the current $\mathbf{Q}_n$ for a new iterate $\mathbf{Q}_{n+1}$ so that $f(\mathbf{Q}_{n+1}) < f(\mathbf{Q}_n)$. The distance to move along $\Delta\mathbf{Q}_n$ should satisfy criteria such as Wolfe's conditions [108]. In particular, *gradient descent* uses the opposite direction of the gradient

$$\Delta\mathbf{Q}_n = -\nabla_{\bar{\mathbf{Q}}}\mathrm{MSE}_{\min}|_{\mathbf{Q}=\mathbf{Q}_n}. \tag{5.42}$$

This method is summarized in Algorithm 5.1.

This algorithm handles different structural constraints conveniently. When $\mathbf{Q}$ is not

Algorithm 5.1: Gradient Descent

Choose P_r and $\epsilon > 0$;
Choose $\bar{\alpha} > 0$, $\rho, c \in (0,1)$; {Line search parameters.}
Initialize the counter $n = -1$ and the equalizer $\mathbf{Q}_0$;
repeat
  Increment counter $n \leftarrow n + 1$;
  Compute $\Delta\mathbf{Q}_n$ from (5.42);
  Set $\alpha \leftarrow \bar{\alpha}$;
  repeat
    $\alpha \leftarrow \rho\alpha$; {Backtracking line search.}
  until $\mathrm{MSE}_{\min}(\mathbf{Q}_n + \alpha\Delta\mathbf{Q}_n) \leq \mathrm{MSE}_{\min}(\mathbf{Q}_n) - c\alpha\|\Delta\mathbf{Q}_n\|_{\mathrm{F}}^2$;
  Update $\mathbf{Q}_{n+1} \leftarrow \mathbf{Q}_n + \alpha\Delta\mathbf{Q}_n$;
until $\|\Delta\mathbf{Q}_n\|_{\mathrm{F}}^2 < \epsilon$.

structurally constrained, the gradient is

$$\nabla_{\bar{\mathbf{Q}}} = \eta^{-2}\mathbf{Q}\mathbf{R}_n - \theta^{-1}\mathbf{E}_g\mathbf{C}\mathbf{E}_g\mathbf{Q}\mathbf{G}\mathbf{G}^H, \tag{5.43}$$

where $\mathbf{C} \triangleq \mathbf{B}^H\mathbf{H}^H\mathbf{R}_x^{-1}\mathbf{H}\mathbf{B}$. When $\mathbf{Q}$ is *diagonal*, we define $\mathbf{q} \triangleq \mathrm{diag}(\mathbf{Q}) = [q_1, \ldots, q_{N_D}]^T$. The gradient of $f(\mathbf{q})$ with respect to $\bar{\mathbf{q}}$ is simply the vector which includes all the diagonal entries of the above equation, that is,

$$\nabla_{\bar{\mathbf{q}}} = \eta^{-2}\mathrm{diag}(\mathbf{Q}\mathbf{R}_n) - \theta^{-1}\mathrm{diag}(\mathbf{E}_g\mathbf{C}\mathbf{E}_g\mathbf{Q}\mathbf{G}\mathbf{G}^H). \tag{5.44}$$

When $\mathbf{Q}$ is *block-diagonal*, the gradient is equal to the Hadamard (element-wise) product of (5.43) and a block-diagonal matrix $\mathbf{I}_{BD}$:

$$\nabla_{\bar{\mathbf{Q}}} = (\eta^{-2}\mathbf{Q}\mathbf{R}_n - \theta^{-1}\mathbf{E}_g\mathbf{C}\mathbf{E}_g\mathbf{Q}\mathbf{G}\mathbf{G}^H) \odot \mathbf{I}_{BD}, \tag{5.45}$$

in which the diagonal blocks of $\mathbf{I}_{BD}$ are matrices whose entries are all one.

Initial search point when $\mathbf{Q}$ *is diagonal.* The choice of the initial $\mathbf{Q}_0$ for gradient descent is very important. For this purpose, we propose to choose a phase rotation vector that can speed up the algorithms significantly. In particular, we temporarily assume that the

diagonal entries of $\mathbf{Q}$ are all phase rotation terms, that is,

$$\mathbf{Q} = \mathrm{diag}\{\mathbf{q}\} = \begin{bmatrix} e^{j\phi_1} & & \\ & \ddots & \\ & & e^{j\phi_{N_D}} \end{bmatrix}. \tag{5.46}$$

By defining the EVD $\mathbf{G}\mathbf{G}^H = \mathbf{U}_G \mathbf{\Lambda}_G \mathbf{U}_G^H$ with the eigenvalues $\lambda_1, \ldots, \lambda_{N_D}$, the objective function in (5.41) becomes

$$f(\mathbf{q}) = \mathrm{tr}\big(\mathbf{A}(\mathbf{I} + \mathbf{\Lambda}_G/\theta)\big)^{-1}\big) = \sum_{k=1}^{N_S} \frac{a_k}{\theta^{-1}\lambda_k + 1}, \tag{5.47}$$

where $\mathbf{A} \triangleq \mathbf{U}_G^H \mathbf{Q}^H (\mathbf{I} - \mathbf{E}_h)\mathbf{Q}\mathbf{U}_G$. The diagonal entries of $\mathbf{A}$ are

$$a_k = \mathbf{u}_k^H \mathbf{Q}^H (\mathbf{I} - \mathbf{E}_h)\mathbf{Q}\mathbf{u}_k = \mathbf{q}^H \mathbf{U}_{(k)}^H (\mathbf{I} - \mathbf{E}_h)\mathbf{U}_{(k)}\mathbf{q}, \tag{5.48}$$

where $\mathbf{U}_G = [\mathbf{u}_1, \ldots, \mathbf{u}_{N_D}]$ and $\mathbf{U}_{(k)} \triangleq \mathrm{diag}(\mathbf{u}_k)$. We then have to minimize

$$f(\mathbf{q}) = \mathbf{q}^H \mathbf{\Psi} \mathbf{q}, \quad \text{where } \mathbf{\Psi} \triangleq \sum_{k=1}^{N_S} \frac{\mathbf{U}_{(k)}^H (\mathbf{I} - \mathbf{E}_h)\mathbf{U}_{(k)}}{\theta^{-1}\lambda_k + 1}. \tag{5.49}$$

This is not a trivial problem because the entries of $\mathbf{q}$ are all of norm one [122]. Without loss of generality, we can assume $\phi_1 = 0$. Define the EVD $\mathbf{\Psi} = \mathbf{U}_\Psi \mathbf{\Lambda}_\Psi \mathbf{U}_\Psi^H$, with the eigenvalues sorted in a descending order. The objective function becomes the following weighted sum

$$f(\mathbf{q}) = \sum_{k=1}^{N_D} \lambda_{\Psi,k} \|\mathbf{u}_{\Psi,k}^H \mathbf{q}\|^2, \tag{5.50}$$

where $\mathbf{u}_{\Psi,k}$ is the kth column of $\mathbf{U}_\Psi$. Since the entries of $\mathbf{q}$ must have the same magnitude, the solution cannot be the eigenvector of $\mathbf{\Psi}$ corresponding to its smallest eigenvalue ($\mathbf{u}_{\Psi,N_D}$). A heuristic suboptimal approach is to allocate as much as possible to the term with the smallest weight λ_{Ψ,N_D}. This is achieved by letting

$$\phi_k = \frac{\mathbf{u}_{\Psi,N_D}(k)}{|\mathbf{u}_{\Psi,N_D}(k)|}. \tag{5.51}$$

We should bear in mind that global optimality is not of much importance here because this ϕ_k is merely used as the initial search point.

5.5 Adaptive Realization

Since optimization of either the precoder $\mathbf{B}$ or the equalizer $\mathbf{Q}$ has been discussed, we now propose an alternating algorithm for the joint the design of these two matrices. This algorithm also exploits the slow variations between successive transmission blocks by means of *inter-block adaptation*. Here, we introduce the block index k to differentiate between transmission blocks, which was temporarily suppressed in the previous sections for notational simplicity.

The alternating algorithm optimizes $\mathbf{B}[k]$ and $\mathbf{Q}[k]$ interchangeably while holding the other matrix constant. This process generates a descending sequence of MSE values and is therefore guaranteed to converge to a local optimum. Afterwards, the optimal $\mathbf{F}^\star[k]$ and $\eta^\star[k]$ are computed as in Theorem 5.3.1. This framework is described in Algorithm 5.2. The sub-procedures of updating $\mathbf{B}_m[k]$ or $\mathbf{Q}_m[k]$ individually have been discussed in Sec. 5.4.

Algorithm 5.2: Joint optimization of $\mathbf{B}[k]$, $\mathbf{Q}[k]$ and $\mathbf{F}[k]$

1: Initialize $\mathbf{Q}_0[k] = \mathbf{Q}^\star[k-1]$, $m = 0$, $\epsilon > 0$;
2: **repeat**
3: Increment the counter $m \leftarrow m + 1$;
4: Optimize and update $\mathbf{B}_m[k]$;
5: Optimize and update $\mathbf{Q}_m[k]$;
6: **until** $|\mathrm{MSE}_{\min}(\mathbf{Q}_m[k], \mathbf{B}_m[k]) - \mathrm{MSE}_{\min}(\mathbf{Q}_{m-1}[k], \mathbf{B}_{m-1}[k])| \leq \epsilon$;
7: Compute $\mathbf{F}^\star[k]$ and $\eta^\star[k]$ according to Theorem 5.3.1;
8: Let $\mathbf{B}^\star[k] \leftarrow \mathbf{B}_m[k]$, $\mathbf{Q}^\star[k] \leftarrow \eta^{\star -1}[k]\mathbf{Q}_m[k]$.

In this algorithm, inter-block adaptation is implemented by setting $\mathbf{Q}_0[k]$ to the optimal equalizer for the *previous* transmission block $\mathbf{Q}^\star[k-1]$. In addition, step 5 is essentially a "procedure call" of Algorithm 5.1 and the initial search point for this step is chosen as $\mathbf{Q}_{m-1}[k]$. As it will be shown in Sec. 5.6, inter-block adaptation speeds up convergence and henceforth reduces computational complexity significantly.

The fact that $\mathbf{Q}$ does not show up in any power constraint is especially important.

Otherwise, the initial point $\mathbf{Q}_0$ and the following iterates would not satisfy the constraints and therefore would not be eligible solutions. If we had optimized $\mathbf{Q}$ in the first place, both $\mathbf{B}$ and $\mathbf{F}$ would be involved in their corresponding power constraints, making it much more difficult to leverage the slow fluctuations between successive transmission blocks. In addition, Algorithm 5.2 is different from a conventional alternating approach widely used to solve similar problems [67,77,117]. The latter needs to alternate between *all three* matrices $\mathbf{B}$, $\mathbf{F}$ and $\mathbf{Q}$. The proposed algorithm converges significantly faster by freeing the relaying matrix $\mathbf{F}$ from being involved in this alternating process, which will be verified numerically later.

Next, we make further comments on the proposed unified framework within the context of a few typical system configurations (*cf.* Table 5.1). These systems may differ in terms of CSI availability, and the numbers of sources, relays or destinations.

Point-to-point channels (1S-1R-1D)

The optimal MMSE relaying matrix for single-user 1S-1R-1D systems is actually in a well-defined SVD form [15,31]. For example, when CSI is unavailable at the source, the precoder $\mathbf{B}$ is fixed beforehand. Without loss of generality, we assume a spatial multiplexing system with $N_S = N_B$ and $\mathbf{B} = \sqrt{P_s/N_s}\mathbf{I}_{N_s}$. In this case, there exists a waterfilling-type solution. Let the SVD of $\mathbf{R}_w^{-1/2}\mathbf{H}\mathbf{B}$ be $\mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^H$ with the singular values sorted in descending order and then the EVD of $\mathbf{I} - \mathbf{E}_h$ is

$$\mathbf{I} - \mathbf{E}_h = \mathbf{V}_1\mathbf{\Lambda}_1\mathbf{V}_1^H. \tag{5.52}$$

where $\mathbf{\Sigma}_1 = \mathbf{I} - (\mathbf{I} + \mathbf{S}_1^H\mathbf{S}_1)^{-1} = \mathrm{diag}(\lambda_{h,1}, \ldots, \lambda_{h,N_B})$. Let the SVD of $\mathbf{R}_n^{-1/2}\mathbf{G}$ be $\mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^H$ and $\mathbf{S}_2\mathbf{S}_2^H = \mathrm{diag}(\lambda_{g,1}, \ldots, \lambda_{g,N_B})$. Define $\tilde{\mathbf{Q}} \triangleq \mathbf{Q}\mathbf{R}_n^{1/2}$. We can prove that the optimal $\tilde{\mathbf{Q}}$ has the form

$$\tilde{\mathbf{Q}}^\star = \mathbf{V}_1\mathbf{S}_Q\mathbf{U}_2^H, \tag{5.53}$$

where $\mathbf{S}_Q = \mathrm{diag}\{s_1, \ldots, s_{N_B}\} \in \mathbb{C}^{N_B \times N_D}$. The singular values are expressed as

$$s_k^2 = \left(\sqrt{\frac{\lambda_{h,k}}{\gamma\lambda_{g,k}}} - \frac{1}{\lambda_{g,k}}\right)^+, \tag{5.54}$$

with γ satisfying

$$\sum_{k=1}^{n} s_k^2 = P_r. \tag{5.55}$$

The corresponding optimal relaying matrix satisfies

$$\mathbf{F}^\star \mathbf{R}_w^{-1/2} = \eta^\star \mathbf{V}_2 \mathbf{S}_F \mathbf{U}_1^H, \tag{5.56}$$

where

$$\mathbf{S}_F \triangleq (\mathbf{S}_2^H \mathbf{S}_Q^H \mathbf{S}_Q \mathbf{S}_2 + \mathbf{I})^{-1} \mathbf{S}_2^H \mathbf{S}_Q \mathbf{S}_1 (\mathbf{S}_1 \mathbf{S}_1^H + \mathbf{I})^{-1}. \tag{5.57}$$

This is equal to the SVD form as in [15, 31]. When the CSI is available at the source, the optimal $\mathbf{B}$, $\mathbf{F}$ and $\mathbf{Q}$ jointly diagonalize the equivalent channel matrix and the diagonal power allocation matrices of $\mathbf{B}$ and $\mathbf{F}$ were designed alternatively [15].

The downside is the need to perform SVDs on the channel matrices for *every* transmission block. Computing the SVD of a matrix from that of another matrix with small perturbation is not much easier than from scratch [123]. In this sense, the numerical algorithm such as Algorithm 5.2 may be competitive with the closed-form SVD approach.

MAC (MS-1R-1D)

The SVD-based transceiver for 1S-1R-1D is also applicable to MS-1R-1D MAC systems with single-antenna source users, or with multiple-antenna source users but without source CSI. For the former case, as discussed before, the precoder $\mathbf{B}$ has to be *diagonal* and each user should transmit the maximum allowable power. For the latter case, although the source users may use precoders, they cannot rely on the knowledge of the channels and hence the block-diagonal precoding matrix $\mathbf{B}$ must be predetermined.

BC (1S-1R-MD)

The MMSE design for BC was considered in [67] with single-antenna users, and in [70,77,78] with multi-antenna users. The latter case usually requires complicated algorithms that iterate through the precoder, the relaying matrix and every equalizers multiple times. The proposed general framework can be applied to these two scenarios. For single-antenna users, different users may apply their own amplitude scaling and phase rotation before decoding. This *diagonal scaling* scheme provides more flexibility than the design in [67] which assumes

the same scaling for these users ($\mathbf{Q} = \mathbf{I}$). The corresponding structural constraints are that $\mathbf{B}$ is arbitrary and $\mathbf{Q}$ is diagonal. The proposed algorithms are immediately applicable and simulation results demonstrate lower BER than [67]. As a special case when $\mathbf{Q} = \mathbf{I}_{N_D}$, our approach provides a closed-form expression for the optimal solution when the users apply the same scaling, whereas in [67], $\mathbf{B}$ and $\mathbf{F}$ were designed using an alternating algorithm. The optimal design for multi-antenna users is similar to that for single-antenna users, except that $\mathbf{Q}$ is block-diagonal and hence the gradient has a different mathematical form as in (5.45).

IC (MS-1R-MD)

In interference networks, multiple sources are communicating with their intended destinations through a relay station. This network topology is much more likely to be found in *ad hoc* networks than in cellular systems. The source users probably do not have access to CSI because of the overhead caused by CSI feedback. Therefore, unlike 1S-1R-MD systems, $\mathbf{B}$ does not need to be optimized and joint design of $\mathbf{F}$ and $\mathbf{Q}$ can be done using Algorithm 5.1.

Multi-relay systems

The inter-block adaptation can be easily applied to 1S-MR-1D and 1S-MR-MD systems as well. In fact, Algorithm 4.2 on page 75 is essentially the multi-relay version of Algorithm 5.1. Hence, adaptation can be done in a similar way to single-relay systems, that is, by setting $\mathbf{Q}_0$ to the optimal equalizer for the previous transmission block. For 1S-MR-1D systems, $\mathbf{Q}$ is block-diagonal and the gradient is the Hadamard product of (4.58) and $\mathbf{I}_{BD}$ (similar to(5.45)).

5.6 Numerical Results

In Sec. 5.6.1 and 5.6.2, we study the convergence and tracking behaviors of the proposed numerical algorithms, followed by BER simulation of a 1S-1R-MD system in Sec. 5.6.3. Our purpose is to demonstrate how the proposed framework works, instead of presenting an exhaustive study of all the possible system configurations. For convenience, we define two SNRs, $\rho_1 = P_s/(N_S\sigma_w^2)$ and $\rho_2 = P_r/(N_R\sigma_n^2)$, that represent the link quality of the first-hop and second-hop transmissions, respectively. The following parameters are chosen

Table 5.2 Parameters of fading channels

Doppler Spectrum	Jakes
Rician K-factor	4
Sampling interval	0.5μs
Sampling frequency	2MHz
Carrier frequency	2GHz
Doppler spread	$f_d = 10$Hz (velocity of 1.5m/s)
Approximate coherence time	$0.5/f_d$ =0.05s
Delay spread	NA (flat fading)
Spatial correlation	0
Transmission block	1/20 × coherence time

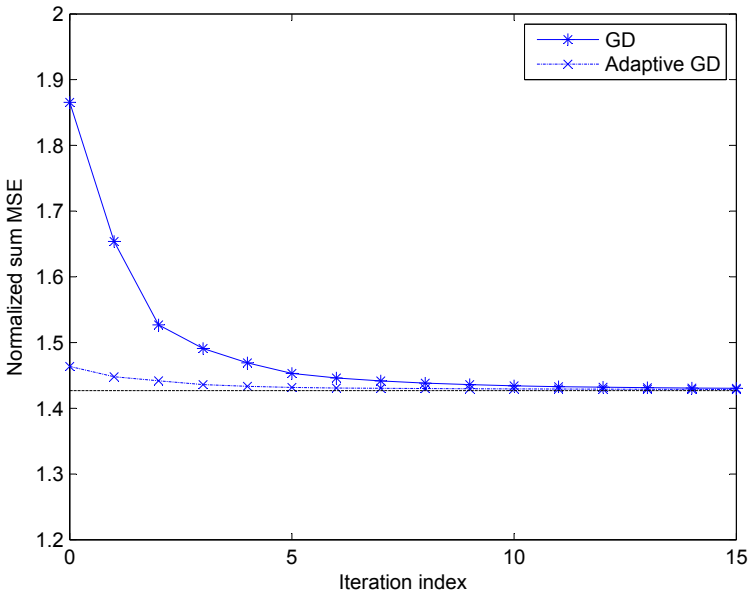

Fig. 5.2 Speed of convergence when $\mathbf{Q}$ is structurally unconstrained.

throughout this section: $N_B = N_S = N_R = N_D = 4$, $\mathbf{R}_w = \sigma_w^2\mathbf{I}$, $\mathbf{R}_n = \sigma_n^2\mathbf{I}$ and $\sigma_w^2 = \sigma_n^2$. We assume slow and flat fading MIMO channels with the parameters listed in Table 5.2. Although the doppler spread corresponds to the velocity of a pedestrian (1.5m/s), the following results still hold for higher speeds because the transmission block would be shorter accordingly.

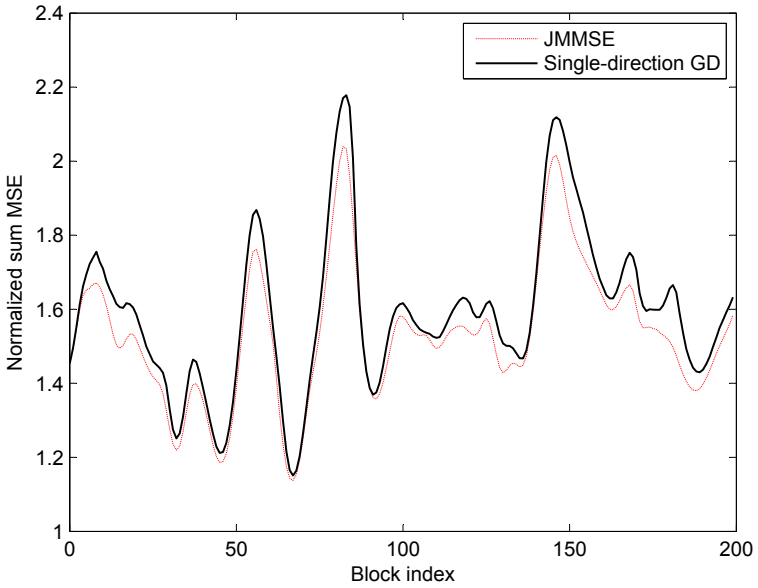

Fig. 5.3 Tracking behaviors of the proposed adaptive algorithms when $\mathbf{Q}$ is structurally unconstrained. JMMSE is the optimal solution from [31]. Single-direction GD computes the gradient only once and searches along this single direction.

5.6.1 Optimization of the Equalizer Q

In Sec. 5.4.2, we proposed the gradient descent algorithm (Algorithm 5.1) which can be used to optimize $\mathbf{Q}$. In this section, we will check convergence and tracking behaviors for this line search algorithm when $\mathbf{Q}$ is either structurally unconstrained or diagonal.

Structurally unconstrained $\mathbf{Q}$. We assume $\rho_1 = \rho_2 = 15$dB and $\mathbf{B} \propto \mathbf{I}$. The parameters in Algorithm 5.1 are chosen as $\rho = 0.9$ and $c = 0.01$. In Fig. 5.2, we compare the speed of convergence between non-adaptive and adaptive versions of gradient descent (GD) under a typical channel realization. As explained earlier, the adaptive version searches from the optimal $\mathbf{Q}^\star$ for the previous transmission block. As shown by the curves, this inter-block adaptation speeds up convergence considerably. Then, we consider a single-direction GD algorithm which computes the gradient only once and searches along the opposite direction of this gradient. The suboptimal equalizer is then used as the initial search point for the next transmission block. As shown in Fig. 5.3, single-direction GD is able to track the optimal joint MMSE method [31] closely as time evolves, though with slight performance loss.

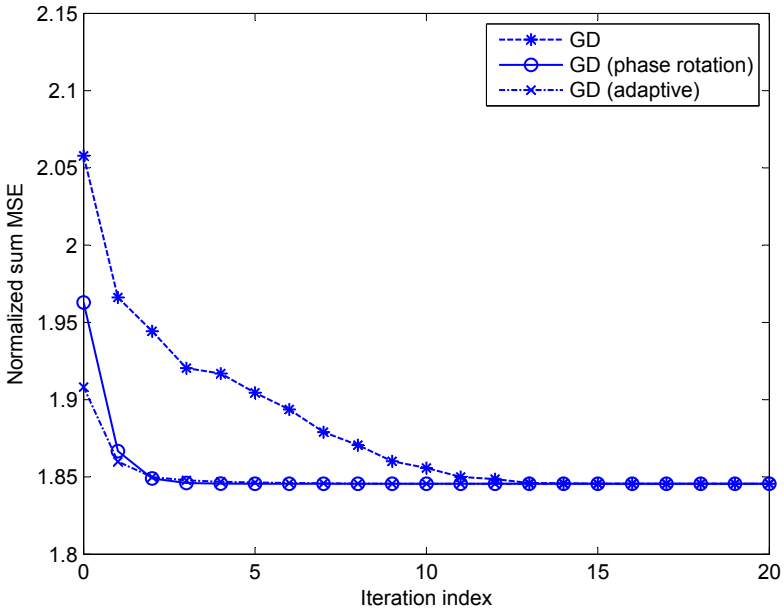

Fig. 5.4 Speed of convergence with diagonal $\mathbf{Q}$. The phase rotation versions search from the phase rotation matrix in (5.46).

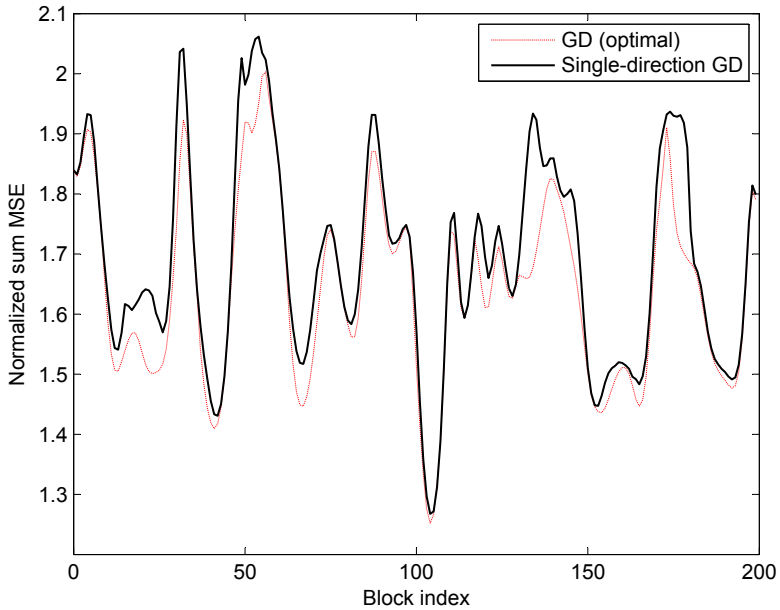

Fig. 5.5 Tracking behaviors of the proposed adaptive algorithms when $\mathbf{Q}$ is diagonal. Single-direction GD computes the gradient only once and searches along this direction.

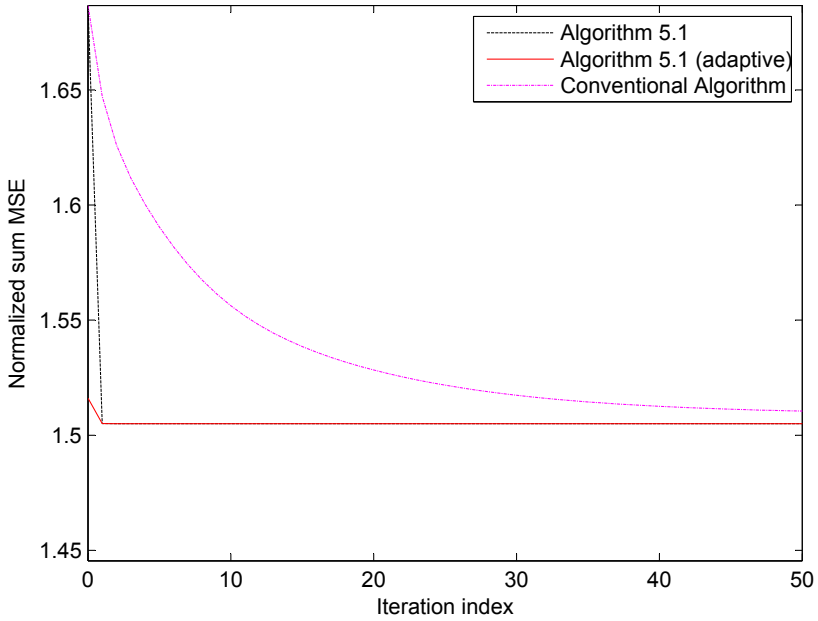

Fig. 5.6 Convergence behaviors of joint design. The conventional algorithm iterates through all three matrices $\mathbf{B}$, $\mathbf{F}$ and $\mathbf{Q}$, whereas Algorithm 5.2 updates $\mathbf{B}$ and $\mathbf{Q}$ alternatively. For the non-adaptive version, the initial values are $\mathbf{B}_0 \propto \mathbf{I}$, $\mathbf{Q}_0 \propto \mathbf{I}$ and $\mathbf{F}_0 \propto \mathbf{I}$; For the adaptive version, the solution for the previous transmission block is used as the initial search point.

Diagonal $\mathbf{Q}$. We choose the same settings as in the previous scenario. In addition to the original and adaptive versions, we consider searching from the phase rotation matrix in (5.46) as well. As shown in Fig. 5.4, the phase rotation version converges considerably faster and are therefore appropriate for the first transmission block. The adaptive version is still the best candidate for the following blocks. As in the previous scenario, single-direction GD tracks the optimal strategy closely in Fig. 5.5.

5.6.2 Joint Optimization

We compare the speed of convergence between several approaches mentioned in Sec. 5.5: Algorithm 5.2, its non-adaptive version and the conventional approach. We assume that $\mathbf{Q}$ is diagonal. As shown in Fig. 5.6, by removing $\mathbf{F}$ from the iterating process, Algorithm 5.2 converges much faster than the conventional algorithm, in fact after just one iteration. A possible explanation is that at high SNRs, the third term on the right-hand side of (5.31)

is negligibly small. Hence, it follows that the optimal $\mathbf{B}^\star$ would almost be independent of $\mathbf{Q}^\star$ and vice versa. We note that although both versions of Algorithm 5.2 converge after just one iteration, the adaptive version converge much faster than the non-adaptive version in Step 5 which is essentially Algorithm 5.1.

5.6.3 BER Simulations

We simulate the BER of a 1S-1R-4D BC system. The BS transmits four independent uncoded QPSK symbol streams to their corresponding single-antenna destination users. The spatial characteristics of the wireless channel between the BS and the relay are described by the Kronecker model: $\mathbf{H} = \mathbf{R}_r^{1/2}\mathbf{H}_w\mathbf{R}_t^{1/2}$. Herein, $\mathbf{H}_w$ has zero-mean, unit-variance, circularly symmetric complex Gaussian entries that are statistically independent. The (i,j)-th entries of $\mathbf{R}_r$ and $\mathbf{R}_t$ are both $0.7^{|i-j|}$. The forward channel is $\mathbf{G} = \mathbf{D}\mathbf{G}_w\mathbf{R}_t^{1/2}$, where $\mathbf{G}_w$ has the same statistical characteristics as $\mathbf{H}_w$ and the diagonal matrix $\mathbf{D}$ represents relative pathlosses of 0, 0, 3 and 6dB for different users. The temporal characteristics of the channels are in Table 5.2.

The methods under comparison are: 1) ZF relaying [21]: $\mathbf{B} = \sqrt{P_S/N_S}\mathbf{I}$, $\mathbf{F} = \eta\mathbf{G}^\dagger\mathbf{H}^\dagger$ and $\mathbf{Q} \propto \mathbf{I}$. 2) MMSE relaying without precoder [13]: $\mathbf{B} = \sqrt{P_S/N_S}\mathbf{I}$, $\mathbf{F}$ as in (5.25) and $\mathbf{Q} \propto \mathbf{I}$. 3) Joint design of $\mathbf{B}$ and $\mathbf{F}$ without diagonal scaling ($\mathbf{Q} \propto \mathbf{I}$) [67]; 4) Joint design of $\mathbf{B}$, $\mathbf{F}$ and $\mathbf{Q}$ with diagonal scaling. The average BERs of the multiple users are shown in Fig. (5.7) and (5.8), where we let $\rho_1 = 25$dB or $\rho_2 = 25$dB and vary the other SNR between 5 and 25dB. As expected, MMSE relaying without precoder always outperforms ZF relaying without precoder. The joint design of $\mathbf{B}$ and $\mathbf{F}$ leads to an SNR gain of 2–3dB over the MMSE relaying without precoder. Furthermore, by including the diagonal $\mathbf{Q}$ in the joint design, the diagonal scaling scheme enables an additional SNR gain of 0.5–2dB at mid-to-high SNRs.

5.7 Summary

In this chapter, we have proposed a transceiver optimization framework which allows us to study various non-regenerative MIMO relay networks in a unified way. This framework leads to new transceiver design algorithms for different network topologies, and also makes it convenient to exploit the slow variations between transmission blocks. First, we formulated a unified system model which can accommodate different types of relay systems by

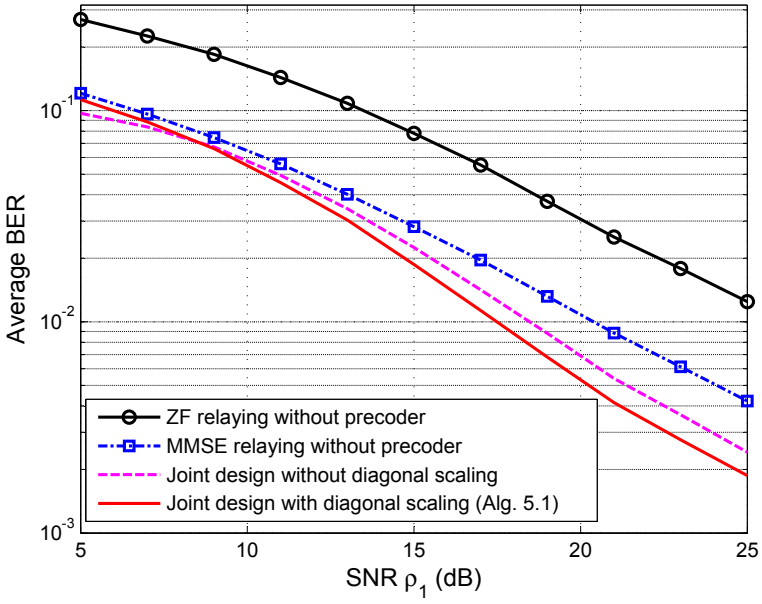

Fig. 5.7 BER versus ρ_1 with $\rho_2 = 25$dB.

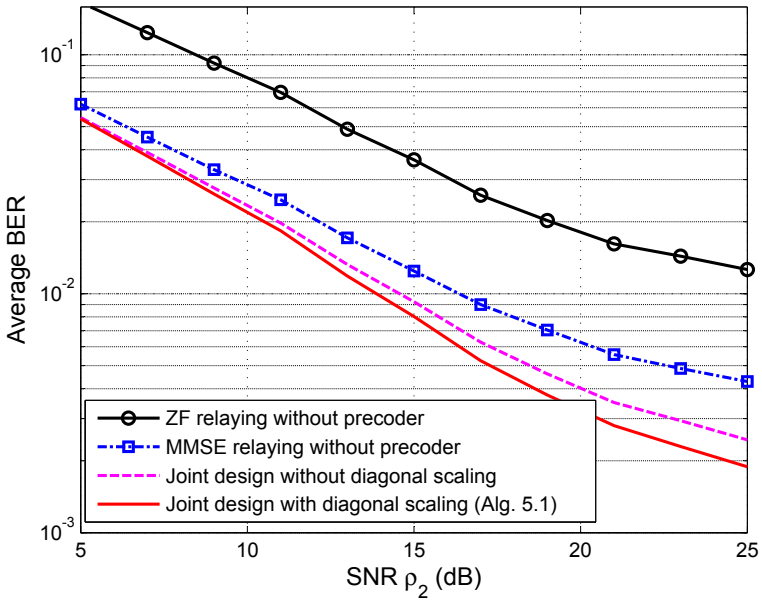

Fig. 5.8 BER versus ρ_2 with $\rho_1 = 25$dB.

imposing proper structural constraints on the precoder, the relaying matrix and the equalizer. Then, we consider the joint design of these modules based on the MMSE criterion. More specifically, the optimal relaying matrix was derived as a closed-form expression of the other two matrices, thereby removing this matrix and its corresponding power constraint from the optimization problem. Subsequently, the precoder or the equalizer was jointly designed using Algorithm 5.2 which updates each of them one at a time. When the equalizer is fixed, the optimal precoder is in a waterfilling-like SVD form if there is no structural constraint, and does not need to be designed if it is diagonal. When the precoder is fixed, the optimal equalizer can be obtained using gradient descent (Algorithm 5.1) regardless of the structural constraint. Inter-block adaptation is implemented by choosing the optimal equalizer from the previous block as the initial search point for the current block. This speeds up convergence and henceforth reduces computational complexity significantly. The proposed framework is explained in detail and verified numerically within the context of different system configurations. For example, for relay-assisted BC with single-antenna users, the proposed framework leads to a new diagonal scaling scheme which provides more flexibility and better BER performance by allowing different users to apply their own amplitude scaling and phase rotation before decoding, in contrast to [67] which assumes the same scaling for these users.

Chapter 6

Conclusions

6.1 Summary and Conclusions

In this thesis, we have studied optimal linear transceiver design for non-regenerative MIMO relay networks, which is crucial to fulfilling the great potential of such networks. We focused on two important aspects: combining and adaptation. The former stands for coherent combining at the destination of the signals transmitted from multiple multi-antenna relays to benefit from a distributed array gain; the latter denotes appropriate exploitation of time-domain properties of wireless channels to reduce algorithmic complexity.

After presenting a detailed survey of relevant literature in Chapter 2, we first studied combining in terms of a 1S-MR-1D system in which the source sends information to the destination, assisted by multiple parallel relays. The SVD-based optimal scheme for single-relay systems does not work due to the structural constraint of block-diagonality. In such systems, whether signals from different relays can *combine constructively* at the destination is of utmost importance. Therefore, we studied optimal design of the multiple relaying matrices for such combining-type 1S-MR-1D systems, aiming for coherent combing but without over-amplifying the noises induced at the relay receivers. The existence of multiple antennas makes it difficult to balance these two aspects, and therefore we considered two different approaches to transceiver design, namely, a structured hybrid framework and a MMSE-based optimization approach.

In Chapter 3, we proposed a low-complexity hybrid framework in which the non-regenerative MIMO relaying matrix at each relay is generated by cascading two substructures, akin to an equalizer for the backward channel and a precoder for the forward channel.

For each of these two substructures, we introduced two one-dimensional parametric families of candidate matrix transformations. The first family, non-cooperative by nature, depends only on the backward or forward channel of the same relay. Specifically, this family includes ZF, linear MMSE and MF as special cases, as well as other intermediate situations, thereby providing a continuous trade-off between interference and noise suppression. The second (cooperative) family also makes use of information derived from the channels of other relays. This hybrid framework allows for the classification and comparison of all possible combinations of these substructure, including several previously investigated methods and their generalizations. Within this hybrid framework, the design parameters $\boldsymbol{\lambda}$ were further optimized, resulting in significant performance improvements. This can be done on-line based on individual channel estimates or off-line based on *a priori* knowledge of the channel statistics. In the latter case, the optimal parameters can be well approximated by linear functions of SNR $[\log \rho_1, \log \rho_2]^T$ with minor performance loss. The optimal $\boldsymbol{\lambda}_{\text{opt}}$ differs significantly from those corresponding to the ZF, MF and linear MMSE relaying strategies. Through simulations, we showed that the capacity of selected hybrid schemes (with optimized parameters) comes within 1bits/s/Hz of the upper bound achieved by the capacity-optimal iterative method in [60]. In the mid-to-high SNR range, the BER performance of C-NC even exceeds that of the MSE-optimal iterative method. The proposed hybrid methods therefore achieve a good balance between performance and complexity: they outperform existing low-complexity strategies by a large margin in terms of both capacity and BER, and at the same time, are significantly simpler than previous iterative algorithms.

In Chapter 4, the other approach minimizes the mean square error (MSE) between the transmitted and received signal symbols. Two types of constraints on the transmit power of the relays were considered separately: 1) a weighted sum power constraint, and 2) per-relay power constraints. As opposed to using general-purpose interior-point methods, we exploited the inherent structure of the problems to develop more efficient algorithms. Under the weighted sum power constraint, the optimal solution was expressed as a function of a Lagrangian parameter. By introducing a complex scaling factor at the destination, we derived a closed-form expression for this parameter, thereby avoiding the need to solve an implicit nonlinear equation numerically. Under the per-relay power constraints, we proposed a power balancing algorithm (Algorithm 4.1) that converts the problem into an equivalent one with the weighted sum power constraint. This is much more efficient than

general-purpose interior-point methods. In addition, we investigated the joint design of a MIMO equalizer at the destination and the relaying matrices, using block coordinate descent or steepest descent. The steepest descent method (Algorithm 4.2) converges much faster than the block coordinate descent method. The BER simulation results demonstrated that all the proposed designs, under either type of constraints, with or without the equalizer, perform much better than previous methods and the hybrid methods. These simulations also illustrated significant performance advantage of multi-relay systems over single-relay ones. This chapter also provides new insights into the design of non-regenerative 1S-MR-1D systems. Firstly, the optimal design does not require global CSI availability: each relay only needs to know its own backward and forward channel, together with a little additional information. Secondly, under the weighted sum power constraint, the optimal strategy tends to allocate more power to those relays with better source-relay links or worse relay-destination links. Lastly, under the per-relay power constraints, the optimal strategy sometimes does not use the maximum power at some relays. Forcing equality in the per-relay power constraints as in [60] and Chapter 3 would result in loss of optimality. Another interesting point is that, no matter how low the SNR is at a particular relay, this relay does not have to be turned off completely.

The other aspect, adaptation, was investigated in Chapter 5 in a unified way for 1S-1R-1D and 1S-MR-1D and *multiuser* relay networks. We proposed a transceiver optimization framework which leads to new design algorithms for certain system configurations, and also makes it convenient to exploit the slow variations of the wireless channels between successive transmission blocks. First, we formulated a general system model which can accommodate various network topologies by imposing appropriate structural constraints on the source precoder, the relaying matrix and the destination equalizer. Next, we derived the optimal MMSE relaying matrix as a function of the other two matrices, thereby removing this matrix and its associated power constraint from the optimization This is the common step for point-to-point and multiuser systems. Subsequently, the precoder or the equalizer was jointly designed using Algorithm 5.2 which updates each of them one at a time. When the equalizer is fixed, the optimal precoder is in a waterfilling-like SVD form if there is no structural constraint, and does not need to be designed if it is diagonal. When the precoder is fixed, the optimal equalizer can be obtained using gradient descent (Algorithm 5.1) regardless of the structural constraint. Inter-block adaptation is implemented by choosing the optimal equalizer from the previous block as the initial search point for the

current block. This speeds up convergence and henceforth reduces computational complexity significantly. The proposed framework is explained in detail and verified numerically within the context of different system configurations. For example, for relay-assisted BC with single-antenna users, the proposed framework leads to a new diagonal scaling scheme which provides more flexibility and better BER performance by allowing different users to apply their own amplitude scaling and phase rotation before decoding, in contrast to [67] which assumes the same scaling for these users.

6.2 Future Works

In this thesis, we have considered transceiver optimization for non-regenerative MIMO relay networks. There are many potential research topics that can be further developed based on this thesis, and they are briefly summarized below:

(1) The MMSE transceiver in Chapter 4 can be extended by assigning different weights to the multiple sub-streams. The resulting weighted MMSE approach is able to provide different QoS to these sub-streams. For example, video streams are more sensitive with errors than speech signals. Furthermore, the weighted MMSE transceiver is equivalent to the minimum BER transceiver or the maximum mutual information transceiver when appropriate weights are chosen [15, 110], which may lead to more efficient algorithms for the design of these transceivers.

(2) The MMSE transceiver can also be extended by considering CSI uncertainty in two different ways. One is to use minimax criterion to optimize the worst-case performance. The general form is

$$\begin{array}{ll} \text{Minimize} & \max\limits_{\mathbf{G}\in\mathcal{R}_{\mathbf{G}},\mathbf{H}\in\mathcal{R}_{\mathbf{H}}} J(\mathbf{F}_k,\mathbf{Q}), \\ \text{Subject to} & \text{Power, QoS or SINR constraints,} \end{array} \tag{6.1}$$

where $\mathcal{R}_{\mathbf{G}}$ and $\mathcal{R}_{\mathbf{H}}$ are the uncertainty regions for $\mathbf{G}$ and $\mathbf{H}$. The other is the Bayesian approach— to optimize the expected performance with respect to CSI errors:

$$\begin{array}{ll} \text{Minimize} & \mathrm{E}_{\Delta\mathbf{G},\Delta\mathbf{H}}\big\{J(\mathbf{F}_j,\mathbf{Q})\big\}, \\ \text{Subject to} & \text{Power, QoS or SINR constraints,} \end{array} \tag{6.2}$$

where $\Delta\mathbf{G}$ and $\Delta\mathbf{H}$ are the CSI error matrices. This can be done in a similar way as in [37].

(3) The inter-block adaptive approach in Chapter 5 was considered in a unified framework for various network topologies. This framework needs to be tweaked and tailored for specific system configurations such as MS-1R-1D with multiantennas users and source CSI.

(5) This thesis focuses on *primitive* network topologies. Practical applications such as multi-cell coordinated multi-point (CoMP) transmissions may bring additional concerns. Interferences from other primitive networks also need special treatment [124]. These practical issues provide both challenges and opportunities for researchers and engineers.

Appendix A

Proofs and Derivations

A.1 Proof of Corollary 4.3.5

The Woodbury matrix identity does not hold for pseudo inverse in general and therefore we prove this corollary by substituting (4.27) into the LHS of (4.24). The kth sub-block of the column vector $\mathbf{\Psi f}^\star$ would be

$$\begin{aligned}\mathbf{T}_k(\mathbf{R}_s^T \otimes \mathbf{I}) \sum_{l=1}^{M} \mathbf{T}_l^H (\mathbf{S}_l + \lambda^\star w_l \mathbf{I}_{N_R^2})^\dagger \mathbf{T}_l \mathbf{\Sigma}^{-1} \mathrm{vec}(\mathbf{I}) \\ + (\mathbf{S}_k + \lambda^\star w_k \mathbf{I}_{N_R^2})(\mathbf{S}_k + \lambda^\star w_k \mathbf{I}_{N_R^2})^\dagger \mathbf{T}_k \mathbf{\Sigma}^{-1} \mathrm{vec}(\mathbf{I}). \end{aligned} \tag{A.1}$$

We always have $(\mathbf{S}_k + \lambda^\star w_k \mathbf{I}_{N_R^2})(\mathbf{S}_k + \lambda^\star w_k \mathbf{I}_{N_R^2})^\dagger \mathbf{T}_k = \mathbf{T}_k$: if $\lambda^\star w_k > 0$, the pseudo inverse operator is replaced by matrix inverse; if $\lambda^\star w_k = 0$,

$$\begin{aligned}\mathbf{S}_k \mathbf{S}_k^\dagger \mathbf{T}_k &= (\mathbf{H}_k^H \mathbf{R}_{\mathbf{x}_k}^{-1/2})^T \otimes \mathbf{G}_k^H \mathbf{Q}^H \mathbf{Q} \mathbf{G}_k (\mathbf{G}_k^H \mathbf{Q}^H \mathbf{Q} \mathbf{G}_k)^\dagger \mathbf{G}_k^H \mathbf{Q}^H \\ &= (\mathbf{H}_k^H \mathbf{R}_{\mathbf{x}_k}^{-1/2})^T \otimes \mathbf{G}_k^H \mathbf{Q}^H = \mathbf{T}_k\end{aligned}$$

because $\mathbf{Q}\mathbf{G}_k(\mathbf{G}_k^H \mathbf{Q}^H \mathbf{Q} \mathbf{G}_k)^\dagger \mathbf{G}_k^H \mathbf{Q}^H$ is a projection matrix so that $\mathbf{G}_k^H \mathbf{Q}^H$ is not changed. By inserting $(\mathbf{R}_s^T \otimes \mathbf{I})(\mathbf{R}_s^{-T} \otimes \mathbf{I})$ between $\mathbf{T}_k$ and $\mathbf{\Sigma}^{-1}$ in the second term of (A.1), the kth sub-block of $\mathbf{\Psi f}^\star$ is equal to $\mathbf{T}_k(\mathbf{R}_s^T \otimes \mathbf{I})\mathbf{\Sigma}\mathbf{\Sigma}^{-1}\mathrm{vec}(\mathbf{I}) = \mathbf{T}_k(\mathbf{R}_s^T \otimes \mathbf{I})\mathrm{vec}(\mathbf{I})$. Therefore, $\mathbf{\Psi f}^\star = \mathbf{T}(\mathbf{R}_s^T \otimes \mathbf{I})\mathrm{vec}(\mathbf{I}) = \mathbf{b}$ and (4.27) is a solution of (4.24).

In addition, (4.27) is indeed the minimum-norm solution, that is, ${\mathbf{f}^\star}^H \mathbf{f}_{(m_k)}^\perp = 0$ for m_k,

$1 \leq k \leq M$. The following equality

$$\mathcal{N}(\mathbf{S}_k^\dagger) = \mathcal{N}(\mathbf{S}_k) = \mathcal{N}(\mathbf{I} \otimes \mathbf{G}_{m_k}^H \mathbf{Q}^H \mathbf{Q} \mathbf{G}_{m_k}) = \mathcal{N}(\mathbf{I} \otimes \mathbf{Q}\mathbf{G}_{m_k}),$$

leads to $\mathbf{S}_k^\dagger \mathbf{f}_{m_k}^\perp = \mathbf{0}$. Since $\lambda^\star w_{m_k} = 0$, the inner product is

$$\mathbf{f}^{\star H} \mathbf{f}_{(m_k)}^\perp = \mathbf{f}_{m_k}^{\star H} \mathbf{f}_{m_k}^\perp = \mathrm{vec}(\mathbf{I})^H \mathbf{\Sigma}^{-1} \mathbf{T}_{m_k}^H \mathbf{S}_{m_k}^\dagger \mathbf{f}_{m_k}^\perp = 0.$$

This completes the proof.

A.2 Proof of Proposition 4.3.6

We prove this proposition by showing that the derivative of $g(\lambda)$ is negative. The matrix $\mathbf{\Psi}$ can be singular, for example, when $w_k = 0$ for a particular k and $N_R > N_S$. Hence, the major difficulty is that the following property for matrix inverse

$$\frac{\mathrm{d}\mathbf{A}(\lambda)^{-1}}{\mathrm{d}\lambda} = -\mathbf{A}(\lambda)^{-1} \frac{\mathrm{d}\mathbf{A}(\lambda)}{\mathrm{d}\lambda} \mathbf{A}(\lambda)^{-1} \tag{A.2}$$

does not hold for the pseudo inverse in general. Our tactics here is to make $\mathbf{\Psi}$ invertible by adding a matrix to it, but without changing the value of $g(\lambda)$. According to Proposition 4.3.3, $\mathcal{N}(\mathbf{\Psi})$ is the direct sum of all $\mathcal{F}_{m_k}$ satisfying $\lambda w_{m_k} = 0$, which does not depend on the specific value of λ as long as $\lambda > 0$. Let $\{\mathbf{u}_1, \cdots, \mathbf{u}_p\}$ be a set of orthonormal basis vectors for $\mathcal{N}(\mathbf{\Psi})$ and define $\mathbf{\Psi}_e \triangleq \mathbf{\Psi} + \sum_{k=1}^p \mathbf{u}_k \mathbf{u}_k^H$ (a function of λ). The orthogonality relationship $\mathcal{R}(\mathbf{\Psi}) \perp \mathcal{N}(\mathbf{\Psi})$ leads to

$$\mathbf{\Psi}_e^{-1} = \left(\mathbf{\Psi} + \sum_{k=1}^p \mathbf{u}_k \mathbf{u}_k^H \right)^{-1} = \mathbf{\Psi}^\dagger + \sum_{k=1}^p \mathbf{u}_k \mathbf{u}_k^H. \tag{A.3}$$

Since $\mathbf{b} \in \mathcal{R}(\mathbf{\Psi})$, we have $\mathbf{u}_k^H \mathbf{b} = 0$ and therefore

$$\mathbf{\Psi}_e^{-1} \mathbf{b} = \mathbf{\Psi}^\dagger \mathbf{b}. \tag{A.4}$$

To this point, all the pseudo inverses ($\mathbf{\Psi}^\dagger$) can be replaced by matrix inverses ($\mathbf{\Psi}_e^{-1}$).

Based on the chain rule, the derivative of $g(\lambda)$ satisfies

$$g'(\lambda) = -2\mathbf{b}^H \boldsymbol{\Psi}_e^{-1} \mathbf{I}_{\text{sum}} \boldsymbol{\Psi}_e^{-1} \mathbf{I}_{\text{sum}} \boldsymbol{\Psi}_e^{-1} \mathbf{b} \leq 0, \tag{A.5}$$

for the matrix $\boldsymbol{\Psi}_e^{-1}$ is positive definite. Next we prove $g'(\lambda) \neq 0$ by contradiction. Assume there exists a $\lambda_0 > 0$ so that $g'(\lambda_0) = 0$. Since $\boldsymbol{\Psi}_e(\lambda_0)^{-1}$ is positive definite, $\mathbf{I}_{\text{sum}} \boldsymbol{\Psi}_e(\lambda_0)^{-1} \mathbf{b}$ must be a zero vector, which leads to

$$\begin{aligned} \mathbf{b} &= (\boldsymbol{\Phi} + \lambda_0 \mathbf{I}_{\text{sum}}) \boldsymbol{\Psi}(\lambda_0)^\dagger \mathbf{b} = (\boldsymbol{\Phi} + \lambda_0 \mathbf{I}_{\text{sum}}) \boldsymbol{\Psi}_e(\lambda_0)^{-1} \mathbf{b} \\ &= (\boldsymbol{\Phi} + \lambda \mathbf{I}_{\text{sum}}) \boldsymbol{\Psi}_e(\lambda_0)^{-1} \mathbf{b}, \qquad \forall \lambda \geq 0. \end{aligned}$$

This means that $\mathbf{f}^\star = \boldsymbol{\Psi}_e(\lambda_0)^{-1} \mathbf{b}$ is the solution of (4.24) for any $\lambda \geq 0$. As a result, the weighted sum power satisfies

$$0 \leq g(\lambda) = \|\mathbf{I}_{\text{sum}} \boldsymbol{\Psi}(\lambda)^\dagger \mathbf{b}\|_2^2 = \|\mathbf{I}_{\text{sum}} \boldsymbol{\Psi}_e(\lambda_0)^{-1} \mathbf{b}\|_2^2 = 0.$$

This contradicts with $g(0) > 0$ and therefore, $g'(\lambda) < 0$ always holds. The limit of $g(\lambda)$ is

$$\lim_{\lambda \to \infty} g(\lambda) = \lim_{\lambda \to \infty} \mathbf{b}^H \mathbf{I}_{\text{sum}}^\dagger \mathbf{b}^H / \lambda^2 = 0.$$

A.3 Gradient of the MSE with Respect to $\bar{\mathbf{Q}}$

From the MSE expression in (4.33b), the partial derivative is

$$\frac{\partial \text{MSE}}{\partial \bar{q}_{ij}} = -\text{vec}(\mathbf{I})^H \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \bar{q}_{ij}} \boldsymbol{\Sigma}^{-1} \text{vec}(\mathbf{I}). \tag{A.6}$$

Using similar techniques to those in the proof of Proposition 4.3.6 and substituting (4.27), we can express (A.6) as

$$\eta_o^{-2} \sum_{k=1}^{M} \mathbf{f}_k^{\star H} \frac{\partial (\mathbf{S}_k + \theta w_k \mathbf{I})}{\partial \bar{q}_{ij}} \mathbf{f}_k^\star - \eta_o^{-1} \sum_{k=1}^{M} \mathbf{f}_k^{\star H} \frac{\partial \mathbf{T}_k}{\partial \bar{q}_{ij}} \boldsymbol{\Sigma}^{-1} \text{vec}(\mathbf{I}).$$

Define an indicator matrix $\mathbf{E}_{ji}$ whose (j,i)th entry is one and other entries are all zero, we have

$$\begin{aligned}
\frac{\partial \mathbf{S}_k}{\partial \bar{q}_{ij}} &= (\mathbf{R}_{x_k}^{-T/2}\mathbf{R}_{w_k}^{T}\mathbf{R}_{x_k}^{-T/2}) \otimes (\mathbf{G}_k^H \mathbf{E}_{ji}\mathbf{Q}\mathbf{G}_k), \\
\frac{\partial \theta}{\partial \bar{q}_{ij}} &= \frac{\mathrm{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{E}_{ji})}{P_R}, \\
\frac{\partial \mathbf{T}_k}{\partial \bar{q}_{ij}} &= (\mathbf{R}_{x_k}^{-T/2}\bar{\mathbf{H}}_k) \otimes \mathbf{G}_k^H \mathbf{E}_{ji}.
\end{aligned}$$

According to the property (4.16c), we have

$$\mathbf{f}_k^{\star H}\frac{\partial \mathbf{S}_k}{\partial \bar{q}_{ij}}\mathbf{f}_k^{\star} = \mathrm{tr}(\mathbf{E}_{ji}\mathbf{Q}\mathbf{G}_k\mathbf{F}_k\mathbf{R}_{w_k}\mathbf{F}_k^H\mathbf{G}_k^H).$$

Because the power constraint is tightly satisfied, we have $\sum_{k=1}^{M} w_k\|\mathbf{f}_k^{\star}\|^2 = P_R$ and therefore

$$\sum_{k=1}^{M}\mathbf{f}_k^{\star H}\Big(\frac{\partial \theta}{\partial \bar{q}_{ij}}w_k\mathbf{I}\Big)\mathbf{f}_k^{\star} = P_R\frac{\partial \theta}{\partial \bar{q}_{ij}} = \mathrm{tr}(\mathbf{E}_{ji}\mathbf{Q}\mathbf{R}_n).$$

The property (4.16c) also leads to

$$\mathbf{f}_k^{\star H}\frac{\partial \mathbf{T}_k}{\partial \bar{q}_{ij}}\mathbf{\Sigma}^{-1}\mathrm{vec}(\mathbf{I}) = \mathrm{tr}\Big(\mathbf{E}_{ji}\ \mathrm{unvec}\big(\mathbf{\Sigma}^{-1}\mathrm{vec}(\mathbf{I})\big)\mathbf{H}_k^H\mathbf{F}_k^H\mathbf{G}_k^H\Big).$$

Since $\mathrm{tr}(\mathbf{E}_{ji}\mathbf{X}) = \mathbf{X}(i,j)$, the gradient can be expressed as in (4.58).

A.4 Proof of Theorem 5.3.1

For any $\eta^{-1}\mathbf{Q}$ with $\eta > 0$, the optimal relaying matrix is obtained by replacing $\mathbf{Q}$ with $\eta^{-1}\mathbf{Q}$ in (5.23). Henceforth, $\mathbf{F}^{\star}$ and the corresponding minimum MSE are functions of η and $\lambda^{\star}$. In turn, $\lambda^{\star}$ is an implicit piecewise function of η, whose critical point $\eta_c > 0$ satisfies

$$\begin{aligned}
P_r &= \mathrm{tr}(\mathbf{F}^{\star}\mathbf{R}_x\mathbf{F}^{\star H})|_{\lambda^{\star}=0} \\
&= \eta_c^2\mathrm{tr}\big(\mathbf{B}^H\mathbf{H}^H\mathbf{R}_x^{-1}\mathbf{H}\mathbf{B}\mathbf{Q}\mathbf{G}((\mathbf{G}^H\mathbf{Q}^H\mathbf{Q}\mathbf{G})^{\dagger})^2\mathbf{G}^H\mathbf{Q}^H\big).
\end{aligned}$$

As a result, the minimum MSE itself is also a piecewise function of η. The next step is to find the particular $\eta^{\star}$ leading to the smallest minimum MSE.

If $\eta \leq \eta_c$, $\lambda^\star = 0$ is the solution to the KKT conditions. Substituting $\eta^{-1}\mathbf{Q}$ into the minimum MSE in (5.24), we found that only the third term, $\eta^{-2}\mathrm{tr}(\mathbf{Q}\mathbf{R}_n\mathbf{Q}^H)$, depends on η and hence the minimum MSE is a decreasing function of η in the interval $(0, \eta_c]$. Therefore, the smallest minimum MSE must be achieved when $\eta \geq \eta_c$.

If $\eta \geq \eta_c$, $\lambda^\star$ has to satisfy $P_r = \mathrm{tr}(\mathbf{F}^\star\mathbf{R}_x\mathbf{F}^{\star H})$. Using this equality to replace η^{-2} in the third term of (5.24) and P_r in the last term, we have

$$\begin{aligned}\mathrm{MSE}_{\min}(\eta) = \mathrm{tr}(\mathbf{I}) - \mathrm{tr}\big(&\mathbf{Q}\mathbf{G}\ (\mathbf{G}^H\mathbf{Q}^H\mathbf{Q}\mathbf{G} + \lambda^\star\eta^2\mathbf{I})^{-2} \\ &\big(\mathbf{G}^H\mathbf{Q}^H\mathbf{Q}\mathbf{G} + (2\lambda^\star\eta^2 - \theta)\mathbf{I}\big)\mathbf{G}^H\mathbf{Q}^H\mathbf{B}^H\mathbf{H}^H\mathbf{R}_x^{-1}\mathbf{H}\mathbf{B}\big),\end{aligned} \tag{A.7}$$

which can be viewed as a function of $\gamma \triangleq \lambda^\star\eta^2$. Let the SVD of $\mathbf{Q}\mathbf{G}$ be $\mathbf{U}\mathbf{S}\mathbf{V}^H$ and define $\mathbf{A} \triangleq \mathbf{U}^H\mathbf{B}^H\mathbf{H}^H\mathbf{R}_x^{-1}\mathbf{H}\mathbf{B}\mathbf{U}$, the problem is reduced to that of maximizing

$$g(\gamma) = \mathrm{tr}\big(\mathbf{S}\big(\mathbf{S}^H\mathbf{S} + \gamma\mathbf{I}\big)^{-2}\big(\mathbf{S}^H\mathbf{S} + (2\gamma - \theta)\mathbf{I}\big)^{-1}\mathbf{S}^H\mathbf{A}\big). \tag{A.8}$$

Let $\beta_k \geq 0$ be the kth diagonal entry of $\mathbf{S}\mathbf{S}^H$ and $a_k \geq 0$ be the (k,k)th entry of $\mathbf{A}$, we have

$$g(\gamma) = \sum_{k=1}^{N_D} a_k\beta_k\frac{2\gamma - \theta + \beta_k}{(\gamma + \beta_k)^2}, \tag{A.9}$$

whose first-order derivative is

$$\frac{\mathrm{d}g}{\mathrm{d}\gamma} = \sum_{k=1}^{N_D} 2a_k\beta_k\frac{\theta - \gamma}{(\gamma + \beta_k)^3} = 0. \tag{A.10}$$

If $0 < \gamma < \theta$, this derivative is positive and $g(\gamma)$ is monotonically increasing. If $\gamma > \theta$, the derivative is negative and $g(\gamma)$ is monotonically decreasing. Therefore, $\gamma^\star = \theta$ is the solitary solution to maximize $g(\gamma)$ and hence to minimize the MSE. This corresponds to a particular combination of $\eta^\star$ and $\lambda^\star$. The optimal relaying matrix in (5.25) can be obtained by replacing $\mathbf{Q}$ with $\mathbf{Q}/\eta^\star$ in (5.23). Since $\eta \geq \eta_c$, the power constraint is tightly satisfied and therefore $\eta^\star$ satisfies $\mathrm{tr}(\mathbf{F}^\star\mathbf{R}_x\mathbf{F}^{\star H}) = P_r$. Substituting $\gamma^\star = \lambda^\star\eta^{\star 2} = \theta$ into (A.7) leads to (5.26).

A.5 Proof of Lemma 5.4.1

Define the EVD $\mathbf{U}^H\mathbf{B}^H\mathbf{V}\boldsymbol{\Lambda}'\mathbf{V}^H\mathbf{B}\mathbf{U} = \tilde{\mathbf{U}}\tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{U}}^H$, with the eigenvalues, $\tilde{\lambda}_k$, sorted in the non-increasing order. The objective function is hence a weighted sum of the nonnegative diagonal entries of $\tilde{\mathbf{U}}(\mathbf{I}+\tilde{\boldsymbol{\Lambda}})^{-1}\tilde{\mathbf{U}}^H$, denoted as x_k:

$$\begin{aligned} f(\mathbf{B}) &= \mathrm{tr}\big(\boldsymbol{\Sigma}\,\tilde{\mathbf{U}}(\mathbf{I}+\tilde{\boldsymbol{\Lambda}})^{-1}\tilde{\mathbf{U}}^H\big) = \sum_{k=1}^{N_D}\lambda_k x_k, \\ &= \lambda_{N_D}\sum_{l=1}^{N_D} x_l + \sum_{k=1}^{N_D-1}\left((\lambda_k-\lambda_{k+1})\sum_{l=1}^{k} x_l\right). \end{aligned} \tag{A.11}$$

Let $\tilde{\mathbf{B}} \triangleq \mathbf{B}\mathbf{U}\tilde{\mathbf{U}}\mathbf{U}^H$. Since both $\mathbf{U}$ and $\tilde{\mathbf{U}}$ are unitary, $\tilde{\mathbf{B}}$ satisfies $\mathrm{tr}(\tilde{\mathbf{B}}\tilde{\mathbf{B}}^H) = \mathrm{tr}(\mathbf{B}\mathbf{B}^H) \le P_S$, and $\mathbf{U}^H\tilde{\mathbf{B}}^H\mathbf{V}\boldsymbol{\Lambda}'\mathbf{V}^H\tilde{\mathbf{B}}\mathbf{U} = \tilde{\boldsymbol{\Lambda}}$ is diagonal. Hence, $f(\tilde{\mathbf{B}})$ can be obtained by replacing x_k in (A.11) with $t_k \triangleq (\tilde{\lambda}_k+1)^{-1}$. For the same matrix $\tilde{\mathbf{U}}(\mathbf{I}+\tilde{\boldsymbol{\Lambda}})^{-1}\tilde{\mathbf{U}}^H$, t_k is a non-decreasing sequence comprising of its eigenvalues, whereas the sequence x_k includes its diagonal entries. The latter can be sorted into a non-decreasing sequence x_{π_k}. According to [27, Thm. 4.3.26], we have

$$\sum\nolimits_{l=1}^{k} t_l \le \sum\nolimits_{l=1}^{k} x_{\pi_l} \le \sum\nolimits_{l=1}^{k} x_l, \tag{A.12}$$

for any $1 \le k \le N_D$ (majorization). As per (A.11), $f(\tilde{\mathbf{B}}) \le f(\mathbf{B})$.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, N.J., USA: Wiley-Interscience, 2006.

[2] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, UK: Cambridge Univ. Press, 2005.

[3] G. G. Raleigh and J. M. Cioffi, "Spatio-temporal coding for wireless communication," *IEEE Trans. Commun.*, vol. 46, pp. 357–366, Mar. 1998.

[4] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, pp. 585–595, Nov. 1999.

[5] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1073–1096, May 2003.

[6] H. Bölcskei, *Space-time Wireless Systems : from Array Processing to MIMO Communications*. Cambridge, UK: Cambridge Univ. Press, 2006.

[7] A. Osseiran, E. Hardouin, A. Gouraud, M. Boldi, I. Cosovic, K. Gosse, J. Luo, S. Redana, W. Mohr, J. Monserrat, T. Svensson, A. Tolli, A. Mihovska, and M. Werner, "The road to IMT-advanced communication systems: State-of-the-art and innovation areas addressed by the winner+ project," *IEEE Commun. Mag.*, vol. 47, pp. 38–47, Jun. 2009.

[8] K. Loa, C. C. Wu, S. T. Sheu, Y. Yuan, M. Chion, D. Huo, and L. Xu, "IMT-advanced relay standards," *IEEE Commun. Mag.*, vol. 48, pp. 40–48, Aug. 2010.

[9] A. Scaglione, D. L. Goeckel, and J. N. Laneman, "Cooperative communications in mobile *ad hoc* networks," *IEEE Signal Process. Mag.*, vol. 23, pp. 18–29, Sep. 2006.

[10] N. Khajehnouri and A. H. Sayed, "Distributed MMSE relay strategies for wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 55, pp. 3336–3348, Jul. 2007.

[11] T. Cover and A. E. Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inf. Theory*, vol. 25, pp. 572–584, Sep. 1979.

[12] R. H. Y. Louie, Y. Li, and B. Vucetic, "Practical physical layer network coding for two-way relay channels: performance analysis and comparison," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 764–777, Feb. 2010.

[13] O. Oyman and A. J. Paulraj, "Design and analysis of linear distributed MIMO relaying algorithms," *IEE Proc.-Commun.*, vol. 153, pp. 565–572, Apr. 2006.

[14] A. S. Tanenbaum and D. J. Wetherall, *Computer Networks*, 5th ed. Pearson Higher Ed, 2010.

[15] Y. Rong, X. Tang, and Y. Hua, "A unified framework for optimizing linear non-regenerative multicarrier MIMO relay communication systems," *IEEE Trans. Signal Process.*, vol. 57, pp. 4837–4851, Dec. 2009.

[16] E. C. van der Meulen, "Three-terminal communication channels," *Advances in Applied Probability*, vol. 3, pp. 120–154, 1971.

[17] ——, "A survey of multi-way channels in information theory: 1961-1976," *IEEE Trans. Inf. Theory*, vol. 23, pp. 1–37, Jan. 1977.

[18] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: a unified framework for convex optimization," *IEEE Trans. Signal Process.*, vol. 51, pp. 2381–2401, Sep. 2003.

[19] B. Wang, J. S. Zhang, and A. Host-Madsen, "On the capacity of MIMO relay channels," *IEEE Trans. Inf. Theory*, vol. 51, pp. 29–43, Jan. 2005.

[20] S. Jin, M. R. McKay, C. Zhong, and K.-K. Wong, "Ergodic capacity analysis of amplify-and-forward MIMO dual-hop systems," *IEEE Trans. Inf. Theory*, vol. 56, pp. 2204–2224, May 2010.

[21] R. Louie, Y. Li, and B. Vucetic, "Zero forcing in general two-hop relay networks," *IEEE Trans. Veh. Technol.*, vol. 59, pp. 191–202, Jan. 2010.

[22] C. Song, K.-J. Lee, and I. Lee, "MMSE based transceiver designs in closed-loop non-regenerative MIMO relaying systems," *IEEE Trans. Wireless Commun.*, pp. 2310–2319, Jul. 2010.

[23] P. Coronel and W. Schott, "Spatial multiplexing in the single-relay MIMO channel," in *Proc. 2007 IEEE Sarnoff Symp.*, Princeton, NJ, USA, Mar. 2007, 4 pages.

[24] C. Li, X. Wang, L. Yang, and W.-P. Zhu, "A joint source and relay power allocation scheme for a class of MIMO relay systems," *IEEE Trans. Signal Process.*, vol. 57, pp. 4852–4860, Dec. 2009.

[25] I. Hammerstrom and A. Wittneben, "Power allocation schemes for amplify-and-forward MIMO-OFDM relay links," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 2798–2802, Aug 2007.

[26] Y. Rong and Y. Hua, "Optimality of diagonalization of multi-hop MIMO relays," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 6068–6077, Dec. 2009.

[27] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, UK: Cambridge Univ. Press, 1985.

[28] O. Muñoz-Medina, J. Vidal, and A. Agustín, "Linear transceiver design in nonregenerative relays with channel state information," *IEEE Trans. Signal Process.*, vol. 55, pp. 2593–2604, Jun. 2007.

[29] X. Tang and Y. Hua, "Optimal design of non-regenerative MIMO wireless relays," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 1398–1407, Apr. 2007.

[30] Z. Fang, Y. Hua, and J. C. Koshy, "Joint source and relay optimization for a non-regenerative MIMO relay," in *Proc. 4th IEEE Workshop Sensor Array Multichannel Signal Process.*, Waltham, MA, USA, Jul. 2006, pp. 239–243.

[31] W. Guan and H. Luo, "Joint MMSE transceiver design in non-regenerative MIMO relay systems," *IEEE Commun. Lett.*, vol. 12, pp. 517–519, Jul. 2008.

[32] Y. Rong, "Multihop nonregenerative MIMO relays QoS considerations," *IEEE Trans. Signal Process.*, vol. 59, pp. 290–303, Jan. 2011.

[33] L. Sanguinetti and A. D'Amico, "Power allocation in two-hop amplify-and-forward MIMO relay systems with QoS requirements," *IEEE Trans. Signal Process.*, vol. 60, pp. 2494–2507, May 2012.

[34] Y. Rong, "Optimal linear non-regenerative multi-hop MIMO relays with MMSE-DFE receiver at the destination," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 2268–2279, Jul. 2010.

[35] A. Millar, S. Weiss, and R. Stewart, "Tomlinson Harashima precoding design for non-regenerative MIMO relay networks," in *Proc. IEEE 73rd Veh. Technol. Conf.*, 2011, 5 pages.

[36] J.-K. Zhang, A. Kavcic, and K. Wong, "Equal-diagonal QR decomposition and its application to precoder design for successive-cancellation detection," *IEEE Trans. Inf. Theory*, vol. 51, pp. 154–172, Jan. 2005.

[37] C. Xing, S. Ma, and Y.-C. Wu, "Robust joint design of linear relay precoder and destination equalizer for dual-hop amplify-and-forward MIMO relay systems," *IEEE Trans. Signal Process.*, vol. 58, pp. 2273–2283, Apr. 2010.

[38] C. Xing, S. Ma, Y.-C. Wu, and T.-S. Ng, "Transceiver design for dual-hop nonregenerative MIMO-OFDM relay systems under channel uncertainties," *IEEE Trans. Signal Process.*, vol. 58, pp. 6325–6339, Dec. 2010.

[39] Y. Rong, "Robust design for linear non-regenerative MIMO relays with imperfect channel state information," *IEEE Trans. Signal Process.*, vol. 59, pp. 2455–2460, May 2011.

[40] M. Herdin, "MIMO amplify-and-forward relaying in correlated MIMO channels," in *Proc. 5th Int. Conf. Inf. Commun. Signal Process.*, Prague, Czech Republic, Aug. 2005, pp. 796–800.

[41] H. W. Je, B. Lee, S. Kim, and K. B. Lee, "Design of non-regenerative MIMO-relay system with partial channel state information," in *International Conference on Communications*, 2008, pp. 4441–4445.

[42] H. W. Je, D. H. Kim, and K. B. Lee, "Joint precoding for MIMO-relay systems with partial channel state information," in *Proc. 2009 IEEE Int. Conf. Commun.*, Dresden, Germany, Jun. 2009, 5 pages.

[43] C. Jeong, B. Seo, S.-R. Lee, H.-M. Kim, and I.-M. Kim, "Relay precoding for non-regenerative MIMO relay systems with partial CSI feedback," *IEEE Trans. Wireless Commun.*, vol. 11, pp. 1698–1711, May. 2012.

[44] L. Tong, B. Sadler, and M. Dong, "Pilot-assisted wireless transmissions: general model, design criteria, and signal processing," *IEEE Signal Process. Mag.*, vol. 21, pp. 12–25, Nov. 2004.

[45] J. Ma, P. Orlik, J. Zhang, and G. Li, "Pilot matrix design for estimating cascaded channels in two-hop MIMO Amplify-and-Forward relay systems," *IEEE Trans. Wireless Commun.*, vol. 10, pp. 1956–1965, Jun. 2011.

[46] T. Kong and Y. Hua, "Optimal design of source and relay pilots for MIMO relay channel estimation," *IEEE Trans. Signal Process.*, vol. 59, pp. 4438–4446, Sep. 2011.

[47] Y. Rong, M. Khandaker, and Y. Xiang, "Channel estimation of dual-hop MIMO relay system via parallel factor analysis," *IEEE Trans. Wireless Commun.*, vol. 11, pp. 2224–2233, Jun. 2012.

[48] Y. Huang, L. Yang, M. Bengtsson, and B. Ottersten, "A limited feedback joint precoding for amplify-and-forward relaying," *IEEE Trans. Signal Process.*, vol. 58, pp. 1347–1357, Mar. 2010.

[49] Y. J. Fan and J. Thompson, "MIMO configurations for relay channels: theory and practice," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 1774–1786, May 2007.

[50] B. Yoo, W. Park, K. Lee, and C. Lee, "A simple linear precoder for selection-type cooperative MIMO systems," *IEEE Commun. Lett.*, vol. 14, pp. 306–308, Apr. 2010.

[51] H. Bolcskei, R. U. Nabar, O. Oyman, and A. J. Paulraj, "Capacity scaling laws in MIMO relay networks," *IEEE Trans. Wireless Commun.*, vol. 5, pp. 1433–1444, Jun. 2006.

[52] S. O. Gharan, A. Bayesteh, and A. K. Khandani, "Asymptotic analysis of amplify and forward relaying in a parallel MIMO relay network," *IEEE Trans. Inf. Theory*, vol. 57, pp. 2070–2082, Apr. 2011.

[53] G. Zheng, K.-K. Wong, A. Paulraj, and B. Ottersten, "Collaborative-relay beamforming with perfect CSI: Optimum and distributed implementation," *IEEE Signal Process. Lett.*, vol. 16, pp. 257–260, Apr. 2009.

[54] ——, "Robust collaborative-relay beamforming," *IEEE Trans. Signal Process.*, vol. 57, pp. 3130–3143, Aug. 2009.

[55] P. U. Sripathi and J. S. Lehnert, "A throughput scaling law for a class of wireless relay networks," in *Proc. 38th Asilomar Conf. Signals, Systems and Computers*, vol. 2, Pacific Grove, CA, USA, Nov. 2004, pp. 1333–1337.

[56] H. Shi, T. Abe, T. Asai, and H. Yoshino, "A relaying scheme using QR decomposition with phase control for MIMO wireless networks," in *Proc. IEEE Int. Conf. Commun.*, vol. 4, Seoul, Korea, May 2005, pp. 2705–2711.

[57] T. Abe, H. Shi, T. Asai, and H. Yoshino, "Relay techniques for MIMO wireless networks with multiple source and destination pairs," *EURASIP J. Wireless Commun. Netw.*, vol. 2006, Article ID 64159, 9 pages, 2006.

[58] H. Shi, T. Abe, T. Asai, and H. Yoshino, "Relaying schemes using matrix triangularization for MIMO wireless networks," *IEEE Trans. Commun.*, vol. 55, pp. 1683–1688, Sep. 2007.

[59] Y. Fu, L. Yang, and W.-P. Zhu, "A nearly optimal amplify-and-forward relaying scheme for two-hop MIMO multi-relay networks," *IEEE Commun. Lett.*, vol. 14, pp. 229–231, Mar. 2010.

[60] K.-J. Lee, H. Sung, E. Park, and I. Lee, "Joint optimization for one and two-way MIMO AF multiple-relay systems," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 3671–3681, Dec. 2010.

[61] Y. Rong, "Joint source and relay optimization for two-way MIMO multi-relay networks," *IEEE Commun. Lett.*, vol. 15, pp. 1329–1331, Dec 2011.

[62] Y. Izi and A. Falahati, "Amplify-forward relaying for multiple-antenna multiple relay networks under individual power constraint at each relay," *EURASIP J. Wireless Commun. Netw.*, vol. 2012, no. 50, pp. 1–10, 2012.

[63] L. Sanguinetti, A. D'Amico, and Y. Rong, "A tutorial on the optimization of amplify-and-forward MIMO relay systems," *IEEE J. Sel. Areas Commun.*, vol. 30, pp. 1331–1346, Sep. 2012.

[64] A. S. Behbahani, R. Merched, and A. M. Eltawil, "Optimizations of a MIMO relay network," *IEEE Trans. Signal Process.*, vol. 56, pp. 5062–73, Oct. 2008.

[65] A. Toding, M. R. A. Khandaker, and R. Yue, "Optimal joint source and relay beamforming for parallel MIMO relay networks," in *Proc. 6th Int. Conf. Wireless Commun. Netw. Mobile Comput.*, Chengdu, China, Sep. 2010, 4 pages.

[66] A. Toding, M. Khandaker, and Y. Rong, "Joint source and relay optimization for parallel MIMO relays using MMSE-DFE receiver," in *Proc. 16th Asia-Pacific Conf. Commun.*, Auckland, New Zealand, Oct.-Nov. 2010, pp. 12–16.

[67] G. Li, Y. Wang, T. Wu, and J. Huang, "Joint linear filter design in multi-user non-regenerative MIMO-relay systems," in *Proc. 2009 IEEE Int. Conf. Commun.*, Dresden, Germany, Jun. 2009, 6 pages.

[68] Y. Yu and Y. Hua, "Power allocation for a MIMO relay system with multiple-antenna users," *IEEE Trans. Signal Process.*, vol. 58, pp. 2823–2835, May 2010.

[69] M. Khandaker and Y. Rong, "Joint source and relay optimization for multiuser MIMO relay communication systems," in *Proc. 4th Int. Conf. Signal Process. Commun. Sys.*, Gold Coast, Australia, Dec. 2010, 6 pages.

[70] C. Xing, S. Ma, M. Xia, and Y.-C. Wu, "Cooperative beamforming for dual-hop amplify-and-forward multi-antenna relaying cellular networks," *Signal Process.*, vol. 92, no. 11, pp. 2689–2699, Nov. 2012.

[71] C. B. Chae, T. Tang, R. W. Heath, and S. Cho, "MIMO relaying with linear processing for multiuser transmission in fixed relay networks," *IEEE Trans. Signal Process.*, vol. 56, pp. 727–738, Feb. 2008.

[72] W. Xu, X. Dong, and W.-S. Lu, "Joint precoding optimization for multiuser multiantenna relaying downlinks using quadratic programming," *IEEE Trans. Commun.*, vol. 59, pp. 1228–1235, May 2011.

[73] R. Zhang, C. C. Chai, and Y. C. Liang, "Joint beamforming and power control for multiantenna relay broadcast channel with QoS constraints," *IEEE Trans. Signal Process.*, vol. 57, pp. 726–737, Feb. 2009.

[74] J. Li, J. Zhang, L. Guan, and Y. Zhang, "A comparison of broadcast strategy in MIMO relay networks," in *Proc. IEEE 68th Veh. Technol. Conf.*, Calgary, Alberta, Canada, Sep. 2008, 5 pages.

[75] W. Liu, C. Li, J.-D. Li, and L. Hanzo, "Block diagonalisation-based multiuser multiple input multiple output-aided downlink relaying," *IET Commun.*, vol. 6, no. 15, pp. 2371–2377, 2012.

[76] C. Zhai, X. Li, and Y. Hei, "A novel decomposed transceiver design for multiuser MIMO relay downlink systems," in *Proc. IEEE Wireless Commun. Networking Conf.*, Paris, France, Apr. 2012, pp. 1921–1924.

[77] Y. Cai, D. Le Ruyet, and D. Roviras, "Joint interference suppression and power allocation techniques for multiuser multiantenna relay broadcast systems," in *Proc. 7th Int. Symp. Wireless Commun. Sys.*, York, UK, Sep. 2010, pp. 265–269.

[78] J. Liu, Z. Liu, and Z. Qiu, "Joint MMSE transceiver design for multiuser non-regenerative MIMO relay downlink systems," in *Proc. 7th Int. Wireless Commun. Mobile Comput. Conf.*, Istanbul, Turkey, Jul. 2011, pp. 877–882.

[79] S. Jang, J. Yang, and D.-K. Kim, "Minimum MSE design for multiuser MIMO relay," *IEEE Commun. Lett.*, vol. 14, pp. 812–814, Sep. 2010.

[80] G. Okeke, W. Krzymien, and Y. Jing, "Beamforming in non-regenerative MIMO broadcast relay networks," *IEEE Trans. Signal Process.*, vol. 60, pp. 6641–6654, Dec. 2012.

[81] H.-J. Choi, K.-J. Lee, C. Song, H. Song, and I. Lee, "Weighted sum rate maximization for multiuser multirelay MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 62, pp. 885–889, Feb. 2013.

[82] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, pp. 461–471, Feb. 2004.

[83] W. Xu, X. Dong, and W.-S. Lu, "MIMO relaying broadcast channels with linear precoding and quantized channel state information feedback," *IEEE Trans. Signal Process.*, vol. 58, pp. 5233–5245, Oct. 2010.

[84] B. Zhang, Z. He, K. Niu, and L. Zhang, "Robust linear beamforming for MIMO relay broadcast channel with limited feedback," *IEEE Signal Process. Lett.*, vol. 17, pp. 209–212, Feb. 2010.

[85] B. K. Chalise, L. Vandendorpe, and J. Louveaux, "MIMO relaying for multi-point to multi-point communication in wireless networks," in *Proc. 2nd IEEE Int. Workshop Comput. Advances Multi-Sensor Adaptive Process.*, Wyndham Sugar Bay Resort & Spa St. Thomas, U.S. Virgin Islands, Dec. 2007, pp. 217–220.

[86] B. K. Chalise and L. Vandendorpe, "Joint optimization of multiple MIMO relays for multi-point to multi-point communication in wireless networks," in *Proc. IEEE 10th Workshop Signal Process. Advances Wireless Commun.*, Perugia, Italy, Jun. 2009, pp. 479–483.

[87] ——, "MIMO relay design for multipoint-to-multipoint communications with imperfect channel state information," *IEEE Trans. Signal Process.*, vol. 57, pp. 2785–2796, Jul. 2009.

[88] B. Chalise and L. Vandendorpe, "Optimization of MIMO relays for multipoint-to-multipoint communications: Nonrobust and robust designs," *IEEE Trans. Signal Process.*, vol. 58, pp. 6355–6368, Dec. 2010.

[89] A. El-Keyi and B. Champagne, "Cooperative MIMO-beamforming for multiuser relay networks," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, Las Vegas, Nevada, USA, Mar.-Apr. 2008, pp. 2749–2752.

[90] ——, "Adaptive linearly constrained minimum variance beamforming for multiuser cooperative relaying using the Kalman filter," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 641–651, Feb. 2010.

[91] M. Khandaker and Y. Rong, "Interference MIMO relay channel: Joint power control and transceiver-relay beamforming," *IEEE Trans. Signal Process.*, vol. 60, pp. 6509–6518, Dec. 2012.

[92] K. Truong, P. Sartori, and R. Heath, "Cooperative algorithms for MIMO amplify-and-forward relay networks," *IEEE Trans. Signal Process.*, vol. 61, pp. 1272–1287, May. 2013.

[93] C. Zhao and B. Champagne, "Non-regenerative MIMO relaying strategies – from single to multiple cooperative relays," in *Proc. 2nd Int. Conf. Wireless Commun. Signal Process.*, Suzhou, China, Oct. 2010, 6 pages.

[94] ——, "A low-complexity hybrid framework for combining-type non-regenerative MIMO relaying," *Wireless Personal Commun.*, vol. 72, no. 1, pp. 635–652, Sep. 2013.

[95] G. Li, Y. Wang, T. Wu, and J. Huang, "Joint linear filter design in multi-user cooperative non-regenerative MIMO relay systems," *EURASIP J. Wireless Commun. Netw.*, vol. 2009, Article ID 670265, 13 pages, 2009.

[96] B. K. Chalise and L. Vandendorpe, "Joint linear processing for an amplify-and-forward MIMO relay channel with imperfect channel state information," *EURASIP J. Adv. Signal Process.*, vol. 2010, Article ID 640186, 13 pages, 2010.

[97] C. Zhao and B. Champagne, "MMSE-based non-regenerative parallel MIMO relaying with simplified receiver," in *Proc. IEEE Global Telecomm. Conf.*, Houston, TX, USA, Dec. 2011, 5 pages.

[98] Y. Attar Izi and A. Falahati, "Robust relaying schemes for multiple-antenna multi-relay networks," *Wireless Personal Commun.*, 21pages, 2012.

[99] W. Guan, H. Luo, and W. Chen, "Linear relaying scheme for MIMO relay system with QoS requirements," *IEEE Signal Process. Lett.*, vol. 15, pp. 697–700, 2008.

[100] A. A. Nasir, H. Mehrpouyan, S. D. Blostein, S. Durrani, and R. A. Kennedy, "Timing and carrier synchronization with channel estimation in multi-relay cooperative networks," *IEEE Trans. Signal Process.*, vol. 60, pp. 793–811, Feb. 2012.

[101] D. R. Pauluzzi and N. C. Beaulieu, "A comparison of SNR estimation techniques for the AWGN channel," *IEEE Trans. Commun.*, vol. 48, pp. 1681–1691, Oct. 2000.

[102] G. H. Golub and C. F. Van Loan, *Matrix Computations.* Baltimore, MD, USA: Johns Hopkins Univ. Press, 1996.

[103] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, pp. 1702–1715, Jul. 2003.

[104] H. L. Van Trees, *Optimum Array Processing.* New York, NY: John Wiley & Sons, 2002.

[105] M. Joham, W. Utschick, and J. A. Nossek, "Linear transmit processing in MIMO communications systems," *IEEE Trans. Signal Process.*, vol. 53, pp. 2700–2712, Aug. 2005.

[106] A. Scaglione, P. Stoica, S. Barbarossa, G. B. Giannakis, and H. Sampath, "Optimal designs for space-time linear precoders and decoders," *IEEE Trans. Signal Process.*, vol. 50, pp. 1051–1064, May 2002.

[107] P. Wolfe, "Convergence conditions for ascent methods," *SIAM review*, pp. 226–235, 1969.

[108] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer Science+Business Media, LLC., 2006.

[109] C. Zhao and B. Champagne, "Joint design of multiple non-regenerative MIMO relaying matrices with power constraints," *IEEE Trans. Signal Process.*, vol. 99, no. 19, pp. 4861–4873, Oct. 2013.

[110] H. Sampath, P. Stoica, and A. Paulraj, "Generalized linear precoder and decoder design for MIMO channels using the weighted MMSE criterion," *IEEE Trans. Commun.*, vol. 49, pp. 2198–2206, Dec. 2001.

[111] R. Gallager, *Principles of Digital Communication*. Cambridge, UK: Cambridge Univ. Press, 2008.

[112] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, 1st ed. Cambridge, UK: Cambridge Univ. Press, 1994.

[113] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge Univ. Press, 2004.

[114] Z.-Q. Luo and W. Yu, "An introduction to convex optimization for communications and signal processing," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1426–1438, Aug. 2006.

[115] J. Löfberg, "YALMIP : A toolbox for modeling and optimization in MATLAB," in *Proc. CACSD Conf.*, Taipei, Taiwan, 2004. [Online]. Available: http://users.isy.liu.se/johanl/yalmip

[116] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," http://cvxr.com/cvx//cvx, Apr. 2011.

[117] S. Ma, C. Xing, Y. Fan, Y.-C. Wu, T.-S. Ng, and H. Poor, "Iterative transceiver design for MIMO AF relay networks with multiple sources," in *Proc. Military Commun. Conf.*, San Jose, CA, USA, Nov. 2010, pp. 369–374.

[118] C. Zhao and B. Champagne, "Linear transceiver design for relay-assisted broadcast systems with diagonal scaling," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, Vancouver, Canada, May 2013, 5 pages.

[119] H. Kim, S. Lee, K. Kwak, H. Min, and D. Hong, "On the design of ZF and MMSE Tomlinson-Harashima precoding in multiuser MIMO amplify-and-forward relay system," in *Proc. 20th IEEE Int. Symp. Personal, Indoor Mobile Radio Commun.*, Tokyo, Japan, Sep. 2009, pp. 2509–2513.

[120] W. Xu, X. Dong, and W.-S. Lu, "Joint optimization for source and relay precoding under multiuser MIMO downlink channels," in *Proc. 2010 Int. Conf. Commun.*, Cape Town, South Africa, May. 2010, 5 pages.

[121] A. Phan, H. Tuan, H. Kha, and H. Nguyen, "Beamforming optimization in multi-user amplify-and-forward wireless relay networks," *IEEE Trans. Commun.*, vol. 11, pp. 1510–1520, Apr. 2012.

[122] Y. Ye, "Approximating quadratic programming with bound and quadratic constraints," *Mathematical Programming*, vol. 84, no. 2, pp. 219–226, 1999.

[123] P. I. Davies and M. I. Smith, "Updating the singular value decomposition," *J. Comput. and Applied Math.*, vol. 170, no. 1, pp. 145–167, Sep. 2004.

[124] C. Jeong, H.-M. Kim, H.-K. Song, and I.-M. Kim, "Relay precoding for non-regenerative MIMO relay systems with partial CSI in the presence of interferers," *IEEE Trans. Wireless Commun.*, vol. 11, pp. 1521–1531, Apr. 2012.