Winter 2015

Advanced wireless communications using large numbers of transmit antennas and receive nodes

Junil Choi
Purdue University

Graduate School Form 30
(Updated 11/20/2014)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Junil Choi

Entitled
Advanced Wireless Communications Using Large Numbers of Transmit Antennas and Receive Nodes

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

David J. Love

Mark R. Bell

Michael D. Zoltowski

Xiaojun Lin

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

David J. Love

Approved by Major Professor(s): _____

_____

Approved by: V. Balakrishnan 02/10/2015

Head of the Department Graduate Program Date

ADVANCED WIRELESS COMMUNICATIONS USING LARGE NUMBERS OF TRANSMIT ANTENNAS AND RECEIVE NODES

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Junil Choi

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2015

Purdue University

West Lafayette, Indiana

To my family

ACKNOWLEDGMENTS

Foremost, I would like to express my deepest gratitude to my advisor, Professor David J. Love. He is such a great mentor who continuously supports and encourages my research. Every time when I was stuck on a certain problem during my studies, he was always there to provide insights and shed light on the problem. In fact, he is not just a fabulous mentor but also like a sincere friend. I really like talking to him whatever the topics are. I truly enjoyed my life at Purdue mostly because of his kind consideration and support.

I am indebted to many other mentors. I am very grateful to Professor Michael D. Zoltowski, Professor Mark R. Bell, and Professor Xiaojun Lin for being my doctoral advisory committee and offering me many impressive courses. Their dedication has truly broadened my knowledge. I was also fortunate enough to start my research with Professor Kwang Bok Lee at Seoul National University. His endless passion toward research inspired my early graduate studies.

The life at Purdue would have been a bit dull without the members in the TASC Lab. We had such a great time both inside and outside of the lab, and I hope the tradition of fun and laughter (of course academic achievement as well!) in the lab would continue. I am also thankful to all my friends in the US and Korea.

Above all, I would like to thank my dear family. I could be who I am thanks to the dedication of my parents. My brother always has been supportive and on my side. I am sure my grandmother would be proud of me in the heaven. My appreciation to my lovely wife, Jongkyung, goes beyond words. I have known her for eighteen years, and I cannot imagine my life without her. Lastly, I praise my Lord Jesus Christ for the many blessings in my life.

TABLE OF CONTENTS

Page

Page

Page

LIST OF FIGURES

ABBREVIATIONS

MIMO	Multiple-Input Multiple-Output
MISO	Multiple-Input Single-Output
SIMO	Single-Input Multiple-Output
SNR	Signal-to-Noise Ratio
CSI	Channel State Information
FDD	Frequency Division Duplexing
TDD	Time Division Duplexing
GLP	Grassmannian Line Packing
GSP	Grassmannian Subspace Packing
RVQ	Random Vector Quantization
MSE	Mean Squared Error
MMSE	Minimum Mean Squared Error
i.i.d.	Independent and Identically Distributed
TCM	Trellis Coded Modulation
TCQ	Trellis Coded Quantization
NTCQ	Noncoherent Trellis Coded Quantization
TEC	Trellis-Extended Codebook
TE-SPA	Trellis-Extended Successive Phase Adjustment
IoT	Internet Of Things
SER	Symbol Error Rate
SCRS	Shortened Concatenated Repetition-Simplex

ABSTRACT

Choi, Junil Ph.D., Purdue University, May 2015. Advanced Wireless Communications Using Large Numbers of Transmit Antennas and Receive Nodes. Major Professor: David J. Love.

The concept of deploying a large number of antennas at the base station, often called massive multiple-input multiple-output (MIMO), has drawn considerable interest because of its potential ability to revolutionize current wireless communication systems. Most literature on massive MIMO systems assumes time division duplexing (TDD), although frequency division duplexing (FDD) dominates current cellular systems. Due to the large number of transmit antennas at the base station, currently standardized approaches would require a large percentage of the precious downlink and uplink resources in FDD massive MIMO be used for training signal transmissions and channel state information (CSI) feedback. First, we propose practical open-loop and closed-loop training frameworks to reduce the overhead of the downlink training phase. We then discuss efficient CSI quantization techniques using a trellis search. The proposed CSI quantization techniques can be implemented with a complexity that only grows linearly with the number of transmit antennas while the performance is close to the optimal case. We also analyze distributed reception using a large number of geographically separated nodes, a scenario that may become popular with the emergence of the Internet of Things. For distributed reception, we first propose coded distributed diversity to minimize the symbol error probability at the fusion center when the transmitter is equipped with a single antenna. Then we develop efficient receivers at the fusion center using minimal processing overhead at the receive nodes when the transmitter with multiple transmit antennas sends multiple symbols simultaneously using spatial multiplexing.

1. INTRODUCTION

The concept of wireless systems employing a large number of transmit antennas, often dubbed massive multiple-input multiple-output (MIMO) systems, has been evolving over the past few years. It was found in [1] that adding more antennas at the base station is always beneficial even with very noisy channel estimation because the base station can recover information even with a low signal-to-noise-ratio (SNR) once it has sufficiently many antennas. This motivates the concept of using a very large number of transmit antennas, where the number of antenna elements can be at least an order of magnitude more than the current cellular systems (10s-100s) [2]. Massive MIMO systems have the potential to revolutionize cellular deployments by accommodating a large number of users in the same time-frequency slot to boost the network capacity [3] and by increasing the range of transmission with improved power efficiency [4]. Recently, fundamental limits, optimal transmit precoding and receive strategies, and real channel measurement issues for massive MIMO systems were studied and summarized in [5] (see also the references therein).

Note that the optimal benefits of MIMO and massive MIMO systems can be achieved only when the base station and the user (or multiple users in multiuser MIMO systems) both know the channel state information (CSI) between the two perfectly. However, it is impossible for the base station and the user to know the CSI perfectly in practice. Instead, the user acquires the CSI through a training phase in which the base station transmits training signals that are known at the user a priori. To provide transmit-side CSI, the base station can learn the CSI from limited feedback in frequency division duplexing (FDD) [6] or leverage channel reciprocity in time division duplexing (TDD) [2].

Most of the massive MIMO research assumes TDD systems that rely on channel reciprocity for the base station to acquire CSI. Ideally, pilot contamination, which is

caused by using non-orthogonal uplink pilot signals in neighbouring cells, is the only factor that limits TDD massive MIMO system performance [2, 7]. Some works that mitigate pilot contamination have been proposed recently [8, 9]. However, in practice, there are other system imperfections that limit the performance of TDD massive MIMO systems. Because of calibration error in the downlink/uplink RF chains, the downlink channel estimated by the uplink channel using channel reciprocity may not be accurate [10]. Hardware impairments also can limit the performance [11,12]. Moreover, the user is not able to learn the instantaneous downlink channel (because there is no downlink training for CSI estimation in TDD massive MIMO) [7], which might cause a significant error in data decoding at the user. In addition, FDD dominates current wireless cellular systems. Thus, it is of great interest to explore backwards compatible massive MIMO upgrades for FDD wireless communication systems.

To implement FDD massive MIMO systems, we need to develop 1) a novel training technique for downlink channel estimation and 2) an efficient CSI quantization method. Note that the overhead of downlink training (relying on conventional unitary training techniques) and CSI quantization must both scale proportionally to the number of transmit antennas to enable accurate channel estimation at the user and to maintain a certain level of CSI quantization loss [13, 14]. Because of the very large number of antennas, the overhead for both unitary training and vector quantized (VQ) codebook-based CSI feedback might overwhelm the downlink and uplink resources in massive MIMO systems. Moreover, the complexity of CSI quantization using unstructured VQ codebooks increases exponentially with the feedback overhead (or the number of transmit antennas). The heavy training/feedback overhead and CSI quantization complexity problems should be solved to implement practical FDD massive MIMO systems.

Although we would be able to deploy a large number of antennas at the base station, it may be difficult to deploy many antennas at a mobile such as a smartphone or tablet due to its limited space. The limitation can be overcome by exploiting the emergence of the Internet of Things (IoT). As more and more internet-enabled things

are commonly used (e.g., computers, smartphones, tablets, home appliances, and more), the IoT will change the paradigm of communication systems [15]. In the IoT environment, devices could be used as distributed transmit and/or receive entities allowing massive distributed MIMO systems to be implemented.

There is a growing need for advanced distributed transmission and/or reception techniques that can be applied to a wide array of wireless signal processing scenarios including cellular systems, target detection in radar systems, wireless sensor networks, and military communications. Coordinated multipoint (CoMP) in the 3GPP standard [16–18] enables multiple base stations to cooperate with each other to support cell edge users using techniques such as joint transmission (JT) [19,20] or coordinated scheduling/coordinated beamforming (CS/CB) [21,22]. Distributed antenna systems (DAS) are also adopted to boost performance in cellular systems [23–25]. The geographically separated radio entities in radar systems can obtain different information of a target (or multiple targets) and make better decisions, e.g., location or speed of the target [26–28]. In wireless sensor networks, transmission/reception techniques are even more crucial because sensors are usually very cheap and only can perform simple operations [29–32]. Military communications where a squad of radio units serves as a distributed array in battlefields can be considered as a form of distributed multiple antenna systems [33, 34].

In what follows, we will develop efficient downlink training and CSI quantization techniques to implement practical FDD massive MIMO systems. Then, we will describe distributed reception taking a large number of geographically separated receive nodes into account. We will focus on two different scenarios, i.e., the cases of single and multiple transmit antennas, for distributed reception.

1.1 Downlink Training for Massive MIMO

Many papers have been dedicated to deriving the optimal training signals for *open-loop/single-shot* training frameworks and verifying their channel estimation per-

formance in FDD MIMO systems [13, 35–38]. Open-loop training means that there is no feedback information about the preferable training signal, and single-shot training refers to the case when the user estimates the channel only based on the current received training signal and discards the past received training signals. In open-loop/single-shot training, it was shown in [13] that training signals should be orthogonal to each other, and the optimal training length in time should be the same as the number of transmit antennas in uncorrelated Rayleigh fading channels. When channels are spatially correlated and the base station knows the correlation statistic perfectly, [35] and [38] showed that the optimal training dimension can be reduced when the number of statistically dominant subspaces is smaller than the number of transmit antennas. In temporally correlated channels, a Kalman filter or particle filter can be used at the user to track the channel variation between the training signal intervals [39, 40].

The amount of temporal overhead for downlink training has been assumed negligible in past MIMO scenarios because past systems used small numbers of transmit antennas. However, in FDD massive MIMO systems, the overhead of the training duration could overwhelm the precious downlink resources due to the large number of transmit antennas. Therefore, we propose practical open-loop and closed-loop training approaches with successive channel estimation for FDD massive MIMO in order to reduce the overhead of the downlink training phase.

We consider practical MIMO channels that are correlated in time *and* space. Moreover, we assume that the long-term channel statistics are known only to the user. This assumption is different from [35, 38, 41] that assume perfect knowledge of the spatial correlation at the base station. Having spatial correlation knowledge at the base station may not be practical for FDD massive MIMO systems because the user would have to explicitly feed back the knowledge of the spatial correlation matrix to the base station. Since the number of entries of the spatial correlation matrix grows quadratically with the number of transmit antennas, feedback of the

spatial correlation matrix might not be acceptable in massive MIMO systems.[1] The spatial and temporal correlation vary in time in practice, even though they are long-term channel statistics, which makes it even harder for the base station to acquire such statistics. Thus, we assume *the base station does not have any knowledge of those statistics* throughout this study.

1.2 CSI Quantization for Massive MIMO with Large Feedback Overhead

There is a large body of literature devoted to accurate CSI quantization for closed-loop MIMO FDD systems with a relatively small number of antennas [6]. Most approaches employ a common VQ codebook at the transmitter and the receiver, and the explicit feedback sequence is simply the binary index of the codeword chosen in the codebook. Thus, the main focus has been on codebook design. For i.i.d. Rayleigh fading channel models, deterministic codebook techniques using Grassmannian line packing (GLP) were developed in [43–45], and the performance of random vector quantization (RVQ) codebooks was analyzed in [14, 46]. Limited feedback codebooks that adapt to spatially correlated channels were studied in [47–49], and temporal correlated channels were developed in [50–57].

It has been shown in [46] that an RVQ codebook is asymptotically optimal for i.i.d. Rayleigh fading channels when the number of transmit antennas gets large, assuming a fixed number of feedback bits per antenna. However, existing codebook-based techniques do not scale to approach the RVQ benchmark. In order to maintain the same level of channel quantization error, the feedback overhead must increase proportional to the number of transmit antennas [14, 58]. While the linear increase in feedback overhead with the number of antennas may be acceptable as we scale to massive MIMO, the corresponding exponential increase in codebook size makes a direct look-up approach for feedback generation infeasible.

[1] When statistical reciprocity is available, it is also possible to estimate the downlink spatial correlation matrix by the uplink correlation matrix to sidestep this problem [42].

In order to address this gap in source coding techniques, it is natural to turn to the duality between source and channel coding. Just as RVQ provides a benchmark for source coding, random coding produces information-theoretic benchmarks for channel coding. However, there are thousands of papers dedicated to practical channel code designs that aim to approach these benchmarks, with codes such as convolutional codes, Reed-Solomon codes, turbo codes, and LDPC codes implemented in practice [59]. While these ideas can and have been leveraged for source coding, the measures of distortion used have been the Hamming or Euclidean distortion. Our contribution in this work is to establish and exploit the connection between source coding on the Grassmannian manifold (which is what is needed for the limited feedback application of interest to us) and channel coding for *noncoherent* communication. We coin the term *noncoherent trellis-coded quantization (NTCQ)* for the class of schemes that we propose and investigate. Our approach avoids the computational bottleneck of look-up based codebooks, with encoding complexity scaling linearly with the number of antennas, and its performance is near-optimal, approaching that of RVQ.

1.3 CSI Quantization for Massive MIMO with Reduced Feedback Overhead

NTCQ relies on standard constellation points such as PSK or QAM to quantize a channel vector, which gives a minimum feedback overhead of one bit per channel entry and can not be formulated as a straightforward extension of existing 3GPP codebooks. Thus, we propose a trellis-extended codebook (TEC) for FDD massive MIMO.

The proposed TEC adopts the same path metric as NTCQ for trellis search, but TEC utilizes low-dimensional VQ codebooks rather than constellation points. Therefore, TEC can easily satisfy backward compatibility by exploiting standard-

ized LTE or LTE-Advanced codebooks[2] and achieve a fractional number of bits per channel entry quantization to allow practical feedback overhead. TEC can utilize other codebooks, e.g., a GLP codebook [43, 44], a RVQ codebook [14, 58], a VQ optimized codebook [60]. We develop a codeword-to-branch mapping rule to maximize the performance of TEC. The numerical results show that the mapping rule gives a non-negligible gain even with the same codebook. We also investigate a codebook design methodology (instead of reusing conventional codebooks) that is optimized for TEC.

Moreover, we propose trellis-extended successive phase adjustment (TE-SPA) which functions as a differential version of TEC. TE-SPA quantizes channels successively in time and can reduce quantization loss while using fewer feedback bits per CSI vector than TEC. We show that TE-SPA can be applied to spatially correlated channels as well without any changes.

1.4 Distributed Reception with Single Transmit Antenna

For distributed reception with a single transmit antenna, we first focus on the techniques to provide diversity advantage in fading channels. We assume that there is a transmitter that wants to send a signal to a fusion center with the help of multiple geographically separated receive nodes. Each node receives the broadcasted signal from the transmitter through a fading channel and forwards the *processed* received signal to the fusion center. The fusion center then tries to decode the transmitted signal using the forwarded information from the receive nodes and, if available, channel information.

This scenario has been studied in [61] and [62] for cases when the number of processing bits at each receive node is greater than or equal to the number of bits representing data symbol constellation. Our focus is on the more practical case when each

[2]Instead of having different CSI quantization methods for different number of transmit antennas, it is desirable to reuse standardized feedback frameworks, e.g., LTE or LTE-Advanced codebooks, in practice.

receive node *quantizes* the received signal before forwarding it to the fusion center. Thus, the scenario is also related to the compress-and-forward relaying scheme [63–67]. However, the works in [63, 64] only show theoretical achievable rate regions without deeply studying any practical compression technique. Implementable schemes are proposed in [65–67] considering only a single receive node. Moreover, compress-and-forward relaying usually assumes that the fusion center (destination) receives signals not only from the receive nodes but also from the transmitter directly, which is different from our assumption. We are interested in developing practical compression techniques that can accommodate a large number of receive nodes without having a direct path between the transmitter and the fusion center.

We show that there is a strong connection between the problem of minimizing symbol error probability at the fusion center in distributed reception and channel coding in coding theory. In coding theory, time diversity in fading channels is achieved by transmitting channel coded data bits over multiple channel instances [68]. Similarly, our approach can obtain spatial diversity by exploiting multiple receive nodes that experience weakly correlated or independent channels in distributed reception. This connection allows us to utilize well-established channel coding techniques to develop good distributed reception strategies and achieve the maximum diversity gain. The achieved diversity gain by distributed reception would give range and/or data rate advantages.

The connection between the distributed reception problem and channel coding has been first explored in [69–71] for the distributed fault-tolerant classification problem in wireless sensor networks. A codeword set matrix is generated by two algorithms, i.e., cyclic column replacement and simulated annealing, for single bit and multiple bits receive node processing in [69] and [70], respectively. Each codeword (or a symbol in a signal constellation set) forms a row in the codeword set matrix and each column of the matrix represents the decision rule employed at each receive node. The proposed approaches in [69, 70], however, are heuristic and do not guarantee optimality for communication in any sense. Moreover, those approaches need complex

offline optimization to generate code matrices for every different number of the receive nodes.

In this work, we consider three general assumptions: 1) fading channels between the transmitter and the receive nodes, 2) arbitrary M-ary data symbol transmission from the transmitter, and 3) multiple bit processing at the receive nodes. To support these scenarios effectively, we propose a unified framework of compression at the receive nodes and decoding at the fusion center. We dub the unified framework as *coded receive diversity*.

1.5 Distributed Reception with Multiple Transmit Antennas

Most of the prior work on distributed reception for wireless communication systems, including our coded receive diversity, considered only detection/estimation of a single-dimensional parameter or single transmitted symbol. To our knowledge, there are few papers that discuss multi-dimensional estimation problems. A few exceptions can be found in [72, 73] which consider the estimation of a multi-dimensional vector in wireless sensor networks with additive noise at each sensor.

Thus, we consider distributed MIMO communication systems where the transmitter is equipped with multiple antennas and simultaneously transmits independent data symbols chosen from a standard M-ary constellation using spatial multiplexing to a set of geographically separated receive nodes deployed with a single receive antenna sent through independent fading channels. Each receive node quantizes its received signal and forwards the quantized signal to the fusion center. The fusion center then attempts to decode the transmitted data by exploiting the quantized signals from the receive nodes and global channel information. This scenario is likely to become popular with the emergence of massive MIMO and IoT because base stations tend to be equipped with a large number of antennas in massive MIMO systems and we can easily have a large number of receive nodes in the IoT environment.

For practical purposes, we assume each receive node quantizes its received signal with one bit per real part and one bit per imaginary part of the received signal to minimize the transmission overhead between the receive nodes and the fusion center. Quantizer design is a non-trivial problem because the receive nodes are not able to decode the transmitted symbols due to the fact that each receive node has only one antenna [74]. Instead, each receive node quantizes a single quantity, i.e., the received signal, regardless of the number of transmitted symbols. In this setup, we develop an optimal maximum likelihood (ML) receiver and a low-complexity zero-forcing (ZF)-type receiver assuming global channel knowledge at the fusion center. The ML receiver outperforms the ZF-type receiver regardless of the number of receive nodes and SNR ranges. However, the complexity of the ML receiver is excessive, especially when the number of transmitted symbols becomes large. On the other hand, the ZF-type receiver can be easily implemented and gives comparable performance to that of the ML receiver when the SNR is low to moderate, although it suffers from an error rate floor when SNR is high. The error rate floor of the ZF-type receiver can be easily mitigated by having more receive nodes.

When the SNR is high, the distributed reception problem is closely tied to work in quantized frame expansion. Linear transformation and expansion by a frame matrix in the presence of coefficient quantization is thoroughly studied in [75, 76]. A linear expansion method, which is similar to our ZF-type receiver, and its performance in terms of the mean-squared error (MSE) are analyzed based on the properties of a frame matrix. An advanced non-linear expansion method relying on linear programming is also studied. The major difference compared to our problem setting is that [75,76] do not assume any additive noise before quantization, while our scenario considers a fading channel (which corresponds to a frame matrix in frame expansion) with additive noise prior to quantization at the receive nodes. We rely on some of the analytical results from [75] for evaluating and modifying the ZF-type receiver later.

1.6 Outline

The dissertation is organized as follows. In Chapter 2, downlink training techniques for FDD massive MIMO systems are considered. We first analyze the limitation of using conventional downlink training techniques in temporally and spatially correlated massive MIMO channels. Then we propose open-loop/closed-loop downlink training with memory to overcome the limitation of conventional training techniques. In Chapter 3, we explain the necessity of designing novel efficient CSI quantization techniques for FDD massive MIMO systems. Using the duality between source coding and channel coding, we propose NTCQ of which the complexity increases linearly with the number of transmit antennas. In Chapter 4, we propose TEC and TE-SPA, which can achieve a fractional number of bits per channel entry quantization, to reduce the feedback overhead of NTCQ. In Chapter 5, we study distributed reception with a single transmit antenna. We exploit the connection between the problem of minimizing the symbol error probability at the fusion center in distributed reception and channel coding in coding theory and design a unified framework for coded distributed diversity reception. In Chapter 6, we consider distributed reception with multiple transmit antennas. We design efficient receivers at the fusion center using minimal quantized information from the receive nodes. Chapter 7 concludes the dissertation.

1.7 Notation

Upper and lower boldface symbols are used to denote matrices and column vectors, respectively. $\mathbf{X}^H$, $\mathbf{X}^T$, $\mathbf{X}^{-1}$, $\mathbf{X}^{\frac{1}{2}}$, and $\mathrm{tr}(\mathbf{X})$ are used as the Hermitian transpose, transpose, inverse, square-root, and the trace of $\mathbf{X}$, respectively. $\mathbf{I}_k$ is the $k \times k$ identity matrix, $\mathbf{0}_m$ represents the $m \times 1$ all zero vector, $\mathbf{1}_m$ denotes an $m \times 1$ all one vector, and $\mathbf{X}_{[k:m]}$ represents the sub-matrix of $\mathbf{X}$ formed by the k-th column to the m-th column, inclusively. $\|\mathbf{X}\|$ and $\|\mathbf{X}\|_F$ are used as the two-norm and the Frobenius norm of a matrix $\mathbf{X}$, respectively. We let $\lambda(\mathbf{X})$ denote the vector with the eigenvalues of the matrix $\mathbf{X}$ in decreasing order as its elements. $\mathrm{Re}(\mathbf{x})$ and $\mathrm{Im}(\mathbf{x})$ denote the real

and complex part of a complex vector $\mathbf{x}$, respectively. $\mathbb{C}^m$ ($\mathbb{R}^m$) and $\mathbb{C}^{m\times n}$ ($\mathbb{R}^{m\times n}$) represent the set of all $m \times 1$ complex (real) vectors and the set of all $m \times n$ complex (real) matrices, respectively. $\mathrm{mod}(a, b)$ is the remainder of a when divided by b and $\mathcal{CN}(\bar{\mathbf{x}}, \mathbf{R})$ is used to denote the complex Gaussian random vector distribution with mean $\bar{\mathbf{x}}$ and covariance $\mathbf{R}$. The expectation operation is denoted by $E[\cdot]$ and $Pr(A)$ denotes the probability of event A.

2. DOWNLINK TRAINING TECHNIQUES FOR FDD MASSIVE MIMO SYSTEMS: OPEN-LOOP AND CLOSED-LOOP TRAINING WITH MEMORY

We study downlink training techniques for FDD massive MIMO in this chapter. We first explain the limitations of conventional single-shot training, which only relies on the most recently received training signal to estimate the channel. The analysis shows that the average received SNR quickly saturates to a certain level as the number of transmit antennas gets large with a fixed training length (that is less than the number of transmit antennas), and this SNR ceiling could preclude its use in massive MIMO systems.

Then, we propose open-loop and closed-loop training frameworks with memory to effectively alleviate the SNR ceiling effect. Although the ceiling effect cannot be perfectly eliminated with a fixed training length, the proposed training frameworks can significantly increase the ceiling level. We assume the base station and the user share a common set of training signals where each training signal has a much lower rank than the number of transmit antennas. In open-loop training, the base station transmits training signals in a round-robin manner, and the user predicts/estimates the channel based on previous channel estimates. Thus, the proposed framework can be considered *open-loop training with memory*. With this approach, we can reduce the number of training channels needed to acquire a *good* channel estimate to a reasonable range even with a large number of transmit antennas.

In *closed-loop training with memory*, which had initial results presented in [77] and was studied for the stationary channel in [78], the user selects the best training signal

based on the prior knowledge of the channel and previously received training signals. The user feeds back the index of the selected training signal and the base station relies on the fed back information for the next training phase. This framework is considerably different from current wireless systems that use pre-determined training signals in time and frequency [8, 79, 80]. By allowing a small amount of feedback, we can further improve channel estimation performance with less training overhead. We develop two objective functions to select the training signal at the user: 1) minimizing MSE and 2) maximizing the average received SNR for the data communication phase. Numerical studies show that the second approach can improve the received SNR when the number of transmit antennas is moderately large. We also develop an effective way of designing the set of training signals used for the proposed training frameworks.

Finally, we identify preferable channel conditions and system parameters for closed-loop training with memory. The performance gain of closed-loop training becomes larger when 1) the SNR is low, 2) the number of transmit antennas is large relative to the length of the training phase, or 3) the prior channel estimate is not accurate at the beginning of the communication setup, all of which could be commonly true for massive MIMO systems. Simulation results confirm these analyses.

2.1 System Model

We consider an N_t transmit antennas and single receive antenna MISO system transmitting over a block-fading channel. Although we only consider MISO channels for simplicity, our framework can be easily extended to general MIMO channels with the vectorization approach in [35, 38]. We assume the block-fading channel has a coherence time of L, which means that the channel is static for L channel uses in each block and changes from block-to-block. The input-output relation for the ℓ-th channel use in the i-th fading block is given by

$$y_i[\ell] = \mathbf{h}_i^H \mathbf{x}_i[\ell] + n_i[\ell], \tag{2.1}$$

where $y_i[\ell]$ is the received signal, $\mathbf{h}_i \in \mathbb{C}^{N_t}$ is the channel vector, $\mathbf{x}_i[\ell] \in \mathbb{C}^{N_t}$ is the transmitted signal with $E[\|\mathbf{x}_i[\ell]\|^2] = \rho$, and $n_i[\ell] \sim \mathcal{CN}(0,1)$ is normalized additive white Gaussian noise at the user.

Each channel block consists of a *training* phase and a *data communication* phase. We assume that the first $T < L$ channel uses and the remaining $L - T$ channel uses are dedicated for training and data communication, respectively. We further assume that $T < N_t$ because we consider massive MIMO. For the i-th fading block, the received training signals $y_i[\ell]$ for $\ell = 0, \ldots, T-1$ can be collected into a vector form as $\mathbf{y}_{i,train} = [y_i[0] \ \cdots \ y_i[T-1]]^T$. Then, the input-output relation in (2.1) can be rewritten as

$$\mathbf{y}_{i,train} = \mathbf{X}_i^H \mathbf{h}_i + \mathbf{n}_{i,train}, \tag{2.2}$$

where $\mathbf{X}_i = [\mathbf{x}_i[0] \ \cdots \ \mathbf{x}_i[T-1]]$ is the transmitted signals collected into an $N_t \times T$ matrix and $\mathbf{n}_{i,train} = [n_i[0] \ \cdots \ n_i[T-1]]^T$.

Note that *unitary training* with equal power allocation per pilot symbol which restricts $\mathbf{X}_i$ such that

$$\mathbf{X}_i \in \mathcal{X} = \left\{ \mathbf{F} : \mathbf{F} \in \mathbb{C}^{N_t \times T}, \ \mathbf{F}^H \mathbf{F} = \rho \mathbf{I}_T \right\}, \tag{2.3}$$

is optimal in i.i.d. Rayleigh fading channels. We also rely on unitary training throughout this work because we assume that the base station does not have any prior knowledge of the channel statistics to adapt training signals.[1]

During the $L - T$ data communication channel uses, we assume that the base station employs beamforming, and the transmitted signal is written as

$$\mathbf{x}_i[\ell] = \mathbf{w}_i s_i[\ell],$$

[1]If the base station knows the channel statistics, then non-unitary training with power allocation can give a better performance than unitary training in spatially correlated channels.

where $\mathbf{w}_i$ is a beamforming vector with $\|\mathbf{w}_i\| = 1$ and $s_i[\ell]$ is a data symbol with $E[|s_i[\ell]|^2] = \rho$. With this setup, the normalized average received SNR of the i-th fading block at the user is

$$\Gamma_i = \frac{1}{\rho} E\left[|\mathbf{h}_i^H \mathbf{x}_i[\ell]|^2\right] = E\left[|\mathbf{h}_i^H \mathbf{w}_i|^2\right]. \tag{2.4}$$

The optimal training signal $\mathbf{X}_i$ is highly dependent on the channel statistics and the desired performance metric. Aside from the few works, including [35, 38] that assume spatially correlated channels and [40] that assumes temporally correlated channels, most research on training considers uncorrelated channels both in time and space. In this work, we consider a general and practical channel model, i.e., spatially *and* temporally correlated channels. We assume $\mathbf{h}_i$ follows a Gauss-Markov distribution according to

$$\begin{aligned} \mathbf{h}_0 &= \mathbf{R}^{\frac{1}{2}} \mathbf{g}_0, \\ \mathbf{h}_i &= \eta \mathbf{h}_{i-1} + \sqrt{1-\eta^2} \mathbf{R}^{\frac{1}{2}} \mathbf{g}_i, \quad i \geq 1, \end{aligned} \tag{2.5}$$

where $\mathbf{R} = E\left[\mathbf{h}_i \mathbf{h}_i^H\right]$ is a spatial correlation matrix,[2] $\mathbf{g}_i$ is an innovation process with independent and identically distributed (i.i.d.) entries distributed according to $\mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_t})$ for all i, and $0 \leq \eta \leq 1$ is a temporal correlation coefficient. We assume $\mathbf{h}_0$ is independent of $\mathbf{g}_i$ for all $i \geq 1$. Because the spatial correlation matrix $\mathbf{R}$ is a Hermitian positive definite matrix, it can be decomposed as $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$ where $\mathbf{U}$ and $\mathbf{\Lambda} = \mathrm{diag}\left([\lambda_1, \lambda_2, \cdots, \lambda_{N_t}]\right)$ are the eigenvector and eigenvalue matrices of $\mathbf{R}$, respectively. We assume the λ_k's are in decreasing order as $\lambda_1 \geq \cdots \geq \lambda_{N_t}$ and $\mathrm{tr}\left(\mathbf{R}\right) = \sum_{t=1}^{N_t} \lambda_t = N_t$. As mentioned in the introduction, we assume *the base station does not have any knowledge of the channel statistics such as* $\mathbf{R}$ *and* η throughout the work.

[2] $\mathbf{R}$ is closely related to the antenna spacing at the base station and the user location. We assume that $\mathbf{R}$ is fixed in time because the user location does not change much with moderate user velocities, e.g., 3-10km/h.

2.2 Single-Shot Training and the Ceiling Effect

In most prior work on training, the user discards the previously received training signals $\{\mathbf{y}_{k,train}\}_{k=0}^{i-1}$ and estimates $\mathbf{h}_i$ based only on the current received training signal $\mathbf{y}_{i,train}$. We first explain the conventional single-shot training framework and derive the structure of the optimal training signal $\mathbf{X}_{i,\text{opt}}$ for single-shot training at the i-th fading block assuming there is an *unlimited feedback channel* for $\mathbf{X}_{i,\text{opt}}$ from the user to the base station ($\mathbf{X}_{i,\text{opt}}$ is available only at the user because the base station does not know the channel statistics). Based on an upper bound on training performance for single-shot training using the optimal training signal, we show that deploying a large number of transmit antennas does not increase performance (i.e., the normalized average received SNR Γ_i in (2.4)) with N_t for most practical channel conditions.

We drop the fading block index i from the notation throughout this section because of the lack of dependence on the specific block during channel estimation.

2.2.1 Structure of the optimal training signal of single-shot training

We focus on MMSE channel estimation at the user. Assuming $\mathbf{h}$ is complex Gaussian with mean $\mathbf{0}$ and covariance $\mathbf{R}$, we can derive the MMSE estimate of the channel $\mathbf{h}$ given the observation $\mathbf{y}_{train}$ in (2.2) as [81]

$$\begin{aligned}\widehat{\mathbf{h}} &= E[\mathbf{h} \mid \mathbf{y}_{train}] \\ &= \mathbf{R}\mathbf{X}\left(\mathbf{I}_T + \mathbf{X}^H\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{y}_{train}.\end{aligned}$$

This estimate $\widehat{\mathbf{h}}$ is complex Gaussian with mean $\mathbf{0}$ and covariance

$$\mathbf{R}_{\widehat{\mathbf{h}}} = \mathbf{R}\mathbf{X}\left(\mathbf{I}_T + \mathbf{X}^H\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^H\mathbf{R}.$$

The MSE of channel estimation is given by

$$\begin{aligned}\mathrm{MSE}\left(\mathbf{X}\right) &= \frac{1}{N_t}E\left[\|\mathbf{h}-\widehat{\mathbf{h}}\|^2\right] \\ &= \frac{1}{N_t}\mathrm{tr}\left(\mathbf{R}-\mathbf{R}\mathbf{X}\left(\mathbf{I}_T+\mathbf{X}^H\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^H\mathbf{R}\right), \end{aligned} \tag{2.6}$$

and MMSE estimation minimizes the MSE between $\mathbf{h}$ and $\widehat{\mathbf{h}}$ for a given $\mathbf{X}$.

As mentioned in (2.3), we assume $\mathbf{X}\in\mathcal{X}$. If the base station relies on a pre-defined $\mathbf{X}$ for training, we call it *open-loop/single-shot training*. If there is a feedback channel from the user to the base station to inform the best training signal for single-shot training, we call this scheme *closed-loop/single-shot training*. Then, similar to the derivation in [35, 38], the following lemma shows the optimal structure of $\mathbf{X}_{\mathrm{ss,opt}}$ that minimizes the $\mathrm{MSE}\left(\mathbf{X}\right)$ in (2.6) for closed-loop/single-shot training with unlimited feedback.

Lemma 2.2.1 *The optimal $N_t\times T$ $(N_t\geq T)$ training signal for closed-loop/single-shot training with full feedback for $\mathbf{X}_{\mathrm{ss,opt}}$ that minimizes the* $\mathrm{MSE}\left(\mathbf{X}\right)$ *is given as*

$$\begin{aligned}\mathbf{X}_{\mathrm{ss,opt}} &= \underset{\mathbf{X}\in\mathcal{X}}{\mathrm{argmin}}\,\mathrm{MSE}\left(\mathbf{X}\right) \\ &= \sqrt{\rho}\mathbf{U}_{[1:T]} \end{aligned} \tag{2.7}$$

where $\mathbf{R}=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$.

Proof See Appendix A.1. ■

Lemma 2.2.1 implies we should transmit the training signal along the first T dominant eigen-directions of $\mathbf{R}$ to minimize the MSE. With $\mathbf{X}_{\text{ss,opt}} = \sqrt{\rho}\mathbf{U}_{[1:T]}$, we can derive the MSE as

$$\begin{aligned}\text{MSE}\left(\mathbf{X}_{\text{ss,opt}}\right) &= 1 - \frac{1}{N_t}\,\text{tr}\left(\mathbf{R}\mathbf{X}_{\text{ss,opt}}\left(\mathbf{I}_T + \mathbf{X}_{\text{ss,opt}}^H\mathbf{R}\mathbf{X}_{\text{ss,opt}}\right)^{-1}\mathbf{X}_{\text{ss,opt}}^H\mathbf{R}\right)\\ &= 1 - \frac{1}{N_t}\,\text{tr}\left(\left(\mathbf{I}_T + \mathbf{X}_{\text{ss,opt}}^H\mathbf{R}\mathbf{X}_{\text{ss,opt}}\right)^{-1}\mathbf{X}_{\text{ss,opt}}^H\mathbf{R}^2\mathbf{X}_{\text{ss,opt}}\right)\\ &\overset{(a)}{=} 1 - \frac{1}{N_t}\sum_{t=1}^{T}\frac{\rho\lambda_t^2}{\rho\lambda_t + 1} \end{aligned} \tag{2.8}$$

where (a) follows from $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$. From (2.8), we state following lemma[3] and corollaries, which are intuitive.

Lemma 2.2.2 *Let $\mathbf{R}_H$ and $\mathbf{R}_L$ denote two $N_t \times N_t$ spatial correlation matrices. We assume $\lambda\left(\mathbf{R}_H\right)$ majorizes $\lambda\left(\mathbf{R}_L\right)$, i.e., $\lambda\left(\mathbf{R}_H\right) \succ \lambda\left(\mathbf{R}_L\right)$ which corresponds to the case when $\mathbf{R}_H$ is more spatially correlated than $\mathbf{R}_L$ [38]. We let $\mathbf{X}_H$ and $\mathbf{X}_L$ denote the optimal $N_t \times T$ orthogonal single-shot training signals for channels correlated with $\mathbf{R}_H$ and $\mathbf{R}_L$, respectively. Then, we have*

$$\text{MSE}\left(\mathbf{X}_H\right) \leq \text{MSE}\left(\mathbf{X}_L\right).$$

Proof See Appendix A.2. ■

Corollary 2.2.1 *If ρ and $\{\lambda_t\}_{t=1}^T$ are fixed and $T_1 > T_2 \geq 1$, then*

$$\text{MSE}\left(\mathbf{X}_{\text{ss,opt}}(T_1)\right) < \text{MSE}\left(\mathbf{X}_{\text{ss,opt}}(T_2)\right).$$

Corollary 2.2.2 *If T and $\{\lambda_t\}_{t=1}^T$ are fixed and $\rho_1 > \rho_2 > 0$, then*

$$\text{MSE}\left(\mathbf{X}_{\text{ss,opt}}(\rho_1)\right) < \text{MSE}\left(\mathbf{X}_{\text{ss,opt}}(\rho_2)\right).$$

[3]The result similar to Lemma 2 was already proven in Theorem 2 of [38]; however, we believe the proof in this work is of value due to its simplicity.

Lemma 2.2.2, Corollary 2.2.1 and 2.2.2 show that the channel can be estimated with lower MSE when channels are highly correlated, using more time on training, or training with higher transmit power. Although the above statements are for the case of $\mathbf{X}_{\text{ss,opt}} = \sqrt{\rho}\mathbf{U}_{[1:T]}$, numerical results in Section 2.4 show that these statements also hold for a general training signal $\mathbf{X}$.

2.2.2 Ceiling effect of single-shot training

We assume that the user can feed back not only $\mathbf{X}_{\text{ss,opt}}$ but also the estimated channel $\widehat{\mathbf{h}}$ perfectly to the base station to focus only on the effect of training. We refer to later chapters and references therein that discuss the downlink CSI quantization problem in FDD massive MIMO systems. The base station can then set the beamforming vector to $\mathbf{w} = \frac{\widehat{\mathbf{h}}}{\|\widehat{\mathbf{h}}\|}$. Based on $\mathbf{X}_{\text{ss,opt}}$ and $\mathbf{w}$, we derive an upper bound of the normalized average received SNR using single-shot training in the following lemma.

Lemma 2.2.3 *With the training signal* $\mathbf{X}_{\text{ss,opt}} = \sqrt{\rho}\mathbf{U}_{[1:T]}$ *and the beamforming vector* $\mathbf{w} = \frac{\widehat{\mathbf{h}}}{\|\widehat{\mathbf{h}}\|}$*, the normalized average received SNR of single-shot training,* $\Gamma_{\text{ss,opt}}$*, can be upper bounded as*

$$\Gamma_{\text{ss,opt}} = E\left[\left|\mathbf{h}^H \frac{\widehat{\mathbf{h}}}{\|\widehat{\mathbf{h}}\|}\right|^2\right] \leq \sum_{t=1}^{T} \frac{\rho\lambda_t^2}{\rho\lambda_t + 1} + \lambda_1 \tag{2.9}$$

where $1 \leq T \leq N_t$ *and* λ_t *is the* t*-th dominant eigenvalue of* $\mathbf{R}$*.*

Proof See Appendix A.3. ■

The upper bound in (2.9) becomes trivial when the rank of $\mathbf{R}$ is 1, i.e., $\text{tr}(\mathbf{R}) = \lambda_1 = N_t$. However, (2.9) is a non-trivial upper bound in general.

Lemma 2.2.3 shows that $\Gamma_{\text{ss,opt}}$ is not a linearly increasing function of N_t although the impact of N_t is implicitly reflected in λ_t. With the extreme case of i.i.d. Rayleigh fading channels where $\lambda_t = 1$ for all t, $\Gamma_{\text{ss,opt}}$ is fixed to a constant with a given T and

ρ even when $N_t \to \infty$. Unless the dominant eigen-directions contain most of the gain of the wireless channel, which is rarely the case even in highly correlated channels in practice, the gain of having a large number of antennas will saturate eventually.

Now we verify Lemma 2.2.3 with Rayleigh fading channels which are spatial correlated with the exponential model of $\mathbf{R}$ that is given as[4]

$$\mathbf{R} = \begin{bmatrix} 1 & a & \cdots & a^{N_t-1} \\ a & 1 & & \\ \vdots & & \ddots & \\ a^{N_t-1} & & & 1 \end{bmatrix} \tag{2.10}$$

where $0 < a < 1$ is a real number. The amount of spatial correlation is controlled by a, i.e., a larger (smaller) value of a corresponds to highly (loosely) correlated channels in space. When $a = 0$, we have i.i.d. Rayleigh fading channels.

Before showing the numerical results, we state the following corollary which use the upper bound of the maximum eigenvalue of $\mathbf{R}$ of the exponential model [82]

$$\lambda_1 \leq \frac{1+a}{1-a}.$$

Corollary 2.2.3 *With the exponential model of* $\mathbf{R}$ *in* (2.10)*,* $\Gamma_{\text{ss,opt}}$ *can be further upper bounded as*

$$\Gamma_{\text{ss,opt}} \leq \sum_{t=1}^{T} \frac{\rho\lambda_t^2}{\rho\lambda_t + 1} + \lambda_1 < (T+1)\lambda_1 \leq (T+1)\frac{1+a}{1-a}. \tag{2.11}$$

Corollary 2.2.3 states that the maximum $\Gamma_{\text{ss,opt}}$ is a function of T and a, not N_t.

In Fig. 2.1, we plot $\Gamma_{\text{ss,opt}}$ (in dB scale) based on simulation and the upper bounds in (2.9) and (2.11). From the figure, we see that $\Gamma_{\text{ss,opt}}$ saturates even with the optimal $\mathbf{X}_{\text{ss,opt}}$ and very highly correlated case of $a = 0.9$. Note that the maximum possible value of normalized average received SNR is the same as the number of transmit

[4]We adopt the exponential model of $\mathbf{R}$ for simulation purposes. Other structures of $\mathbf{R}$ such as a Kronecker model can be adopted as well.

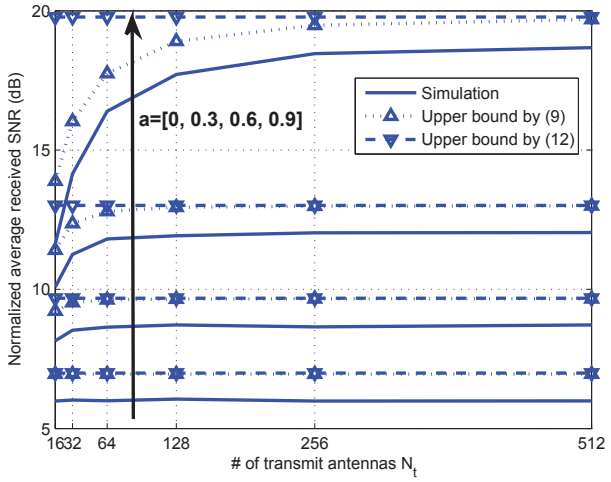


Fig. 2.1.: Plots of $\Gamma_{\text{ss,opt}}$ (in dB scale) with simulation results and the upper bounds in (2.9) and (2.11) with $\rho = 20$dB and $T = 4$. The ordered a values by the arrow correspond with the curves moving from bottom to top.

antennas N_t. We can increase $\Gamma_{\text{ss,opt}}$ by using a large number of channel uses T for training, but this will decrease the number of channel uses $T - L$ for the actual data communication.

The ceiling effect can be effectively reduced by exploiting the temporal channel correlation. Although temporal correlation is present essentially for all wireless communication systems, this correlation is not widely exploited in most MIMO channel estimation and training works. Training for massive MIMO systems should leverage temporal correlation of the channel to maximize the benefit of having a large number of antennas.

2.3 Proposed Training Frameworks

In this section, we first explain *open-loop training with memory* that does not require any feedback from the user to the base station. We then propose the framework of *closed-loop training with memory*. We derive a performance upper bound of closed-loop training with memory assuming the perfect feedback of the training signal from

the user to the base station. We also present an effective way of designing the set of training signals. Finally, we derive preferable system parameters for closed-loop training compared to open-loop training with memory.

2.3.1 Open-loop training with memory

In the proposed open-loop training with memory, we assume that the base station and the user share a common set of training signals that can be indexed with B bits given by[5] $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_{2^B}\}$. Then, training signal for the i-th fading block $\mathbf{X}_i$ is given as

$$\mathbf{X}_i = \mathbf{P}_{\mathrm{mod}(i,2^B)+1}, \quad i = 0, \ldots, T-1, \tag{2.12}$$

in a round-robin manner, which requires no feedback for the training signal from the user to the base station. However, the user estimates the channel $\mathbf{h}_i$ based not only on $\mathbf{y}_{i,train}$ but also on $\{\mathbf{y}_{k,train}\}_{k=0}^{i-1}$ and the channel statistics η and $\mathbf{R}$. Note that this problem is similar to state prediction in dynamical systems. With the training problem formulation, (2.5) specifies the state evolution and (2.2) is the input-output equation [81]. Thus, the user can rely on the Kalman filter (or a more advanced filter such as the particle filter in [40]) to track the channel evolution and provide a more accurate channel estimate.

To begin with, we denote

$$\widehat{\mathbf{h}}_{i_1|i_2} = E\left[\mathbf{h}_{i_1} \mid \{\mathbf{y}_{k,train}\}_{k=0}^{i_2}\right]$$

as the predicted value of $\mathbf{h}_{i_1}$ given $\{\mathbf{y}_{k,train}\}_{k=0}^{i_2}$ for $i_1 \geq i_2$. Then, we can define the sequential MMSE estimator $\widehat{\mathbf{h}}_{i|i}$ based on $\{\mathbf{y}_{k,train}\}_{k=0}^{i}$ as in Table 2.1. Note that the distribution of $\widehat{\mathbf{h}}_{i|i}$ given $\{\mathbf{y}_{k,train}\}_{k=0}^{i-1}$ is complex Gaussian with mean $\widehat{\mathbf{h}}_{i|i-1}$ and covariance

$$\mathbf{R}_{p,i} = \mathbf{R}_{i|i-1}\mathbf{X}_i \left(\mathbf{I}_T + \mathbf{X}_i^H \mathbf{R}_{i|i-1}\mathbf{X}_i\right)^{-1} \mathbf{X}_i^H \mathbf{R}_{i|i-1}.$$

[5]Our framework can also be combined with a time-varying $\mathbf{P}$ similar to how differential codebooks are used in CSI quantization [55, 56] for better performance.

Table 2.1.: Sequential MMSE channel estimation based on the Kalman filter [81].

Initialization:

$$\widehat{\mathbf{h}}_{0|-1} = \mathbf{0},$$
$$\mathbf{R}_{0|-1} = \mathbf{R} = E\left[\mathbf{h}_0\mathbf{h}_0^H\right].$$

Prediction:

$$\widehat{\mathbf{h}}_{i|i-1} = \eta\widehat{\mathbf{h}}_{i-1|i-1}.$$

Minimum prediction MSE matrix ($N_t \times N_t$):

$$\mathbf{R}_{i|i-1} = \eta^2\mathbf{R}_{i-1|i-1} + (1-\eta^2)\mathbf{R}.$$

Kalman gain matrix ($N_t \times T$):

$$\mathbf{K}_i = \mathbf{R}_{i|i-1}\mathbf{X}_i\left(\mathbf{I}_T + \mathbf{X}_i^H\mathbf{R}_{i|i-1}\mathbf{X}_i\right)^{-1}.$$

Correction:

$$\widehat{\mathbf{h}}_{i|i} = \widehat{\mathbf{h}}_{i|i-1} + \mathbf{K}_i\left(\mathbf{y}_{i,train} - \mathbf{X}_i^H\widehat{\mathbf{h}}_{i|i-1}\right).$$

Minimum MSE matrix ($N_t \times N_t$):

$$\mathbf{R}_{i|i} = \left(\mathbf{I}_{N_t} - \mathbf{K}_i\mathbf{X}_i^H\right)\mathbf{R}_{i|i-1}.$$

Because we assume perfect CSI feedback from the user to the base station, the beamforming vector becomes

$$\mathbf{w}_i = \frac{\widehat{\mathbf{h}}_{i|i}}{\|\widehat{\mathbf{h}}_{i|i}\|}. \tag{2.13}$$

From the numerical results in Section 2.4, open-loop training with memory can significantly increase the channel estimation performance.

2.3.2 Closed-loop training with memory

We assume the channels are correlated in time and space. Thus, the training signal at the i-th fading block can be adapted using the channel statistics and the previously received training signals $\{\mathbf{y}_{k,train}\}_{k=0}^{i-1}$ if they are available to the transmitter. Because the base station will not have direct access to the channel statistics and $\{\mathbf{y}_{k,train}\}_{k=0}^{i-1}$,

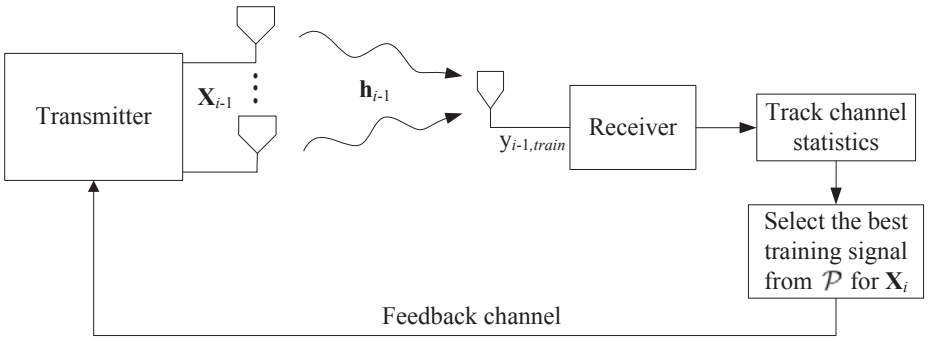


Fig. 2.2.: Concept of closed-loop training.

the best training signal $\mathbf{P}_{i,\text{best}}$ is selected from a predefined set of training signals $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_{2^B}\}$ at the user and sent back to the base station with B bits of feedback. The base station then uses the fed back signal as the training signal for the i-th fading block. The training signal selection at the user is based on using channel prediction to track the statistics of the channel at the i-th fading block conditioned on the user's side information as explained in open-loop training with memory. The conceptual explanation of closed-loop training with memory is given in Fig. 2.2.

We propose two metrics for selecting $\mathbf{P}_{i,\text{best}}$ at the user, i.e., minimizing the MSE of channel estimation and maximizing the normalized average received SNR for the data communication phase.

1) Minimizing the MSE (MSE-based): It is easy to show that the MSE between $\mathbf{h}_i$ and $\widehat{\mathbf{h}}_{i|i}$ is a function of $\mathbf{X}_i$ and given as

$$\begin{aligned}\text{MSE}\left(\mathbf{X}_i\right) &= \frac{1}{N_t} E\left[\left\|\mathbf{h}_i - \widehat{\mathbf{h}}_{i|i}\right\|^2\right] \\ &= \frac{1}{N_t}\text{tr}\left(\mathbf{R}_{i|i}\right) \\ &= \frac{1}{N_t}\text{tr}\left(\mathbf{R}_{i|i-1} - \mathbf{R}_{p,i}\right). \end{aligned} \tag{2.14}$$

Therefore, the user selects $\mathbf{P}_{i,\text{best}}$ for the i-th block that minimizes the MSE as

$$\begin{aligned}\mathbf{P}_{i,\text{best}} &= \underset{\mathbf{P}_k \in \mathcal{P}}{\operatorname{argmin}} \operatorname{MSE}(\mathbf{P}_k) \\ &= \underset{\mathbf{P}_k \in \mathcal{P}}{\operatorname{argmax}} \operatorname{tr}(\mathbf{R}_{p,i}),\end{aligned} \tag{2.15}$$

and feeds back the B-bit index of $\mathbf{P}_{i,\text{best}}$ to the transmitter. Then, the base station uses $\mathbf{X}_i = \mathbf{P}_{i,\text{best}}$ for the training signal for the i-th block.

2) Maximizing the normalized average received SNR (SNR-based): Using the beamforming vector $\mathbf{w} = \frac{\widehat{\mathbf{h}}_{i|i}}{\|\widehat{\mathbf{h}}_{i|i}\|}$ as in (2.13), Γ_i given $\{\mathbf{X}_k, \mathbf{y}_{k,train}\}_{k=0}^{i}$ becomes

$$\begin{aligned}\Gamma_i\left(\{\mathbf{X}_k, \mathbf{y}_{k,train}\}_{k=0}^{i}\right) &= E\left[\left|\mathbf{h}_i^H \mathbf{w}\right|^2 \mid \{\mathbf{X}_k, \mathbf{y}_{k,train}\}_{k=0}^{i}\right] \\ &= \mathbf{w}^H\left(\widehat{\mathbf{h}}_{i|i}\widehat{\mathbf{h}}_{i|i}^H + \mathbf{R}_{i|i}\right)\mathbf{w} \\ &= \left\|\widehat{\mathbf{h}}_{i|i}\right\|^2 + \frac{\widehat{\mathbf{h}}_{i|i}^H \mathbf{R}_{i|i}\widehat{\mathbf{h}}_{i|i}}{\left\|\widehat{\mathbf{h}}_{i|i}\right\|^2}.\end{aligned} \tag{2.16}$$

The user maximizes the expected value of (2.16) averaged over $\mathbf{y}_{i,train}$ by selecting $\mathbf{P}_{i,\text{best}}$ as

$$\begin{aligned}\mathbf{P}_{i,\text{best}} &= \underset{\mathbf{P}_i \in \mathcal{P}}{\operatorname{argmax}} E\left[\Gamma_i(\mathbf{P}_i, \mathbf{y}_{i,train}) \,|\{\mathbf{X}_k, \mathbf{y}_{k,train}\}_{k=0}^{i-1}\right] \\ &= \underset{\mathbf{P}_i \in \mathcal{P}}{\operatorname{argmax}} \Gamma_i\left(\mathbf{P}_i, \{\mathbf{X}_k, \mathbf{y}_{k,train}\}_{k=0}^{i-1}\right),\end{aligned} \tag{2.17}$$

with the expectation taken over $\mathbf{y}_{i,train}$.

We can evaluate $\Gamma_i\left(\mathbf{P}_i, \{\mathbf{X}_k, \mathbf{y}_{k,train}\}_{k=0}^{i-1}\right)$ in (2.17) as

$$\Gamma_i\left(\mathbf{P}_i, \{\mathbf{X}_k, \mathbf{y}_{k,train}\}_{k=0}^{i-1}\right) = \operatorname{tr}(\mathbf{R}_{p,i}) + \left\|\widehat{\mathbf{h}}_{i|i-1}\right\|^2 + q(\mathbf{P}_i)$$

where $q(\mathbf{P}_i)$ is defined as

$$q(\mathbf{P}_i) = E\left[\left.\frac{\widehat{\mathbf{h}}_{i|i}^H \mathbf{R}_{i|i} \widehat{\mathbf{h}}_{i|i}}{\left\|\widehat{\mathbf{h}}_{i|i}\right\|^2}\right| \{\mathbf{X}_k, \mathbf{y}_{k,train}\}_{k=0}^{i-1}\right].$$

By defining $\alpha_1 = \widehat{\mathbf{h}}_{i|i}^H \mathbf{R}_{i|i} \widehat{\mathbf{h}}_{i|i}$ and $\alpha_2 = \left\|\widehat{\mathbf{h}}_{i|i}\right\|^2$, we can approximate $q(\mathbf{P}_i)$ as [83]

$$q(\mathbf{P}_i) \approx \frac{E[\alpha_1]}{E[\alpha_2]}\left(1 - \frac{Cov(\alpha_1, \alpha_2)}{E[\alpha_1] \cdot E[\alpha_2]} + \frac{Var(\alpha_2)}{(E[\alpha_2])^2}\right)$$

where

$$\begin{aligned}
E[\alpha_1] &= \widehat{\mathbf{h}}_{i|i-1}^H \mathbf{R}_{i|i} \widehat{\mathbf{h}}_{i|i-1} + \operatorname{tr}\left(\mathbf{R}_{i|i}\mathbf{R}_{p,i}\right), \\
E[\alpha_2] &= \|\widehat{\mathbf{h}}_{i|i-1}\|^2 + \operatorname{tr}\left(\mathbf{R}_{p,i}\right), \\
Var(\alpha_2) &= 4\widehat{\mathbf{h}}_{i|i-1}^H \mathbf{R}_{p,i} \widehat{\mathbf{h}}_{i|i-1} + 2\operatorname{tr}\left(\mathbf{R}_{p,i}\right)^2, \\
Cov(\alpha_1, \alpha_2) &= 4\widehat{\mathbf{h}}_{i|i-1}^H \mathbf{R}_{i|i}\mathbf{R}_{p,i} \widehat{\mathbf{h}}_{i|i-1} + 2\operatorname{tr}\left(\mathbf{R}_{i|i}\mathbf{R}_{p,i}\mathbf{R}_{p,i}\right).
\end{aligned}$$

Thus, $\mathbf{P}_{i,\text{best}}$ can be selected by the user according to

$$\mathbf{P}_{i,\text{best}} = \underset{\mathbf{P}_i \in \mathcal{P}}{\operatorname{argmax}} \left(\operatorname{tr}\left(\mathbf{R}_{p,i}\right) + \left\|\widehat{\mathbf{h}}_{i|i-1}\right\|^2 + q(\mathbf{P}_i)\right), \tag{2.18}$$

and the B-bit index of $\mathbf{P}_{i,\text{best}}$ can be sent as feedback from the user to the base station.

Note that maximizing (2.18) is the same as minimizing the MSE in (2.15) augmented with the term $q(\mathbf{P}_i)$ ($\widehat{\mathbf{h}}_{i|i-1}$ is a constant regardless of $\mathbf{P}_i$). Numerical studies in Section 2.4 show that $q(\mathbf{P}_i)$ has a non-negligible impact on the received SNR when N_t is moderately large, the channel is highly correlated in space, and the SNR is low. For other cases, however, the difference between the two metrics is negligible.

2.3.3 Closed-loop training with memory with full feedback to minimize MSE

In this subsection, we derive the optimal training signal $\mathbf{X}_{i,\text{opt}}$ of closed-loop training with memory that minimizes the MSE of the i-th fading block in (2.14). Note that $\mathbf{X}_{i,\text{opt}}$ is possible only when closed-loop training supports unlimited feedback overhead. Thus, $\mathbf{X}_{i,\text{opt}}$ only gives an MSE lower bound of the proposed closed-loop with memory.

Because the MSE in (2.14) has the same formulation as (2.6) once $\mathbf{R}_{i|i-1}$ is replaced by $\mathbf{R}$, the same arguments employed in Lemma 2.2.1 can be used to show that the optimal training signal is given as

$$\mathbf{X}_{i,\text{opt}} = \sqrt{\rho}\mathbf{U}_{i[1:T]} \tag{2.19}$$

where $\mathbf{R}_{i|i-1} = \mathbf{U}_i\mathbf{\Lambda}_i\mathbf{U}_i^H$ with $\mathbf{\Lambda}_i = \text{diag}\left([\lambda_{i,1}, \cdots, \lambda_{i,N_t}]\right)$. Comparing (2.7) and (2.19), the optimal training signal is now the first T dominant eigen-directions of the prediction matrix $\mathbf{R}_{i|i-1}$.

It is interesting to point out that, using the recursive derivation of $\mathbf{R}_{i|i-1}$ and $\mathbf{R}_{i|i}$, we can easily show that $\mathbf{U}_i$ is column-wise permutation of $\mathbf{U}$ (the eigenvector matrix of $\mathbf{R}$), which means the T dominant eigenvectors varies with i. Thus, the full-feedback scheme can be thought of as a training technique that scans among the eigen-directions of the original spatial correlation matrix $\mathbf{R}$. This property has been exploited in [41] for FDD massive MIMO training when the base station has perfect knowledge of $\mathbf{R}$.

We now derive the MSE of the i-th fading block using $\mathbf{X}_{i,\text{opt}}$ to provide a lower bound on the MSE of closed-loop training with memory in the following lemma.

Lemma 2.3.1 *Recall* $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$ *and let* $\mathbf{U}_0 = \mathbf{U}$ *and* $\mathbf{\Lambda}_0 = \mathbf{\Lambda}$. *Using the Kalman filter update in Table 2.1 and the optimal training signal* $\mathbf{X}_{i,\text{opt}} = \sqrt{\rho}\mathbf{U}_{i[1:T]}$, *the MSE at the* i*-th fading block is given as*

$$\text{MSE}_i(\mathbf{X}_{i,\text{opt}}) = 1 - \frac{1}{N_t}\sum_{k=0}^{i}\sum_{t=1}^{T}\frac{\eta^{2(i-k)}\rho\lambda_{k,t}^2}{\rho\lambda_{k,t}+1}, \tag{2.20}$$

where $\lambda_{k,t}$ *is the* t*-th dominant eigenvalue of* $\mathbf{R}_{k|k-1}$.

Proof See Appendix A.4. ■

When $i = 0$, (2.20) simplifies down to (2.8). Lemma 2.3.1 clearly shows that in temporally correlated channels with $\eta \approx 1$, the MSE_i in (2.20) is always lower than the MSE of closed-loop/single-shot training in (2.8) for $i > 0$. Thus, channel prediction with an optimized training signal selection will improve channel estimation performance. Although it is hard to analyze the normalized received SNR with $\mathbf{X}_{i,\text{opt}}$, we can expect from Lemma 2.3.1 that closed-loop training with memory can effectively reduce the ceiling effect of single-shot training discussed in Section 2.2.2.

2.3.4 Design of training signal set $\mathcal{P}$

Now, we discuss an effective way of generating a set of training signals $\mathcal{P}$. We again restrict $\mathcal{P}$ to be a subset of the set $\mathcal{X}$ meaning

$$\mathcal{P} \subset \mathcal{X} = \left\{\mathbf{F} : \mathbf{F} \in \mathbb{C}^{N_t \times T},\ \mathbf{F}^H\mathbf{F} = \rho\mathbf{I}_T\right\}.$$

It is shown in the previous subsection that the optimal training method that minimizes the MSE scans over the eigen-directions of $\mathbf{R}$ that are orthogonal to each other. To mimic this, the training signals in $\mathcal{P}$ should be as orthogonal as possible. This can be numerically achieved by Grassmannian subspace packing (GSP).

The chordal distance between the two matrices $\mathbf{X}$ and $\mathbf{Y}$ is given as

$$d_c\left(\mathbf{X},\mathbf{Y}\right) \triangleq \frac{1}{\sqrt{2}}\left\|\mathbf{X}\mathbf{X}^H-\mathbf{Y}\mathbf{Y}^H\right\|_F,$$

and the minimum chordal distance of a candidate training set $\mathcal{P}_c=\left\{\mathbf{P}_{c,1},\ldots,\mathbf{P}_{c,2^B}\right\}$ as

$$d_{c,\min}\left(\mathcal{P}_c\right) \triangleq \min_{1\le m\le n\le 2^B} d_c\left(\mathbf{P}_{c,m},\mathbf{P}_{c,n}\right).$$

Then, the GSP training set $\mathcal{P}_{\mathrm{GSP}}$ can be given as

$$\mathcal{P}_{\mathrm{GSP}} = \underset{\mathcal{P}_c\subset\mathcal{P}_{\mathrm{all}}}{\operatorname{argmax}}\, d_{c,\min}\left(\mathcal{P}_c\right),$$

where $\mathcal{P}_{\mathrm{all}}$ is a set of all possible candidate sets $\mathcal{P}_c$. We adopt numerically optimized $\mathcal{P}_{\mathrm{GSP}}$ for performance evaluation in Section 2.4.

2.3.5 Impact of system parameters on closed-loop training

In this subsection, we give explanations of scenarios when closed-loop training with memory has a gain compared to open-loop training with memory. The explanations are based on the optimal training signal $\mathbf{X}_{i,\mathrm{opt}}$ that minimizes the MSE for tractable analyses.

1) Variation with SNR (ρ): The minimization of the MSE in (2.14) can be first converted to the maximization of

$$\mathrm{tr}\left(\mathbf{R}_{i|i-1}\mathbf{X}_i\left(\mathbf{I}_T+\mathbf{X}_i^H\mathbf{R}_{i|i-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^H\mathbf{R}_{i|i-1}\right)$$

and approximated as

$$\mathrm{tr}\left(\mathbf{X}_i^H\mathbf{R}_{i|i-1}^2\mathbf{X}_i\right)$$

in the low-SNR regime and

$$\mathrm{tr}\left(\left(\mathbf{X}_i^H\mathbf{R}_{i|i-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^H\mathbf{R}_{i|i-1}^2\mathbf{X}_i\right)$$

in the high-SNR regime. The optimal training signal in both cases is again $\mathbf{X}_{i,\text{opt}} = \sqrt{\rho}\mathbf{U}_{i[1:T]}$. However, if we plug in $\mathbf{X}_{i,\text{opt}}$ into each approximated objective function, we have $\sum_{t=1}^{T} \rho\lambda_{i,t}^2$ in the low-SNR regime and $\sum_{t=1}^{T} \rho\lambda_{i,t}$ in the high-SNR regime. Assuming $\lambda_{i,t} > 1$ for $t = 1, \ldots, T$, which is typically true for spatially correlated massive MIMO channels, the subspace spanned by the columns of the training signal is more important in the low-SNR regime than the high-SNR regime. Thus, it is expected that closed-loop training with memory would be more beneficial in the low-SNR regime.

2) Variation with length of training phase (T): When $T = 1$, the direction of the optimal training signal is the dominant eigenvector of $\mathbf{R}_{i|i-1}$ so that $\mathbf{x}_{i,\text{opt}} = \mathbf{U}_{i[1:1]}$. However, when $T = N_t$, it is easy to show in a similar manner as (2.8) that any scaled unitary matrix $\mathbf{X}_{i,\text{opt}} = \sqrt{\rho}\mathbf{V}$ is optimal giving

$$\text{tr}\left(\left(\mathbf{I}_{N_t} + \mathbf{X}_{i,\text{opt}}^H \mathbf{R}_{i|i-1} \mathbf{X}_{i,\text{opt}}\right)^{-1} \mathbf{X}_{i,\text{opt}}^H \mathbf{R}_{i|i-1}^2 \mathbf{X}_{i,\text{opt}}\right) = \sum_{t=1}^{N_t} \frac{\rho\lambda_{i,t}^2}{\rho\lambda_{i,t} + 1}.$$

This means that there is no preferable direction for $\mathbf{X}_{i,\text{opt}}$ when $T = N_t$. In the case of $1 < T < N_t$, it is obvious that any other combination of T columns of $\mathbf{U}_i$ (except rearranging the first T columns of $\mathbf{U}_i$) for $\mathbf{X}_i$ gives inferior results than $\mathbf{U}_{i[1:T]}$. However, the gap between $\mathbf{U}_{i[1:T]}$ and other combinations will reduce as T increases. Thus, the subspace spanned by the columns of the training signal is more important when T (or the ratio $\frac{T}{N_t}$) is small, and closed-loop training with memory is most beneficial in this scenario.

3) Variation with fading block index i: Intuitively, the subspace spanned by the columns of the training signal seems to be more important at the beginning of channel estimation when the user lacks accurate channel knowledge. To explain this rigorously, we know from (2.20) in Lemma 2.3.1 that the MSE is decreased by $\sum_{t=1}^{T} \frac{\rho\lambda_{i,t}^2}{\rho\lambda_{i,t}+1}$ at the i-th block when $\mathbf{X}_{i,\text{opt}}$ is used as a training signal. We also know from the proof of Lemma 2.3.1 that the first T eigenvalues of $\mathbf{R}_{i|i-1}$ are decreasing with i such that

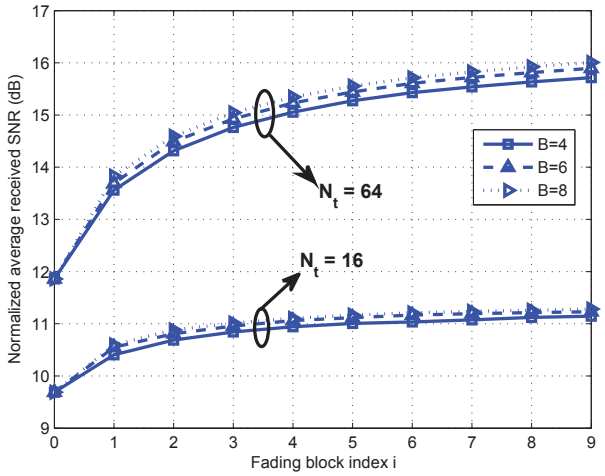


Fig. 2.3.: $\Gamma_i^{(\mathrm{dB})}$ of SNR-based closed-loop training according to the fading block index i with $\rho = 0$dB, $T = 2$, $a = 0.9$, and different B and N_t values.

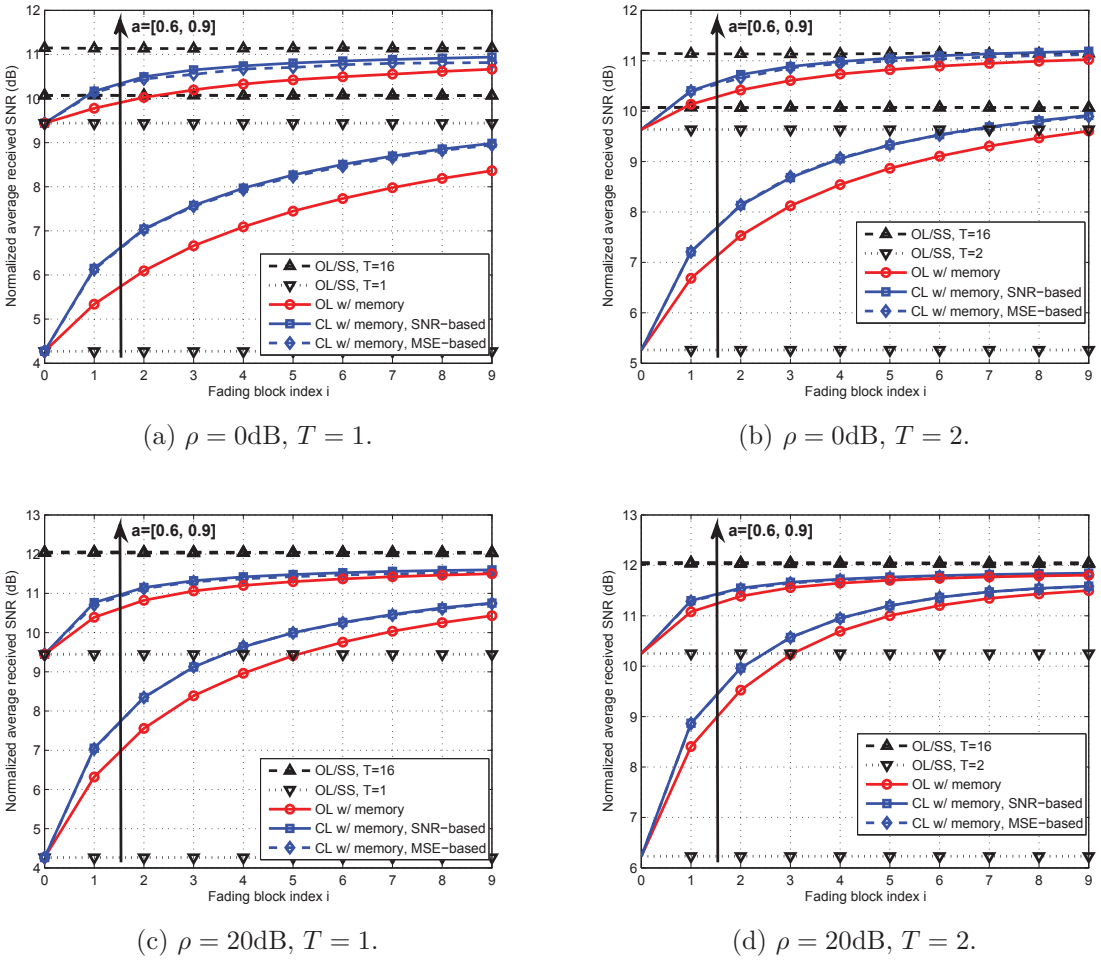
$\lambda_{i-1,t} \geq \lambda_{i,t}$ for $t = 1, \ldots, T$. Because of the Schur-convexity of $\sum_{t=1}^{T} \frac{\rho\lambda_{i,t}^2}{\rho\lambda_{i,t}+1}$ as shown in Appendix B, we have

$$\sum_{t=1}^{T} \frac{\rho\lambda_{i-1,t}^2}{\rho\lambda_{i-1,t}+1} \geq \sum_{t=1}^{T} \frac{\rho\lambda_{i,t}^2}{\rho\lambda_{i,t}+1}.$$

Thus, having the right subspace spanned by the columns of the training signal can reduce the MSE more effectively when i is small, and closed-loop training with memory has more gain when a prior channel estimate is not accurate at the beginning of channel estimation.

2.4 Simulation Results and Discussions

To evaluate the proposed training frameworks, we present Monte-Carlo simulation results with 10000 iterations in this section. Each iteration consists of 10 fading blocks which are temporally and spatially correlated as shown in (2.5). We adopt Jakes' model [84] for the temporal correlation coefficient $\eta = J_0(2\pi f_D \tau)$ where $J_0(\cdot)$ is the 0-th order Bessel function of the first kind, $\tau = 5$ms is the channel instantiation interval, and $f_D = \frac{v f_c}{c}$ denotes the maximum Doppler frequency. With the user speed



(a) $\rho = 0$dB, $T = 1$.

(b) $\rho = 0$dB, $T = 2$.

(c) $\rho = 20$dB, $T = 1$.

(d) $\rho = 20$dB, $T = 2$.

Fig. 2.4.: $\Gamma_i^{(\mathrm{dB})}$ according to the fading block index i with $N_t = 16$ and different ρ, T, and a values.

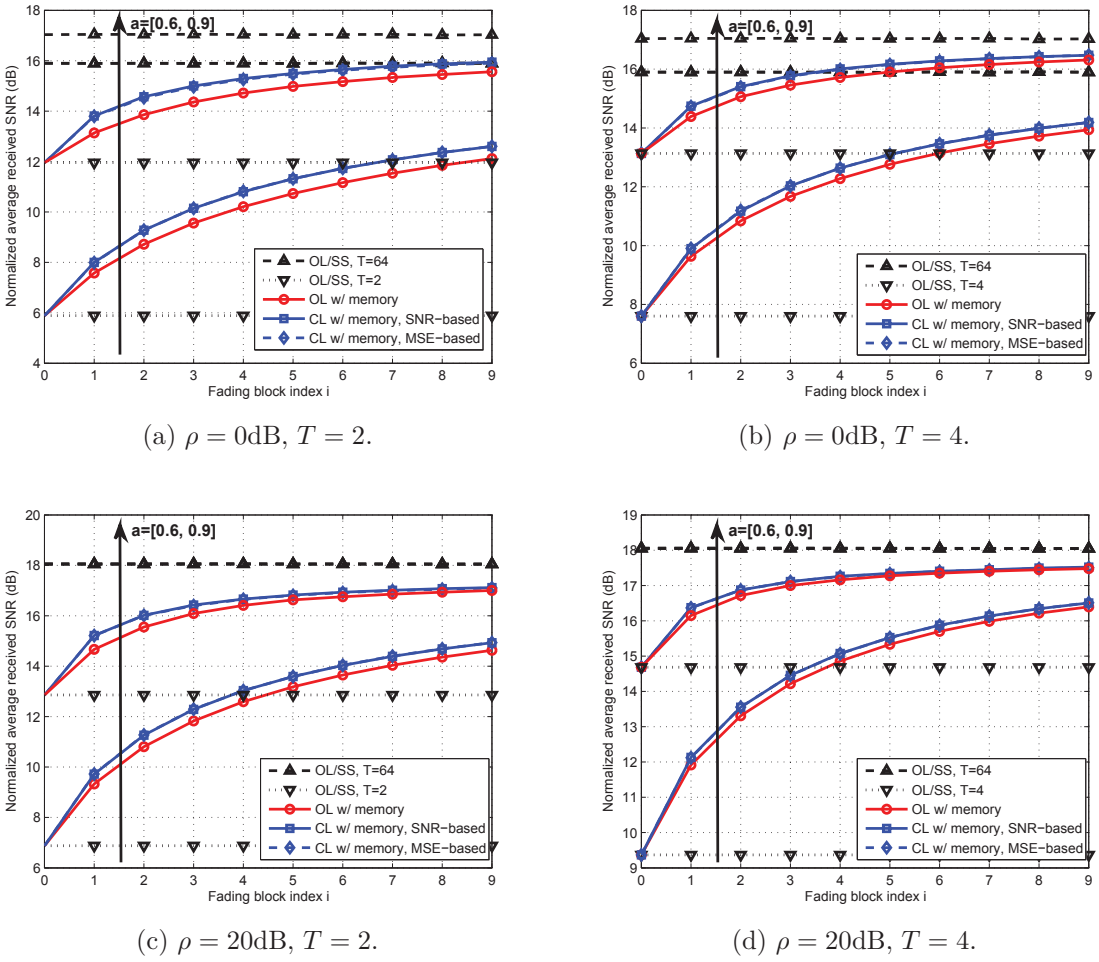
$v = 3$km/h, the carrier frequency $f_c = 2.5$GHz, and the speed of light $c = 3 \times 10^8$m/s, the temporal correlation coefficient becomes $\eta = 0.9881$. Assuming a 5ms coherence time and the frame structure of 3GPP LTE FDD systems [79], each fading block consists of $L \approx 10$ static channel uses. We adopt the same spatial correlation matrix $\mathbf{R}$ as in (2.10), and the numerically optimized GSP training set $\mathcal{P}_{\mathrm{GSP}}$ that is used in both open-loop and closed-loop training with memory. The dB scale of the normalized average received SNR in (2.4), $\Gamma_i^{(\mathrm{dB})}$, is used for the performance metric.

We first compare $\Gamma_i^{(\mathrm{dB})}$ of closed-loop training with memory based on the SNR metric with different values of B for $\mathcal{P}_{\mathrm{GSP}}$ in Fig. 2.3. We set the signal power $\rho = 0$dB, the number of channel uses for training $T = 2$, and the spatial correlation parameter $a = 0.9$. As the size of $\mathcal{P}_{\mathrm{GSP}}$ increases, $\Gamma_i^{(\mathrm{dB})}$ also increases in both $N_t = 16$ and 64 cases. The gain of having larger B is more prominent when N_t is large; however, it is expected that having B less than 10 bits seems to be enough to have a notable gain. This means that the computational complexity of training signal selection might not be a big issue in practice. We set $B = 6$ for other simulations in this section.

In Figs. 2.4 and 2.5, we plot $\Gamma_i^{(\mathrm{dB})}$ of open-loop/single-shot (OL/SS), open-loop with memory (OL w/ memory), and closed-loop with memory based on the MSE metric (CL w/ memory, MSE-based) and the SNR metric (CL w/ memory, SNR-based) training schemes according to the fading block index i with $N_t = 16$ and 64 and different values of ρ, T, and a. We randomly reorder the indices of $\mathcal{P}_{\mathrm{GSP}}$ at each iteration to preclude the effect of a specific ordering of training signals in open-loop training.

From the figures, it is easy to verify that with the same T the proposed training frameworks outperform open-loop/single-shot training, which adopts the training signal for the i-th block with a round-robin manner in (2.12) with $\mathcal{P}_{\mathrm{GSP}}$ but only relies on $\mathbf{y}_{i,train}$ for the i-th block channel estimation. Moreover, in Fig. 2.4b, closed-loop training with memory is slightly better than open-loop/single-shot training with $T = N_t$ when $a = 0.9$ and $i = 9$. This shows that the successive channel estimation approach of closed-loop training with memory can effectively alleviate the impact of noise.

Comparing open-loop and closed-loop training with memory, the gain of closed-loop training with memory becomes larger when 1) ρ is low, 2) T is small relative to N_t, and 3) i is small, which are inline with the discussions in Section 2.3.5. It is shown in Figs. 2.4 that closed-loop training with memory based on the SNR metric



(a) $\rho = 0$dB, $T = 2$.

(b) $\rho = 0$dB, $T = 4$.

(c) $\rho = 20$dB, $T = 2$.

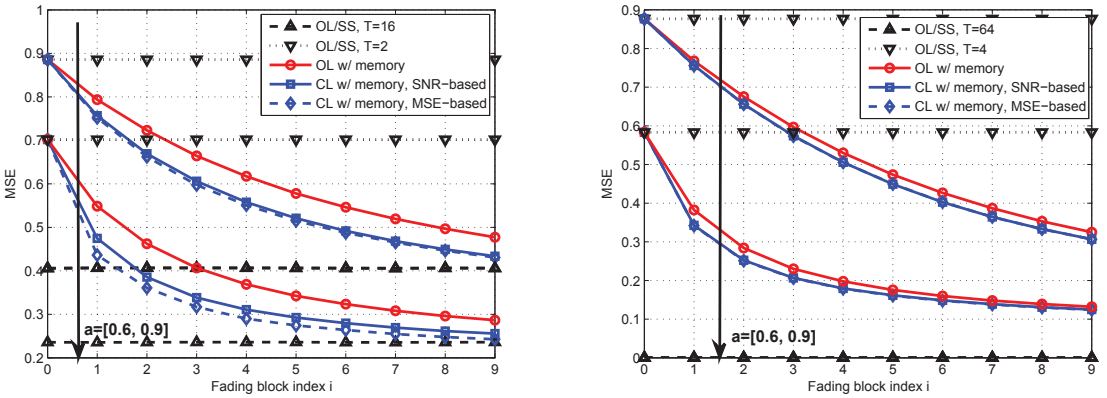
(d) $\rho = 20$dB, $T = 4$.

Fig. 2.5.: $\Gamma_i^{(\mathrm{dB})}$ according to the fading block index i with $N_t = 64$ and different ρ, T, and a values.

gives non-negligible gain compared to closed-loop training based on the MSE metric when N_t is moderately large and ρ is small in highly correlated case.

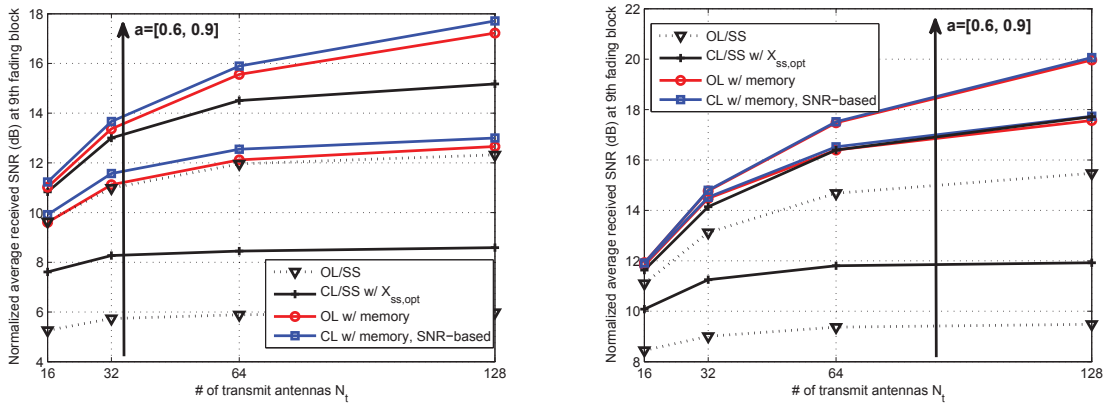
The performance of all schemes increases as a increases, i.e., when channels are highly correlated in space. This certainly shows that the spatial correlation *helps* in estimating the channel, which is pointed out in Lemma 2.2.2 and [35, 38].

We also plot the MSE of each scheme in Fig. 2.6. Similar to the previous figures of $\Gamma_i^{(\mathrm{dB})}$, the proposed training frameworks give far lower MSE than open-loop/single-shot training. Note that the MSE of closed-loop training with memory based on



(a) $N_t = 16$, $\rho = 0$dB, $T = 2$. (b) $N_t = 64$, $\rho = 20$dB, $T = 4$.

Fig. 2.6.: MSE according to the fading block index i with different N_t, ρ, T, and a values.



(a) $\rho = 0$dB, $T = 2$. (b) $\rho = 20$dB, $T = 4$.

Fig. 2.7.: $\Gamma_i^{(\mathrm{dB})}$ according to N_t with different ρ, T, and a values.

the MSE metric is smaller than that of closed-loop training based on the SNR metric when $N_t = 16$, which shows the tradeoff between SNR and MSE metric in closed-loop training.

In Fig. 2.7, we plot $\Gamma_i^{(\mathrm{dB})}$ of the 9th fading block according to N_t. Note that $\Gamma_i^{(\mathrm{dB})}$ of open-loop/single-shot training quickly saturates as N_t increases. We also plot the results of closed-loop/single-shot training with full feedback of $\mathbf{X}_{\mathrm{ss,opt}}$ (CL/SS w/$\mathbf{X}_{\mathrm{ss,opt}}$)

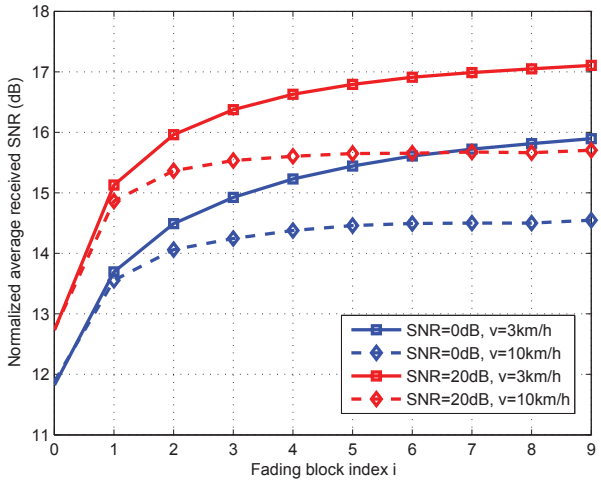


Fig. 2.8.: $\Gamma_i^{(\mathrm{dB})}$ of SNR-based closed-loop training according to the fading block index i with $T = 2$, $a = 0.9$, $B = 6$, $N_t = 64$ and different ρ and v values.

discussed in Section 2.2, which also experiences the ceiling effect, for comparison. It is obvious that open-loop and closed-loop training with memory can effectively reduce the ceiling effect even with small T compared to L, especially when a is large. This clearly shows that the gain of the proposed training schemes for massive MIMO systems.

Finally, we plot closed-loop training with memory based on the SNR metric with different user velocities in Fig. 2.8. Note that $v = 10$km/h corresponds to $\eta = 0.8721$. The loss from the high velocity is severe, i.e., almost 1.4dB loss of the received SNR in the saturation regime. When the user velocity is high, instead of relying on the closed-loop training framework, the base station should transmit sounding signals more frequently in an open-loop manner in practice. For example, four sounding signals (or reference signals) would be transmitted within a 1ms time period to support 350km/h user velocity in 3GPP LTE systems [79]. Even in this case, the proposed open-loop training with memory can be exploited.

3. NONCOHERENT TRELLIS CODED QUANTIZATION: A PRACTICAL LIMITED FEEDBACK TECHNIQUE FOR MASSIVE MIMO SYSTEMS

In this chapter, we explain NTCQ that can solve the CSI quantization problem in FDD massive MIMO. Our NTCQ approach relies on two key observations: (a) Quantization for beamforming requires finding a quantized vector, from among the available choices, that is best aligned with the true channel vector, in terms of maximizing the magnitude of their normalized inner product. This corresponds to a search on the Grassmann manifold rather than in Euclidean space. We point out, as have others before us, that this source coding problem maps to a channel coding problem of *noncoherent* sequence detection, where we try to find the most likely transmitted codeword subject to an unknown multiplicative complex-valued channel gain. (b) We know from prior work on noncoherent communication that a noncoherent block demodulator can be implemented near-optimally using a bank of coherent demodulators, each with a different hypothesis on the unknown channel gain. Furthermore, signal designs and codes for coherent communication are optimal for noncoherent communication, as long as we adjust our encoding and decoding slightly to account for the ambiguity caused by the unknown channel gain.

The relationship between quantization based on a mean squared error cost function and channel coding for *coherent* communication over the AWGN channel has been exploited successfully in the design of trellis coded quantization (TCQ) [85], in which the code symbols take values from a standard finite constellation used for communication, such as phase shift keying (PSK) or quadrature amplitude modulation

[0]

(QAM). The quantized code vector can then be found by using a Viterbi algorithm for trellis decoding. Our observation (b) allows us to immediately extend this strategy to the noncoherent setting. The code vectors for NTCQ can be exactly the same as in standard TCQ, but the encoder now consists of several Viterbi algorithms (in practice, a very small number) running in parallel, with a rule for choosing the best output. Thus, while approximating a beamforming vector on the Grassmann manifold as in (a) appears to be difficult, it can be easily solved by using several parallel searches in Euclidean space. Furthermore, just as noncoherent channel codes inherit the good performance of the coherent codes they were constructed from, NTCQ inherits the good quantization performance of TCQ.

We first show that channel codes, and by analogy, source codes developed in a coherent setting can be effectively leveraged in the noncoherent setting of interest in CSI generation for beamforming. As shown through both analysis and simulations, the resulting NTCQ strategy provides near-optimal beamforming gain, and has encoding complexity which is linear in the channel dimension. We also develop adaptive NTCQ techniques that are optimized for spatial and temporal correlations. A differential version of NTCQ utilizes the temporal correlation of the channel to successively refine the quantized channel to decrease the quantization error. A spatially adaptive version of NTCQ exploits the spatial correlation of the channel so that it only quantizes the local area of the dominant direction of the spatial correlation matrix. Utilization of channel statistics using such advanced schemes can significantly improve the performance or decrease the feedback overhead by utilizing channel statistics.

An important feature of NTCQ is its flexibility, which makes it an attractive candidate for potentially providing a common channel quantization approach for heterogeneous fifth generation (5G) wireless communication systems, which could involve a mix of advanced network entities such as massive MIMO, coordinated multipoint (CoMP) transmission, relay, distributed antenna systems (DAS), and femto/pico cells. For example, massive MIMO systems could be implemented using a two-dimensional (2D) planar antenna array at the base station to reduce the size of antenna ar-

ray [86]. Depending on the channel quality, the base station could turn on and off the rows/columns of this 2D array to achieve better performance. The same situation could be encountered in CoMP and DAS because the number of coordinating transmit stations may vary over time. NTCQ can easily adjust to such scenarios, since it can adapt to different numbers of transmit antennas (or more generally, space-time channel dimension) by changing the number of code symbols, and can adapt CSI accuracy and feedback overhead by changing the constellation size and the coded modulation scheme.

We have already mentioned conventional look-up based quantization approaches and discussed why they do not scale. Trellis-based quantizers for CSI generation have been proposed previously in [87–90], but the path metrics used for the trellis search are *ad hoc*. On the other hand, the mapping to noncoherent sequence detection, similar to NTCQ, has been pointed out in [91]. Depending on the number of constellation points used for the candidate codewords, the proposed algorithms in [91] are dubbed as PSK & QAM singular vector quantization (SVQ). Although PSK/QAM-SVQ adopt similar codeword search methods as NTCQ, they do not consider coding. The use of nontrivial trellis codes as proposed here significantly enhances performance compared to PSK/QAM-SVQ with the same amount of feedback overhead. Furthermore, [91] employs optimal noncoherent block demodulation, derived in [92,93], for quantization, incurring complexity $O(M_t^3)$ for QAM-SVQ and $O(M_t \log M_t)$ for PSK-SVQ, where M_t denotes the number of antennas. Our NTCQ scheme exhibits better complexity scaling: near-optimal demodulation in $O(M_t)$ complexity by running a small number of coherent decoders in parallel, as proposed in [94], suffices for providing near-optimal quantization performance.

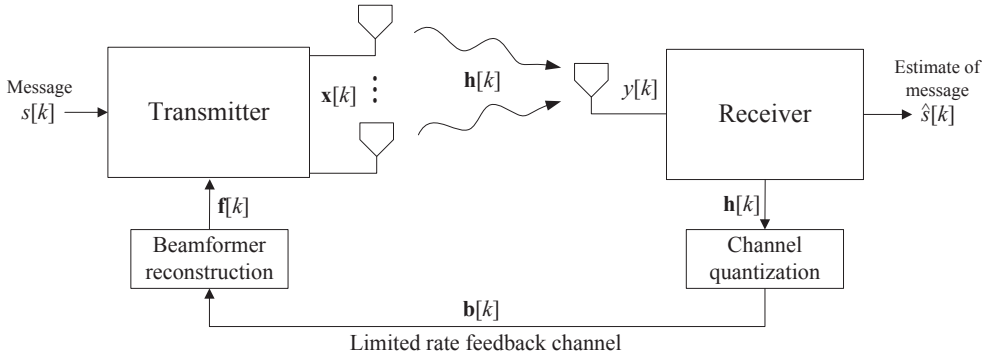


Fig. 3.1.: Multiple-input, single-output communications system with feedback.

3.1 System Model and Theory

3.1.1 System setup

We consider a block fading MISO communications system with M_t transmit antennas at the transmitter as in Fig. 3.1. The received signal, $y_\ell[k] \in \mathbb{C}$, for a channel use index ℓ in the kth fading block can be written as

$$y_\ell[k] = \mathbf{h}^H[k]\mathbf{f}[k]s_\ell[k] + z_\ell[k],$$

where $\mathbf{h}[k] \in \mathbb{C}^{M_t}$ is the MISO channel vector, $\mathbf{f}[k] \in \mathbb{C}^{M_t}$ is the beamforming vector with $\|\mathbf{f}[k]\|_2^2 = 1$, $s_\ell[k] \in \mathbb{C}$ is the message signal with $E\left[s_\ell\left[k\right]\right] = 0$ and $E\left[|s_\ell[k]|^2\right] = \rho$, and $z_\ell[k] \in \mathbb{C}$ is additive complex Gaussian noise such that $z_\ell[k] \sim \mathcal{CN}(0, \sigma^2)$. A number of different models for $\mathbf{h}[k]$ will be considered in the design and performance evaluation of quantization schemes, but for now, we allow it to be arbitrary. The receiver quantizes its estimate of $\mathbf{h}[k]$ into a B_{tot}-dimensional binary vector $\mathbf{b}[k]$, which is sent over a limited rate feedback channel. The transmitter uses this feedback to construct a beamforming vector $\mathbf{f}[k]$. In order to focus attention on channel quantization, we do not model channel estimation errors at the receiver or errors over the feedback channel.

Since we do not consider temporal correlation in $\{\mathbf{h}[k]\}$ for quantizer design in this section, we drop the time index k for the remainder of this section. Assuming an average power constraint at the transmitter, we wish to choose $\mathbf{f}$ so as to maximize the *normalized beamforming gain* that is defined as

$$J(\mathbf{f}, \mathbf{h}) = \frac{|\mathbf{h}^H \mathbf{f}|^2}{\|\mathbf{h}\|_2^2 \|\mathbf{f}\|_2^2}. \tag{3.1}$$

Although $\|\mathbf{f}\|_2 = 1$, we still normalize with $\|\mathbf{f}\|_2$ in (3.1) to maintain notational generality. An equivalent approach is to minimize the *chordal distance* between $\mathbf{f}$ and $\mathbf{h}$, defined as

$$d_c^2(\mathbf{f}, \mathbf{h}) = 1 - J(\mathbf{f}, \mathbf{h}) = 1 - \frac{|\mathbf{h}^H \mathbf{f}|^2}{\|\mathbf{h}\|_2^2 \|\mathbf{f}\|_2^2}.$$

These performance measures require searching for codewords on the Grassmann manifold, a projective space in which vectors are mapped to one-dimensional complex subspaces.

Conventional VQ codebook-based channel quantization typically employs exhaustive search to select a codeword from an unstructured and fixed B_{tot}-bit codebook $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{2^{B_{\text{tot}}}}\}$ according to

$$\mathbf{c}_{\text{opt}} = \underset{\mathbf{c} \in \mathcal{C}}{\operatorname{argmax}}\, J(\mathbf{c}, \mathbf{h}) = \underset{\mathbf{c} \in \mathcal{C}}{\operatorname{argmin}}\, d_c^2(\mathbf{c}, \mathbf{h}), \tag{3.2}$$

and the binary sequence $\mathbf{b} = \text{bin}(\text{opt})$ is fed back to the transmitter where $\text{bin}(\cdot)$ converts an integer to its binary representation. Then the beamforming vector is reconstructed at the transmitter as

$$\mathbf{f} = \frac{\mathbf{c}_{\text{int}(\mathbf{b})}}{\|\mathbf{c}_{\text{int}(\mathbf{b})}\|_2}$$

where $\text{int}(\cdot)$ converts a binary string into an integer. Exhaustive search, which does not require geometric interpretation of the performance metric, incurs computational complexity $O(M_t 2^{B_{\text{tot}}})$, which is exponential in the number of bits. We shall see that

utilizing the geometry of the Grassmann manifold, and in particular, relating it to Euclidean geometry, is key to more efficient quantization procedures.

Since our performance criterion is independent of the codeword norm, one could, without loss of generality, normalize the codewords to unit norm up front (i.e., set $\|\mathbf{c}\|_2 \equiv 1$). However, for the code constructions and quantizer designs of interest to us, it is useful to allow codewords to have different norms (the performance criterion, of course, remains independent of codeword scaling).

3.1.2 Feedback overhead

The relation between the feedback overhead B_{tot} (or codebook size $2^{B_{\text{tot}}}$) and the performance of MIMO systems has been thoroughly investigated for i.i.d. Rayleigh fading channels. In single user (SU) MISO channels with the B_{tot} bits RVQ codebook, the loss in normalized beamforming gain is given as [14]

$$\begin{aligned} E\left[1 - \max_{\mathbf{f}\in\mathcal{F}_{\text{RVQ}}} J(\mathbf{f},\mathbf{h})\right] &= 2^{B_{\text{tot}}}\beta\left(2^{B_{\text{tot}}}, \frac{M_t}{M_t-1}\right) \\ &\approx 2^{-\frac{B_{\text{tot}}}{M_t-1}} \end{aligned} \tag{3.3}$$

where $\mathcal{F}_{\text{RVQ}}$ is an RVQ codebook, $\beta(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ is the Beta function, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ is the Gamma function, and expectation is taken over $\mathbf{h}$ and $\mathcal{F}_{\text{RVQ}}$. The expression in (3.3) indicates that the feedback overhead needs to be increased proportional to M_t to maintain the loss in normalized beamforming gain at a certain level.

For MU-MIMO zero-forcing beamforming (ZFBF), a similar conclusion is drawn in [95,96]: in order to achieve the full multiplexing gain of M_t, the number of feedback bits per user, B_{user}, must scale linearly with SNR (in dB) and M_t as

$$B_{\text{user}} = (M_t - 1)\log_2\rho \approx \frac{M_t-1}{3}\rho_{dB}.$$

We therefore assume that at each channel use, the receiver sends back a binary feedback sequence of length

$$B_{\text{tot}} \triangleq BM_t + q$$

where B is the number of quantization bits used per transmit antenna and q is a small, fixed number of auxiliary feedback bits, which does not scale with M_t.

While linear scaling of feedback bits with the number of transmit elements is typically acceptable in terms of overhead, a VQ codebook-based limited feedback is computationally infeasible for massive MIMO systems with large M_t because of the exponential growth of codeword search complexity with M_t as $O(M_t 2^{BM_t})$. Thus, we need to develop new techniques to quantize CSI for large M_t.

In order to develop an efficient CSI quantization method for massive MIMO systems, we draw an analogy between searching for a candidate beamforming vector to maximize beamforming gain as in (3.2) and noncoherent sequence detection (e.g., [87, 91]). We then employ prior work relating noncoherent and coherent detection to map quantization on the Grassmann manifold to quantization in Euclidean space, which can be accomplished far more efficiently. This line of reasoning, which corresponds to the *process* of quantization, has been previously established in [91], but we provide a self-contained derivation in Section 3.1.3 pointing to a low-complexity, near-optimal source encoding strategy. We then show, in Section 3.1.4 that structured quantization codebooks for Euclidean metrics are effective for quantization on the Grassmann manifold. This leads to a CSI quantization framework which is efficient in terms of both overhead and computation.

3.1.3 Efficient Grassmannian encoding using Euclidean metrics

Consider a single antenna noncoherent, block fading, additive white Gaussian noise (AWGN) channel with received vector

$$\mathbf{y} = \beta\mathbf{x} + \mathbf{n},$$

where $\beta \in \mathbb{C}$ is an unknown complex channel gain, $\mathbf{x} \in \mathbb{C}^N$ is a vector of N transmitted symbols, $\mathbf{n} \in \mathbb{C}^N$ is complex Gaussian noise, and $\mathbf{y} \in \mathbb{C}^N$ is the received signal. Using the generalized likelihood ratio test (GLRT) as in [91, 94], the estimate of the transmitted vector, $\hat{\mathbf{x}}$, is given by

$$\begin{aligned}
\hat{\mathbf{x}} &= \underset{\mathbf{x}\in\mathbb{C}^N}{\operatorname{argmin}} \min_{\beta\in\mathbb{C}} \|\mathbf{y} - \beta\mathbf{x}\|_2^2 && (3.4)\\
&= \underset{\mathbf{x}\in\mathbb{C}^N}{\operatorname{argmin}} \min_{\alpha\in\mathbb{R}^+} \min_{\theta\in[0,2\pi)} \|\mathbf{y}\|_2^2 + \alpha^2\|e^{j\theta}\mathbf{x}\|_2^2 - 2\alpha\,\mathrm{Re}(e^{j\theta}\mathbf{y}^H\mathbf{x}) && (3.5)\\
&= \underset{\mathbf{x}\in\mathbb{C}^N}{\operatorname{argmin}} \min_{\alpha\in\mathbb{R}^+} \|\mathbf{y}\|_2^2 + \alpha^2\|\mathbf{x}\|_2^2 - 2\alpha|\mathbf{y}^H\mathbf{x}| && (3.6)\\
&= \underset{\mathbf{x}\in\mathbb{C}^N}{\operatorname{argmax}} \frac{|\mathbf{y}^H\mathbf{x}|^2}{\|\mathbf{x}\|_2^2}, && (3.7)
\end{aligned}$$

where we decomposed the entire complex plain $\beta = \alpha e^{j\theta}$ with $\alpha \in \mathbb{R}^+$ and $\theta \in [0, 2\pi)$ in (3.5), and (3.6) comes from

$$\min_{\theta\in[0,2\pi)} \left\{-\,\mathrm{Re}(e^{j\theta}\mathbf{y}^H\mathbf{x})\right\} = -|\mathbf{y}^H\mathbf{x}|.$$

To derive (3.7), we differentiate (3.6) with respect to α and set to 0 which gives $\alpha^\star = \frac{|\mathbf{y}^H\mathbf{x}|}{\|\mathbf{x}\|_2^2}$. Note that $\alpha^\star$ is the global minimizer of (3.6) because (3.6) is a quadratic function of α. We can derive (3.7) after plugging $\alpha^\star$ into (3.6) and some basic algebra.

We can easily check from (3.2) and (3.7) that finding the optimal codeword for a MISO beamforming system and the noncoherent sequence detection problems are equivalent (although this relation is already shown in [91], we proved the duality of (3.4) and (3.2) more explicitly than [91]). Therefore, we can find $\mathbf{c}_{\text{opt}}$ for a MISO

beamforming system with a Euclidean distance quantizer (or noncoherent block demodulator)

$$\min_{\alpha\in\mathbb{R}^+}\min_{\theta\in[0,2\pi)}\min_{\mathbf{c}_i\in\mathcal{C}}\|\bar{\mathbf{h}}-\alpha e^{j\theta}\mathbf{c}_i\|_2^2. \tag{3.8}$$

where $\bar{\mathbf{h}}=\frac{\mathbf{h}}{\|\mathbf{h}\|_2}$ is the normalized channel direction.

Moreover, instead of searching over the entire complex plane by having $\alpha\in\mathbb{R}^+$ and $\theta\in[0,2\pi)$, we know from prior work on noncoherent communication [94] that the noncoherent block demodulator in (3.8) can be implemented near-optimally using a bank of coherent demodulators over the optimized discrete sets of $\alpha\in\mathbb{A}=\{\alpha_1,\alpha_2,\ldots,\alpha_{K_\alpha}\}$ and $\theta\in\Theta=\{\theta_1,\theta_2,\ldots,\theta_{K_\theta}\}$. While *optimal* noncoherent detection can be accomplished with quadratic complexity in M_t [91], as we show through our numerical results, a small number of parallel coherent demodulators (which incurs complexity linear in M_t) is all that is required for excellent quantization performance.

The preceding development tells us that we can apply coherent demodulation, which maps to quantization using Euclidean metrics, to noncoherent demodulation, which maps to quantization on the Grassmann manifold. However, we must still determine how to choose the quantization codebook. Next, we present results indicating that we can simply use codes optimized for Euclidean metrics for this purpose.

3.1.4 Efficient Grassmannian codebooks based on Euclidean metrics

We begin with an asymptotic result for i.i.d. Rayleigh fading coefficients, which relies on the well-known rate-distortion theory for i.i.d. Gaussian sources.

Theorem 3.1.1 *If we quantize an $M_t\times 1$ i.i.d. Rayleigh fading MISO channel $\mathbf{h}\sim\mathcal{CN}(\mathbf{0},\sigma_h^2\mathbf{I})$ with a Euclidean distance quantizer using B bits per entry (which corresponds to $\frac{B}{2}$ bits per each of real and imaginary dimension) as*

$$\mathbf{g}_{\mathrm{ED}}=\min_{\mathbf{g}_i\in\mathcal{G}}\|\mathbf{h}-\mathbf{g}_i\|_2^2 \tag{3.9}$$

where $\mathcal{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_{2^{B_{\text{tot}}}}\}$, $B_{\text{tot}} = BM_t$, $\mathbf{g}_i \sim \mathcal{CN}(\mathbf{0}, (\sigma_h^2 - 2D)\mathbf{I})$ *for all* i, *and* $D = \frac{1}{2}\sigma_h^2 2^{-B}$, *then the asymptotic loss in normalized beamforming gain, or chordal distance, is given by*

$$d_c^2(\mathbf{h}, \mathbf{g}_{\text{ED}}) \overset{M_t \to \infty}{\longrightarrow} 2^{-B}. \tag{3.10}$$

Proof By expanding $\|\mathbf{h} - \mathbf{g}_{\text{ED}}\|_2^2$, we have

$$\|\mathbf{h} - \mathbf{g}_{\text{ED}}\|_2^2 = \sum_{t=1}^{M_t} \left[\{\text{Re}(h_t) - \text{Re}(g_{\text{ED},t})\}^2 + \{\text{Im}(h_t) - \text{Im}(g_{\text{ED},t})\}^2\right]$$

where h_t and $g_{\text{ED},t}$ are the t^{th} entry of $\mathbf{h}$ and $\mathbf{g}_{\text{ED}}$, respectively. Note that $\text{Re}(h_t)$ and $\text{Im}(h_t)$ are from the same distribution $\mathcal{N}(0, \frac{1}{2}\sigma_h^2)$, and $\text{Re}(g_{\text{ED},t})$ and $\text{Im}(g_{\text{ED},t})$ are from the distribution $\mathcal{N}(0, \frac{1}{2}\sigma_h^2 - D)$. Assuming $\frac{B}{2}$ bits are used to quantize each of $\text{Re}(h_t)$ and $\text{Im}(h_t)$ for all t, by rate-distortion theory for i.i.d. Gaussian sources [97], we can achieve the rate-distortion bound

$$\begin{aligned} E\left[\{\text{Re}(h_t) - \text{Re}(g_{\text{ED},t})\}^2\right] &= E\left[\{\text{Im}(h_t) - \text{Im}(g_{\text{ED},t})\}^2\right] \\ &= D \end{aligned}$$

as $M_t \to \infty$. Thus, by the weak law of large numbers, the following convergences hold[1]

$$\begin{aligned} \frac{1}{M_t}\|\mathbf{h} - \mathbf{g}_{\text{ED}}\|_2^2 &\xrightarrow{P} 2E\left[\{\text{Re}(h_t) - \text{Re}(g_{\text{ED},t})\}^2\right] = 2D, \\ \frac{1}{M_t}\|\mathbf{h}\|_2^2 &\xrightarrow{P} 2E[\{\text{Re}(h_t)\}^2] = \sigma_h^2, \\ \frac{1}{M_t}\|\mathbf{g}_{\text{ED}}\|_2^2 &\xrightarrow{P} 2E[\{\text{Re}(g_{\text{ED},t})\}^2] = \sigma_h^2 - 2D \end{aligned}$$

[1]Let $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ and $\mu = E[X_i]$ for all i. We say $\bar{X}_n$ converges to μ in probability as $\bar{X}_n \xrightarrow{P} \mu$ for $n \to \infty$ when $\lim_{n\to\infty} Pr\left(|\bar{X}_n - \mu| > \epsilon\right) = 0$ for any $\epsilon > 0$.

as $M_t \to \infty$. Moreover, $\left|\frac{\mathbf{h}^H \mathbf{g}_{\mathrm{ED}}}{M_t}\right|^2$ can be lower bounded as

$$\begin{aligned}\left|\frac{\mathbf{h}^H \mathbf{g}_{\mathrm{ED}}}{M_t}\right|^2 &\geq \left(\frac{\mathrm{Re}(\mathbf{h}^H \mathbf{g}_{\mathrm{ED}})}{M_t}\right)^2 \\ &= \left(\frac{\|\mathbf{h}\|_2^2 + \|\mathbf{g}_{\mathrm{ED}}\|_2^2 - \|\mathbf{h} - \mathbf{g}_{\mathrm{ED}}\|_2^2}{2M_t}\right)^2 \\ &\xrightarrow{P} \left(\sigma_h^2 - 2D\right)^2.\end{aligned}$$

Then, the normalized beamforming gain loss relative to the unquantized beamforming case is bounded as

$$\begin{aligned}d_c^2(\mathbf{h}, \mathbf{g}_{\mathrm{ED}}) &= 1 - \frac{|\mathbf{h}^H \mathbf{g}_{\mathrm{ED}}|^2}{\|\mathbf{h}\|_2^2 \|\mathbf{g}_{\mathrm{ED}}\|_2^2} \leq \frac{2D}{\sigma_h^2} = 2^{-B}, \\ d_c^2(\mathbf{h}, \mathbf{g}_{\mathrm{ED}}) &\overset{(a)}{\geq} 2^{-\frac{BM_t}{M_t - 1}}\end{aligned}$$

where (a) follows from the optimality of the RVQ codebook in large asymptotic regime [46]. As $M_t \to \infty$, the lower bound of $d_c^2(\mathbf{h}, \mathbf{g}_{\mathrm{ED}})$ converges to the upper bound 2^{-B}, which finishes the proof. ■

Note that the loss in (3.10) is asymptotically the same as that of the RVQ codebook in (3.3). Since the RVQ codebook is known to be asymptotically optimal as $M_t \to \infty$ (fixing the number of bits per antenna) [46], we conclude that coherent Euclidean distance quantization as in (3.9) with a rich, rotationally invariant constellation such as a Gaussian codebook $\mathcal{G}$, is also an asymptotically optimal way to quantize the channel vector $\mathbf{h}$. Of course, in practice, for finite constellations and number of antennas, we must "align" the codewords $\mathbf{g}_i$ with the channel $\mathbf{h}$, using parallel branches with different amplitude scaling α and phase rotations θ as in (3.8), prior to computing the Euclidean metric, in order to maximize the beamforming gain.

We also note that the use of nontrivial codes is implicit in Theorem 3.1.1, hence the uncoded constellations employed in [91] do not achieve optimal quantization per-

formance. The constellation expansion employed in the NTCQ schemes considered here is required to approach optimal performance.

We now provide a *non-asymptotic* result regarding the chordal distances associated with Grassmannian line packing (GLP) attained by codebooks optimized using Euclidean metrics. Let $N = 2^{B_{\text{tot}}}$ and $\mathcal{U}_{M_t}^N \in \mathbb{C}^{M_t \times N}$ denote the set of $M_t \times N$ complex matrices with unit vector columns. To minimize the average quantization error of (3.8) or (3.9) in Euclidean space with a fixed codebook $\mathcal{C}$, we have to maximize the minimum Euclidean distance between all possible codeword pairs

$$d_{E,\min}^2(\mathcal{C}) \triangleq \min_{1 \le k < l \le N} d_E^2(\mathbf{c}_k, \mathbf{c}_l)$$

where $d_E(\mathbf{x}, \mathbf{y}) \triangleq \|\mathbf{x} - \mathbf{y}\|_2$, and $\{\mathbf{c}_i\}_{i=1}^N$ are column vectors of $\mathcal{C}$. Let $\mathcal{C}_{\text{ED}}$ denote an optimized Euclidean distance (ED) codebook that maximizes the minimum Euclidean distance as

$$\mathcal{C}_{\text{ED}} = \underset{\mathcal{C} \in \mathcal{U}_{M_t}^N}{\operatorname{argmax}} \, d_{E,\min}^2(\mathcal{C}).$$

On the other hand, beamforming codebooks are ideally designed for i.i.d. Rayleigh fading channels to maximize the minimum chordal distance between codewords as

$$d_{c,\min}^2(\mathcal{C}) \triangleq \min_{1 \le k < l \le N} d_c^2(\mathbf{c}_k, \mathbf{c}_l),$$

and a GLP codebook is given as [43, 44]

$$\mathcal{C}_{\text{GLP}} = \underset{\mathcal{C} \in \mathcal{U}_{M_t}^N}{\operatorname{argmax}} \, d_{c,\min}^2(\mathcal{C}).$$

Note that the optimization metrics of $\mathcal{C}_{\text{GLP}}$ and $\mathcal{C}_{\text{ED}}$ are different, the former is the chordal distance and the latter is the Euclidean distance. The following lemma shows the relation of the two metrics.

Lemma 3.1.1 *For any two unit vectors* $\mathbf{x}$ *and* $\mathbf{y}$*, the squared chordal distance between* $\mathbf{x}$ *and* $\mathbf{y}$ *is upper bounded by a function of their Euclidean distance as*

$$\begin{aligned} d_c^2(\mathbf{x},\mathbf{y}) &\leq 1 - \left(1 - \frac{1}{2}d_E^2(\mathbf{x},\mathbf{y})\right)^2 \\ &= d_E^2(\mathbf{x},\mathbf{y}) - \frac{1}{4}d_E^4(\mathbf{x},\mathbf{y}). \end{aligned}$$

Proof Let us define $d_\theta^2(\mathbf{x},\mathbf{y})$ as

$$\begin{aligned} d_\theta^2(\mathbf{x},\mathbf{y}) &\triangleq \min_{\theta\in[0,2\pi)} d_E^2(\mathbf{x}, e^{j\theta}\mathbf{y}) \\ &= \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2\max_{\theta\in[0,2\pi)} \mathrm{Re}\left\{e^{j\theta}\mathbf{x}^H\mathbf{y}\right\} \\ &= 2 - 2|\mathbf{x}^H\mathbf{y}| \leq d_E^2(\mathbf{x},\mathbf{y}). \end{aligned}$$

Then, the squared chordal distance of $\mathbf{x}$ and $\mathbf{y}$ is upper bounded as

$$\begin{aligned} d_c^2(\mathbf{x},\mathbf{y}) &= 1 - |\mathbf{x}^H\mathbf{y}|^2 \\ &= 1 - \left(1 - \frac{1}{2}d_\theta^2(\mathbf{x},\mathbf{y})\right)^2 \\ &\leq 1 - \left(1 - \frac{1}{2}d_E^2(\mathbf{x},\mathbf{y})\right)^2, \end{aligned}$$

which finishes the proof. ■

Moreover, Lemma 3.1.1 can be directly extended to the following corollary.

Corollary 3.1.1 *The minimum chordal distance of* $\mathcal{C}_{\mathrm{ED}}$*,* $d_{c,\min}^2(\mathcal{C}_{\mathrm{ED}})$*, is upper bounded by the minimum Euclidean distance of* $\mathcal{C}_{\mathrm{ED}}$*,* $d_{E,\min}^2(\mathcal{C}_{\mathrm{ED}})$ *as*

$$d_{c,\min}^2(\mathcal{C}_{\mathrm{ED}}) \leq d_{E,\min}^2(\mathcal{C}_{\mathrm{ED}}).$$

Although Corollary 3.1.1 does not say that $\mathcal{C}_{\mathrm{ED}}$ maximizes the minimum chordal distance between its codewords, $\mathcal{C}_{\mathrm{ED}}$ is expected to have a *good* chordal distance property. We verify this by simulation with numerically optimized $\mathcal{C}_{\mathrm{GLP}}$ and $\mathcal{C}_{\mathrm{ED}}$ in

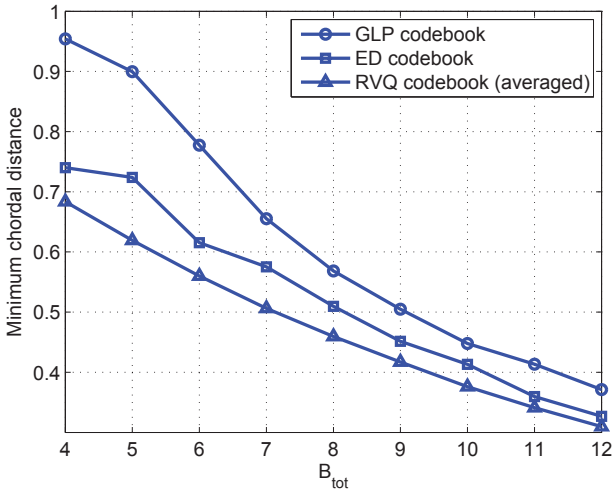


Fig. 3.2.: The minimum chordal distances of different codebooks with $M_t = 8$. GLP and Euclidean distance (ED) codebook are numerically optimized according to their metrics, while the minimum distance of RVQ codebook is averaged over 1000 different RVQ codebooks.

Fig. 3.2. It is shown that the minimum chordal distance of $\mathcal{C}_{\text{ED}}$ is larger than the (averaged) minimum chordal distance of the RVQ codebook for all B_{tot} values.

3.2 Noncoherent Trellis-Coded Quantization (NTCQ)

3.2.1 Euclidean distance codebook design

The observations in the preceding section provide the following practical guidelines for quantization on the Grassmann manifold: (a) find a good codebook in Euclidean space whose structure permits efficient encoding (or, equivalently, find a good, efficiently decodable channel code); (b) use parallel versions of the Euclidean encoder with different amplitude scalings and phase rotations, and choose the best output (or, equivalently, implement block noncoherent decoding efficiently with a number of parallel coherent decoders). The proposed NTCQ emerges naturally from application of these guidelines.

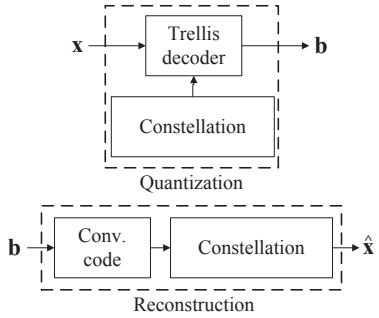


Fig. 3.3.: Quantization and reconstruction processes for a Euclidean distance quantizer using trellis-coded quantization (TCQ).

NTCQ relies on TCQ which was originally proposed in [85], exploiting the functional duality between source coding and channel coding to leverage the well-known trellis-coded modulation (TCM) channel codes designed for coherent communication over AWGN channels [98]. TCM integrates the design of convolutional codes with modulation to maximize the minimum Euclidean distance between modulated codewords. This is done by coding over partitions of the source constellation. Let $\mathcal{C}_{\mathrm{TCM}}$ denote a fixed codebook with N codewords generated by a TCM channel code. Then $\mathcal{C}_{\mathrm{TCM}}$ can be mathematically expressed as

$$\mathcal{C}_{\mathrm{TCM}} = \underset{\mathcal{C}\in\mathcal{V}_{M_t}^N}{\operatorname{argmax}}\, d^2_{\mathrm{E,min}}(\mathcal{C})$$

where $\mathcal{V}_{M_t}^N \subset \mathcal{U}_{M_t}^N$ is the set of $M_t \times N$ complex matrices generated by a given trellis structure with a finite number of constellation points of interest for entries of the matrix. Note that $\mathcal{C}_{\mathrm{TCM}}$ is a Euclidean distance codebook within a given set $\mathcal{V}_{M_t}^N$. Thus, $\mathcal{C}_{\mathrm{TCM}}$ is expected to have a *good* chordal distance property as well.

In TCQ, the decoder and encoder of TCM are used to quantize and reconstruct a given source, respectively. From Fig. 3.3, we see that the TCQ system consists of a source constellation, a trellis-based decoder (for source quantization), and a convolutional encoder (for source reconstruction). Quantization is performed by passing a source vector $\mathbf{x} \in \mathbb{C}^N$ through a trellis-based optimization whose goal is to minimize

a mean square error distortion between the quantized output and the source message input. The additive structure of the square of Euclidean distance implies that the Viterbi algorithm can be employed to efficiently search for a codebook vector that minimizes the Euclidean distance from a given source vector as

$$\mathbf{c}_{\text{opt}} = \underset{\mathbf{c}_i \in \mathcal{C}_{\text{TCM}}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{c}_i\|_2^2, \tag{3.11}$$

which is then mapped to a binary sequence $\mathbf{b} = \text{bin}(\text{opt})$. The quantized source vector $\hat{\mathbf{x}}$ is reconstructed by passing the binary sequence $\mathbf{b}$ into the convolutional encoder and mapping the binary output of the convolutional encoder to points on the source constellation (as if modulating the signal). Due to the linearity of the convolutional code, each unique binary sequence $\mathbf{b}$ represents a unique quantized vector $\hat{\mathbf{x}}$.

NTCQ adopts TCQ to quantize CSI. Note that (3.11) is the same optimization problem as (3.8) with a given $\alpha \in \mathbb{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_{K_\alpha}\}$ and $\theta \in \Theta = \{\theta_1, \theta_2, \ldots, \theta_{K_\theta}\}$. Thus, the minimization (3.8) can be performed using $K_\alpha \cdot K_\theta$ parallel instances of the Viterbi algorithm. This is the same paradigm proposed as in TCQ except for the search over α and θ parameters; due to the presence of these terms, the process is coined *noncoherent trellis-coded quantization*. Note that with PSK constellations, we can set $\alpha = 1$ because all the candidate beamforming vectors $\mathbf{c}_i$'s have the same norm.

We explain the implementation of NTCQ with 8PSK and 16QAM constellations next (we also report results for QPSK, but do not describe the corresponding NTCQ procedure, since it is similar to that for 8PSK). Before explaining the actual implementation, it should be pointed out that, because of the inherited TCM structure, the number of constellation points is larger than 2^B in NTCQ where B is the number of quantization bits per channel entry. We explicitly list the relationship between B and the constellations in Table 3.1. This issue will become clear as we explain the 8PSK implementation.

Table 3.1.: Mapping of quantizing bits/entry (B) and constellations.

B	1 bit/entry	2 bits/entry	3 bits/entry
Constellation	QPSK	8PSK	16QAM

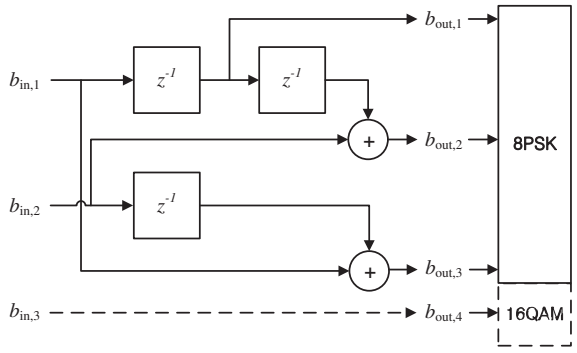


Fig. 3.4.: This rate 2/3 convolutional code corresponds to the trellis in Fig. 3.6. In the figure, the smaller the index the less significant the bit, e.g., $b_{\text{in},1}$ is the least significant input bit and $b_{\text{in},3}$ is the most significant input bit.

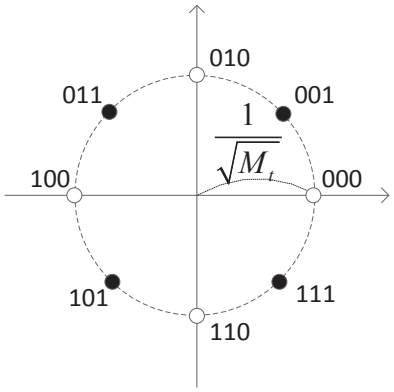


Fig. 3.5.: 8PSK constellation points used in NTCQ are labeled with binary sequences.

3.2.2 NTCQ with 8PSK (2 bits/entry)

We adopt the rate 2/3 convolutional code in [98], as shown in Fig. 3.4. The source constellation is assumed to be 8PSK as in Fig. 3.5. Note that all constellation points are normalized with the number of transmit antennas M_t.

The construction of the feedback sequence is done using a trellis decoder. As is done in traditional decoding of convolutional codes, the encoding process is repre-

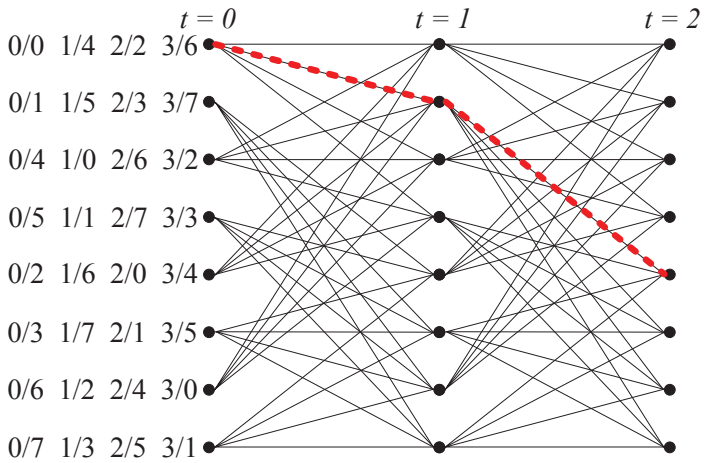


Fig. 3.6.: The Ungerboeck trellis with $S = 8$ states corresponding to the convolutional encoder in Fig. 3.4. The input/output relations using decimal numbers correspond to state transitions from the top to bottom. The example path $\mathbf{p}_2 = [1, 2, 5]$ that corresponds to binary input sequence $[01, 00]^T$ (or decimal input $[1, 0]^T$) and binary output sequence $[100, 001]^T$ (or decimal output $[4, 1]^T$) is highlighted.

sented using a trellis showing the relationship between states of the encoder along with input and output transitions. The trellis with input/output state transitions corresponding to the convolutional code in Fig. 3.4 is shown in Fig. 3.6.

We select candidate beamforming vectors using an M_t-stage trellis where each stage selects an entry in each of the candidate vectors. Thus, each path through the trellis corresponds to a unique candidate beamforming vector. It is important to note that there are only four state-transitions from any of the eight states in Fig. 3.6. Each transition is mapped to one point of the 8PSK constellation. Therefore, even though the source constellation is 8PSK, each element of $\bar{\mathbf{h}}$ is quantized with one of the QPSK subconstellations marked by black or white circles in Fig. 3.5, which results in 2 bits quantization per entry as shown in Table 3.1.

The path choices are enumerated with binary labels, and each path also corresponds to a unique binary sequence. The candidate vector or path that is chosen for output is the one that optimizes the given path metric. The path metric is chosen to reflect the desired Euclidean distance minimization regarding codeword $\mathbf{c}_i$ in (3.8) for

a given α and θ. The output of the quantization is the binary sequence corresponding to the best candidate path.

Each transition from each state at the t^{th} stage, $s_t \in \{1, 2, \ldots, S\}$, in the trellis to a state at the $(t+1)^{th}$ stage, s_{t+1}, corresponds to a point in the source constellation. For example, a transition from state 4 to state 8 corresponds to the binary output sequence 011 which corresponds to the constellation point $\frac{1}{\sqrt{2M_t}}(-1+j)$ in Fig. 3.5. Note that, in this setup, a single entry is chosen at each stage where it is possible to choose more; this is done by using intermediate codebooks for each stage of the trellis. For more details on this method and the design of the codebooks, the reader is referred to [87].

To optimize over the trellis, the first task is to define a path metric. Let $\mathbf{p}_t$ be a partial path, or a sequence of states, up to the stage t. For example, the path $\mathbf{p}_2 = [1, 2, 5]$ using state indices is highlighted in Fig. 3.6. Also, define the two functions $\text{in}(\cdot)$ and $\text{out}(\cdot)$ such that $\text{in}(\mathbf{p}_t)$ outputs the binary input sequence corresponding to path $\mathbf{p}_t$, and $\text{out}(\mathbf{p}_t)$ gives the sequence of output constellation points corresponding to the path $\mathbf{p}_t$. Again, using the sample path $\mathbf{p}_2$ in Fig. 3.6, we can see that

$$\text{in}(\mathbf{p}_2) = [01, 00]^T, \;\; \text{out}(\mathbf{p}_2) = \frac{1}{\sqrt{M_t}}\left[-1, \frac{1}{\sqrt{2}}(1+j)\right]^T.$$

With these definitions, we can define the path metric, $m(\cdot)$, as

$$m(\mathbf{p}_t, \theta) = \|\bar{\mathbf{h}}_t - e^{j\theta}\,\text{out}(\mathbf{p}_t)\|_2^2,$$

where $\theta \in [0, 2\pi)$ and $\bar{\mathbf{h}}_t$ is the vector created by truncating of normalized MISO channel vector $\bar{\mathbf{h}}$ to the first t entries. Note that $\alpha = 1$ because all constellation points have the same magnitude in the 8PSK case. It is easy to check that minimizing over

the path metric will minimize the Euclidean distance. It is also important to notice that the path metric can be written recursively as

$$m(\mathbf{p}_t, \theta) = m(\mathbf{p}_{t-1}, \theta) + \left| \bar{h}_t - e^{j\theta} \operatorname{out}\left([p_{t-1}\ p_t]^T \right) \right|^2,$$

where $\bar{h}_t$ and p_t are the t^{th} entry of $\bar{\mathbf{h}}$ and $\mathbf{p}_t$, respectively. The above path metric can be efficiently computed via the Viterbi algorithm. The path metric is computed in parallel for each quantized value of $\theta \in \Theta = \{\theta_1, \theta_2, \ldots, \theta_{K_\theta}\}$. Then the best path $\mathbf{p}_{\text{best}}$ and the phase θ_{best} that minimize the path metric can be found as

$$\min_{\theta \in \Theta} \min_{\mathbf{p}_{M_t} \in \mathbb{P}_{M_t}} m(\mathbf{p}_{M_t}, \theta)$$

where $\mathbb{P}_{M_t}$ denotes all possible paths up to stage M_t. Finally, the beamforming vector $\mathbf{f}$ is calculated as

$$\mathbf{c}_{\text{opt}} = \operatorname{out}(\mathbf{p}_{\text{best}}),\ \mathbf{f} = \frac{\mathbf{c}_{\text{opt}}}{\|\mathbf{c}_{\text{opt}}\|_2}.$$

Note that $\|\mathbf{c}_{\text{opt}}\|_2 = 1$ for 8PSK; therefore $\mathbf{f} = \mathbf{c}_{\text{opt}}$.

It is important to point out that minimizing over θ only increases the complexity of quantization, not the feedback overhead because the transmitter does not have to know the value of θ_{best} that minimizes the path metric during the beamforming vector reconstruction process. However, there is additional feedback overhead with NTCQ. Since we test all paths in the trellis, the transmitter has to know the starting state of $\mathbf{p}_{\text{best}}$, which causes additional $\log_2 S$ bits of feedback overhead where S is the number of states in the trellis. Therefore, the total feedback overhead is

$$B_{\text{tot}} = BM_t + \log_2 S.$$

The additional feedback overhead $\log_2 S$ bits can vary depending on the trellis used in NTCQ.

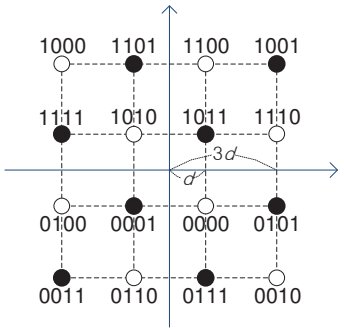


Fig. 3.7.: 16QAM constellation points used in NTCQ are labeled with binary sequences.

3.2.3 NTCQ with 16QAM (3 bits/entry)

For the 16QAM constellation, the rate 3/4 convolution encoder is shown in Fig. 3.4. The source constellation is shown in Fig. 3.7 where $d = \frac{\triangle}{2\sqrt{M_t}}$ with $\triangle = \sqrt{\frac{6}{M-1}}$ with $M = 16$ to have $E[\|\mathbf{c}_i\|_2^2] = 1$ where expectation is taken over $\mathbf{c}_i$ assuming all constellation points are selected with equal probability.

The procedure of NTCQ using 16QAM is basically the same as the 8PSK case. The difference arising for 16QAM is that we have to take α into account during the path metric computation as

$$m(\mathbf{p}_t, \alpha, \theta) = \|\bar{\mathbf{h}}_t - \alpha e^{j\theta} \operatorname{out}(\mathbf{p}_t)\|_2^2 \tag{3.12}$$

where $\theta \in \Theta = \{\theta_1, \theta_2, \ldots, \theta_{K_\theta}\}$ and $\alpha \in \mathbb{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_{K_\alpha}\}$. Similar to the 8PSK case, additional $\log_2 S$ feedback bits are needed to indicate the starting state of $\mathbf{p}_{\text{best}}$ to the transmitter in the 16QAM case.

3.2.4 Complexity

NTCQ relies on a trellis search to quantize the beamforming vector, and the trellis search is performed by the Viterbi algorithm. In each state transition of the trellis, one channel entry is quantized with one of 2^B constellation points. This

computation is performed for S states in each state transition (stage) and there are M_t state transitions in total. Thus, the complexity of the Viterbi algorithm becomes $O(2^B S M_t)$.

The Viterbi algorithm has to be executed $K_\theta \cdot K_\alpha$ times in NCTQ, which gives the overall complexity of $O(K_\theta K_\alpha 2^B S M_t)$. In the limit of large M_t, Theorem 3.1.1 tells us that we can get away with $K_\theta \to 1$ and $K_\alpha \to 1$ without performance loss. However, even for moderate values of M_t, our results in Section 3.4.1 show that small values of K_θ and K_α can be employed with minimal performance degradation. The key aspect to note is the linear scaling of complexity with the number of transmit antennas M_t, which makes NTCQ particularly attractive for massive MIMO systems for which conventional look-up based approaches are computationally infeasible.

3.2.5 Variations of NTCQ

We can also construct several variations of NTCQ with minor tradeoffs between the total number of feedback bits, B_{tot}, and performance. We explain one of the variations briefly below.

- *Variation: Fixing the starting state for the trellis search.*

Because NTCQ searches paths which start from every possible state in the first stage in the trellis, we need an additional $\log_2 S$ bits of feedback overhead to indicate the starting state of $\mathbf{p}_{\text{best}}$. One variation is to fix the first state to eliminate these additional bits, so that the total feedback overhead incurred is exactly BM_t bits. We do incur a small performance loss by doing this, since allowing starting from different states effectively leads to considering more possible values of the scaling parameters α and θ. However, this loss becomes negligible as M_t gets large (consistent with Theorem 3.1.1).

For other variations, we can fix the first entry of $\mathbf{c}_{\text{opt}}$ to a constant in the trellis search or adopt a tail-biting convolutional code.

3.3 Advanced NTCQ Exploiting Channel Correlations

In practice, channels are temporally and/or spatially correlated. In this section, we propose advanced NTCQ schemes that exploit these correlations to improve the performance or reduce the feedback overhead.

3.3.1 Differential scheme for temporally correlated channels

A useful model of this correlation is the first-order Gauss-Markov process [99]

$$\mathbf{h}[k] = \eta \mathbf{h}[k-1] + \sqrt{1-\eta^2}\mathbf{g}[k]$$

where $\mathbf{g}[k] \in \mathbb{C}^{M_t}$ denotes the process noise, which is modeled as having i.i.d. entries distributed with $\mathcal{CN}(0,1)$. We assume that the initial state $\mathbf{h}[0]$ is independent of $\mathbf{g}[k]$ for all k. The temporal correlation coefficient η $(0 \leq \eta \leq 1)$ represents the correlation between elements $h_t[k-1]$ and $h_t[k]$ where $h_t[k]$ is the t^{th} entry of $\mathbf{h}[k]$.

If η is close to one, two consecutive channels are highly correlated and the difference between the previous channel $\mathbf{h}[k-1]$ and the current channel $\mathbf{h}[k]$ might be small. Differential codebooks in [50–57] utilize this property to reduce the channel quantization error with an assumption that both the transmitter and the receiver know η perfectly. Most of the previous literature, however, focused on the case with a fixed and small number of transmit antennas and moderate feedback overhead, e.g., $M_t = 4$ and $B_{\text{tot}} = 4$. Therefore, we have to come up with a new differential feedback scheme to accommodate massive MIMO with large feedback overhead.

We denote $\mathbf{f}[k-1]$ as the quantized beamforming vector at block $k-1$ and

$$\mathbf{f}_{\text{opt}}[k] = \frac{\mathbf{h}[k]}{\|\mathbf{h}[k]\|_2}$$

as the unquantized optimal beamforming vector at time k. In our differential NTCQ scheme, instead of quantizing $\mathbf{h}[k]$ directly at time k, the receiver quantizes $\mathbf{f}_{\text{diff}}[k]$ which is given as

$$\mathbf{f}_{\text{diff}}[k] = \left(\mathbf{I}_{M_t} - \mathbf{f}[k-1]\mathbf{f}^H[k-1]\right)\mathbf{f}_{\text{opt}}[k].$$

Note that $\mathbf{f}_{\text{diff}}[k]$ is a projection of $\mathbf{f}_{\text{opt}}[k]$ to the null space of $\mathbf{f}[k-1]$. We let $\hat{\mathbf{f}}_{\text{diff}}[k]$ denote the quantized version of $\mathbf{f}_{\text{diff}}[k]$ by NTCQ with $\|\hat{\mathbf{f}}_{\text{diff}}[k]\|_2^2 = 1$. The receiver then constructs candidate beamforming vectors $\mathbf{f}_{\bar{\alpha},\bar{\theta}}$ with weights $\bar{\alpha} \in \bar{\mathbb{A}} = \{\bar{\alpha}_1, \ldots, \bar{\alpha}_{K_{\bar{\alpha}}}\}$ and $\bar{\theta} \in \bar{\Theta} = \left\{\bar{\theta}_1, \ldots, \bar{\theta}_{K_{\bar{\theta}}}\right\}$ as

$$\mathbf{f}_{\bar{\alpha},\bar{\theta}} = \frac{\eta\mathbf{f}[k-1] + \bar{\alpha}e^{j\bar{\theta}}\sqrt{1-\eta^2}\hat{\mathbf{f}}_{\text{diff}}[k]}{\left\|\eta\mathbf{f}[k-1] + \bar{\alpha}e^{j\bar{\theta}}\sqrt{1-\eta^2}\hat{\mathbf{f}}_{\text{diff}}[k]\right\|_2}. \tag{3.13}$$

The receiver selects the optimal weights $\bar{\alpha}_{\text{opt}}$ and $\bar{\theta}_{\text{opt}}$ by optimizing

$$\max_{\bar{\alpha}\in\bar{\mathbb{A}}} \max_{\bar{\theta}\in\bar{\Theta}} \left|\bar{\mathbf{h}}^H[k]\mathbf{f}_{\bar{\alpha},\bar{\theta}}\right|^2, \tag{3.14}$$

and the final beamforming vector is given as

$$\mathbf{f}[k] = \mathbf{f}_{\bar{\alpha}_{\text{opt}},\bar{\theta}_{\text{opt}}}.$$

To construct candidate beamformaing vectors as in (3.13), we have to define sets of weights $\bar{\mathbb{A}}$ and $\bar{\Theta}$. It is easy to conclude that $\bar{\Theta} = [0, 2\pi)$ because the quantization process uses beamformer phase invariance. To derive the range of the set $\bar{\mathbb{A}}$, we make the following proposition.

Proposition 3.3.1 *When $\eta \to 1$, the range of $\bar{\mathbb{A}}$ can be set as*

$$\frac{1-\eta}{\sqrt{1-\eta^2}} \le \bar{\alpha} \le \frac{1+\eta}{\sqrt{1-\eta^2}}. \tag{3.15}$$

Proof First, we define $\mathbf{f}_{\bar{\alpha},\bar{\theta}}^{\text{nom}}$ as the numerator of (3.13) as

$$\mathbf{f}_{\bar{\alpha},\bar{\theta}}^{\text{nom}} = \eta \mathbf{f}[k-1] + \bar{\alpha} e^{j\bar{\theta}} \sqrt{1-\eta^2} \hat{\mathbf{f}}_{\text{diff}}[k].$$

Then, the norm square of $\mathbf{f}_{\bar{\alpha},\bar{\theta}}^{\text{nom}}$ becomes

$$\|\mathbf{f}_{\bar{\alpha},\bar{\theta}}^{\text{nom}}\|_2^2 = \eta^2 + \bar{\alpha}^2(1-\eta^2) + 2\bar{\alpha}\sqrt{1-\eta^2}\text{Re}\left\{e^{j\bar{\theta}}\mathbf{f}^H[k-1]\hat{\mathbf{f}}_{\text{diff}}[k]\right\}.$$

Because $-1 \leq \text{Re}\left\{e^{j\bar{\theta}}\mathbf{f}^H[k-1]\hat{\mathbf{f}}_{\text{diff}}[k]\right\} \leq 1$, we have

$$\left(\eta - \bar{\alpha}\sqrt{1-\eta^2}\right)^2 \leq \|\mathbf{f}_{\bar{\alpha},\bar{\theta}}^{\text{nom}}\|_2^2 \leq \left(\eta + \bar{\alpha}\sqrt{1-\eta^2}\right)^2. \tag{3.16}$$

Note that $\mathbf{f}^H[k-1]\hat{\mathbf{f}}_{\text{diff}}[k] \approx 0$ with a good quantizer. Moreover, with the assumption of a slowly varying channel which is typically assumed in the differential codebook literature, we approximate $\eta \approx 1$. Then we have $\|\mathbf{f}_{\bar{\alpha},\bar{\theta}}^{\text{nom}}\|_2^2 = 1$, and plugging this into (3.16) gives the range of $\bar{\alpha}$ in (3.15). ■

Note that the range in (3.15) can be further optimized numerically. In Section 3.4.2, we set $\frac{1-\eta}{\sqrt{1-\eta^2}} \leq \bar{\alpha} \leq \frac{1+\eta}{3\sqrt{1-\eta^2}}$ for simulation. Once the receiver selects the optimal weights $\bar{\alpha}_{\text{opt}}$ and $\bar{\theta}_{\text{opt}}$ by (3.14), it feeds back $\hat{\mathbf{f}}_{\text{diff}}[k]$, $\bar{\alpha}_{\text{opt}}$ and $\bar{\theta}_{\text{opt}}$ to the transmitter over the feedback link and the transmitter reconstructs $\mathbf{f}[k]$ as in (3.13). Additional feedback overhead caused by $\bar{\alpha}_{\text{opt}}$ and $\bar{\theta}_{\text{opt}}$ can be very small compared to the feedback overhead for $\hat{\mathbf{f}}_{\text{diff}}[k]$. Simulation indicates that 1 bit for $\bar{\alpha}_{\text{opt}}$ and 3 bits for $\bar{\theta}_{\text{opt}}$ is sufficient to have near-optimal performance in a low mobility scenario.

3.3.2 Adaptive scheme for spatially correlated channels

If the transmit antennas are closely spaced, which is likely for a massive MIMO scenario, channels tend to be spatially correlated and can be modeled as

$$\mathbf{h}[k] = \mathbf{R}^{\frac{1}{2}}\mathbf{h}_w[k]$$

where $\mathbf{h}_w[k]$ is an uncorrelated MISO channel vector with i.i.d. complex Gaussian entries and $\mathbf{R} = E\left[\mathbf{h}[k]\mathbf{h}^H[k]\right]$ is a correlation matrix of the channel where expectation is taken over k. We assume that $\mathbf{R}$ is a full-rank matrix. For spatially correlated MISO channels, codebook skewing methods were proposed in [47–49] such that codewords in a VQ codebook are rotated and normalized with respect to $\mathbf{R}$ to quantize only the local space of the dominant eigenvector of $\mathbf{R}$. It was shown in [47–49] that this skewing method can significantly reduce the quantization error with the same feedback overhead. With NTCQ, however, there are no fixed VQ codewords for channel quantization which precludes the normal approach for skewing. Therefore, we propose the following method to mimic skewing with NTCQ for spatially correlated MISO channels.

We assume that both the transmitter and the receiver know $\mathbf{R}$ in advance[2]. At the receiver side, $\mathbf{h}_w[k]$ is obtained by decorrelating $\mathbf{h}[k]$ with $\mathbf{R}^{-\frac{1}{2}}$, i.e.,

$$\mathbf{h}_w[k] = \mathbf{R}^{-\frac{1}{2}}\mathbf{h}[k].$$

Then the receiver quantizes $\mathbf{h}_w[k]$ with NTCQ and get $\hat{\mathbf{h}}_w[k]$. The receiver feeds back $\hat{\mathbf{h}}_w[k]$, and the transmitter reconstructs $\mathbf{f}[k]$ as

$$\mathbf{f}[k] = \frac{\mathbf{R}^{\frac{1}{2}}\hat{\mathbf{h}}_w[k]}{\left|\left|\mathbf{R}^{\frac{1}{2}}\hat{\mathbf{h}}_w[k]\right|\right|_2}.$$

This procedure effectively decouples the procedure of exploiting spatial correlation from that of quantization, while providing the same performance gain as standard skewing of fixed codewords.

[2]In practice, the transmitter can acquire an approximate knowledge of $\mathbf{R}$ by averaging $\mathbf{f}[k]$, i.e., $\mathbf{R} \approx E\left[\mathbf{f}[k]\mathbf{f}^H[k]\right]$ where expectation is taken over k.

3.4 Performance Evaluation and Discussions

In this section, we present Monte-Carlo simulation results to evaluate the performance of NTCQ in i.i.d. channels, temporally correlated channels, and spatially correlated channels. In each scenario, we simulate the original NTCQ and its variation, differential NTCQ, and spatially adaptive NTCQ explained in Sections 3.2, 3.3.1, and 3.3.2, respectively. We use the average beamforming gain in dB scale

$$J_{\text{avg}}^{\text{dB}} = 10 \log_{10} \left(E[|\mathbf{h}^H \mathbf{f}|^2] \right)$$

as a performance metric where the expectation is over $\mathbf{h}$.

3.4.1 i.i.d. Rayleigh fading channels

For i.i.d. Rayleigh fading channels, $\mathbf{h}[k]$ is drawn from i.i.d. complex Gaussian entries (i.e., $\mathbf{h}[k] \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$). In Fig. 3.8, we first plot $J_{\text{avg}}^{\text{dB}}$ of NTCQ and its variation in i.i.d. channels with $M_t = 20$ transmit antennas depending on different quantization levels for θ_k and α_k. Clearly, the variation of NTCQ gives strictly lower $J_{\text{avg}}^{\text{dB}}$ than the original NTCQ. Note that it is enough to have $K_\theta = 4$ (2 bits for θ_k) for 1 bit/entry (QPSK) to achieve near-maximal performance of NTCQ and its variation. Interestingly, we can fix $\alpha_k = 1$ with 3 bits/entry (16QAM) for NTCQ and its variation without having any performance loss. This is because when optimizing (3.12), it is likely to have $E\left[\|\mathbf{c}_{\text{opt}}\|_2^2\right] = 1$ since the objective variable is the normalized channel vector $\bar{\mathbf{h}}$ which has a unit norm, i.e., $\|\bar{\mathbf{h}}\|_2^2 = 1$. We fix $K_\theta = 16$ (4 bits for θ_k) for simulations afterward regardless of the number of bits per entry to have a fair comparison. We also fix $\alpha_k = 1$ for 3 bits/entry quantization.

In Fig. 3.9, we plot $J_{\text{avg}}^{\text{dB}}$ for variation of NTCQ (to have the same feedback overhead $B_{\text{tot}} = BM_t$ with the other limited feedback schemes) as a function of the number of quantization bits per entry, B, in i.i.d. Rayleigh channel realizations. We also plot $J_{\text{avg}}^{\text{dB}}$ for unquantized beamforming, RVQ, PSK-SVQ in [91], scalar quantization, and

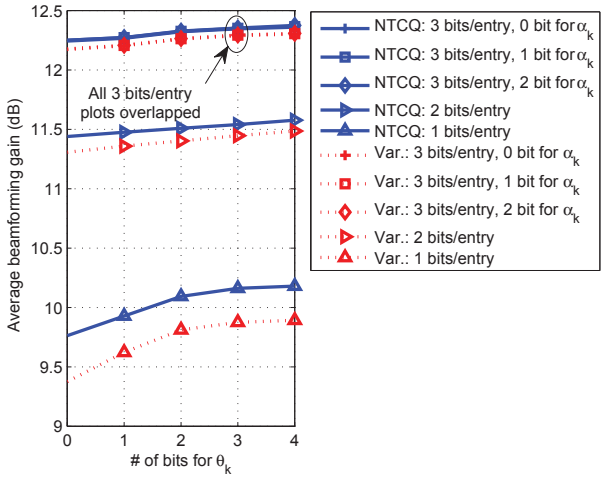


Fig. 3.8.: $J_{\text{avg}}^{\text{dB}}$ vs. different quantization levels of θ_k and α_k with $M_t = 20$ in i.i.d. Rayleigh fading channels.

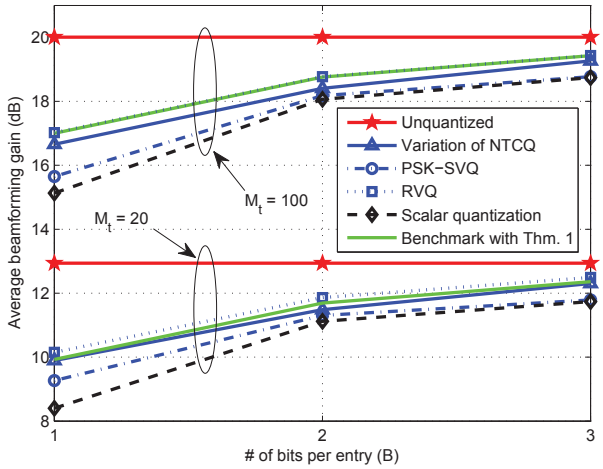


Fig. 3.9.: $J_{\text{avg}}^{\text{dB}}$ vs. B with $M_t = 20$ and 100 in i.i.d. Rayleigh fading channels. PSK-SVQ is from [91]. All limited feedback schemes have the same B_{tot}.

the benchmark from Theorem 3.1.1 which is given as $M_t\left(1-2^{-B}\right)$ (in linear scale). The performance of RVQ is plotted using the analytical approximation in (3.3) as $M_t\left(1-2^{-\frac{B_{\text{tot}}}{M_t-1}}\right)$ (in linear scale), because it is computationally infeasible to simulate when the number of feedback bits grows large. In scalar quantization, B bits are

used to quantize only the phase, not the amplitude, of each channel entry because the phase is generally more important than the amplitude in beamforming [100].

As the number of feedback bits increases, the gap between the unquantized case and all limited feedback schemes decreases as expected. RVQ gives the best performance among limited feedback schemes with the same number of feedback bits. However, the difference between $J_{\text{avg}}^{\text{dB}}$ for RVQ and variation of NTCQ is small for all B. The plots of the benchmark using Theorem 3.1.1 well approximate $J_{\text{avg}}^{\text{dB}}$ of NTCQ for all B and M_t, which shows the near-optimality of NTCQ. Note that variation of NTCQ achieves better $J_{\text{avg}}^{\text{dB}}$ than PSK-SVQ regardless of B and M_t, and the gap becomes larger as M_t increases. This gap comes from the coding gain of NTCQ. As shown in Table 3.1, NTCQ can exploit 2^{B+1} constellation points while PSK-SVQ only utilizes 2^B constellation points with B bits quantization per entry. The coding gain of variation of NTCQ is around 0.25 to 1dB depending on M_t and B. Although we do not plot the performance of QAM-SVQ which relies on QAM constellations, it has the same structure as PSK-SVQ meaning that QAM-SVQ roughly experiences the same performance degradation compared to NTCQ.

3.4.2 Temporally correlated channels

To simulate the differential feedback schemes with the original NTCQ algorithm in temporally correlated channels, we adopt Jakes' model [84] to generate the temporal correlation coefficient $\eta = J_0(2\pi f_D \tau)$, where $J_0(\cdot)$ is the 0th order Bessel function of the first kind, f_D denotes the maximum Doppler frequency, and τ denotes the channel instantiation interval. We assume a carrier frequency of 2.5 GHz and $\tau = 5ms$. We set the quantization level for the combiners $\bar{\theta}$ and $\bar{\alpha}$ in (3.13) as 3 bits and 1 bit, respectively, which causes 4 bits of additional feedback overhead.

In Fig. 3.10, we plot the performance of the proposed differential NTCQ feedback schemes with the velocity $v = 3km/h$ ($\eta = 0.9881$) assuming no feedback delay. The differential NTCQ schemes, even with 1 bit/entry quantization, achieve almost the

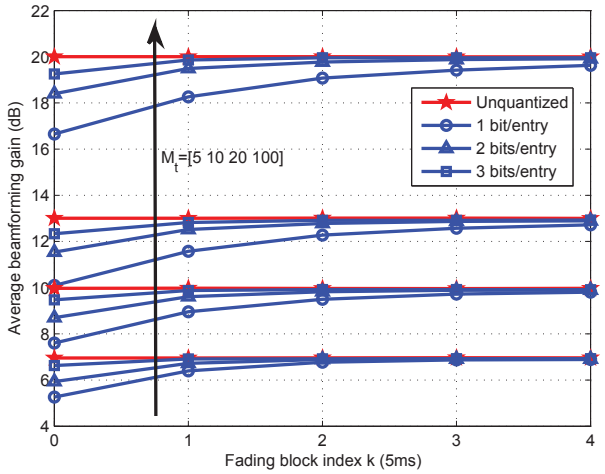


Fig. 3.10.: $J_{\text{avg}}^{\text{dB}}$ vs. fading block index k with $v = 3km/h$ in temporally correlated channels. Without feedback delay.

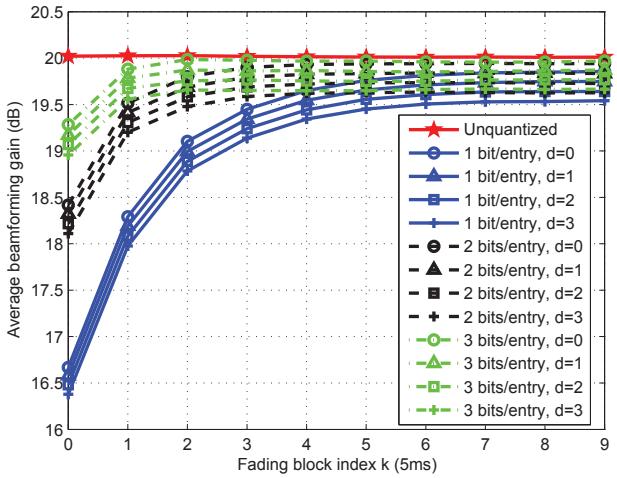


Fig. 3.11.: $J_{\text{avg}-\text{delay}}^{\text{dB}}[d]$ vs. fading block index k with $M_t = 100$, d blocks of feedback delay, and $v = 3km/h$ in temporally correlated channels.

same performance as unquantized beamforming regardless of M_t. Thus, if we can adjust the feedback overhead as a function of time, we can switch from NTCQ with 2 or 3 bits/entry quantization to 1bit/entry quantization in differential NTCQ to reduce the overall feedback overhead.

To see the effect of feedback delay in temporally correlated channels, we simulate the $M_t = 100$ case with different numbers of delay d measured in fading blocks (one fading block corresponds to $5ms$) in Fig. 3.11 such that

$$J^{\mathrm{dB}}_{\mathrm{avg-delay}}[d] = 10\log_{10}\left(E[|\mathbf{h}^H[k]\mathbf{f}[k-d]|^2)]\right).$$

It is shown that the effect of feedback delay is negligible, i.e., around 0.1dB loss with one additional block delay for all cases, which confirms the practicality of the differential NTCQ scheme. Moreover, we can reduce the frequency of the feedback updates to reduce the total amount of feedback overhead without significant performance degradation when the velocity of the receiver is low.

3.4.3 Spatially correlated channels

To generate spatially correlated channels, we adopt the Kronecker model for the spatial correlation matrix $\mathbf{R}$ which is given as $\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$ where $\mathbf{U}$ and $\mathbf{\Sigma}$ are $M_t \times M_t$ eigenvector and diagonal eigenvalue matrices, respectively. The performance of the adaptive scheme will highly depend on the amount of spatial correlation. To see the effect of spatial correlation, we assume the eigenvalue matrix $\mathbf{\Sigma}$ has a structure given by

$$\mathbf{\Sigma} = \mathrm{diag}\left\{\lambda_1, \frac{M_t - \lambda_1}{M_t - 1}, \cdots, \frac{M_t - \lambda_1}{M_t - 1}\right\}$$

where $1 \leq \lambda_1 < M_t$ is the dominant eigenvalue of $\mathbf{R}$. If λ_1 is small (large), the channels are loosely (highly) correlated in spatial domain. Note that channels are i.i.d. when $\lambda_1 = 1$.

In Fig. 3.12, and 3.13, we plot $J^{\mathrm{dB}}_{\mathrm{avg}}$ as a function of λ_1 for $M_t = 10$ and 20 cases. The performance of spatially adaptive NTCQ become closer to that of unquantized beamforming as λ_1 increases with the same feedback overhead as original NTCQ. This shows the effectiveness of the proposed adaptive NTCQ scheme for spatially correlated channels.

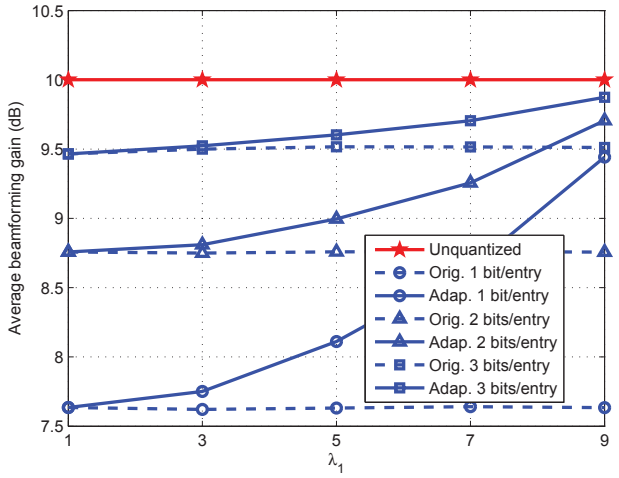


Fig. 3.12.: $J_{\text{avg}}^{\text{dB}}$ vs. λ_1 with $M_t = 10$ in spatially correlated channels.

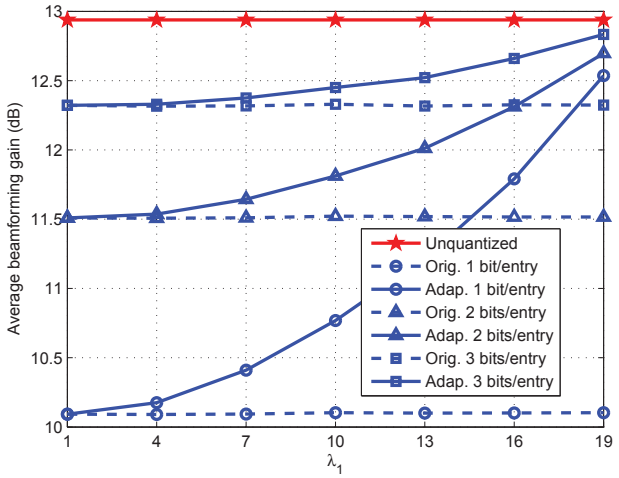


Fig. 3.13.: $J_{\text{avg}}^{\text{dB}}$ vs. λ_1 with $M_t = 20$ in spatially correlated channels.

While we have developed an efficient channel quantization method for massive MIMO systems, we note that limitations on feedback overhead would typically prevent scaling to an indefinitely large number of antennas. However, the feedback overhead may be reasonable for the moderately large number of antennas (32 to 64) expected in initial deployments [86], and NCTQ represents a computationally efficient approach

to generating such feedback. Moreover, we propose TEC and TE-SPA that can reduce the feedback overhead of NTCQ in the next chapter.

4. TRELLIS-EXTENDED CODEBOOKS AND SUCCESSIVE PHASE ADJUSTMENT: A PATH FROM LTE-ADVANCED TO FDD MASSIVE MIMO SYSTEMS

Note that the minimum feedback overhead of NTCQ is one bit per channel entry, which could prevent its use in certain scenarios that need small feedback overhead. In this chapter, we explain TEC and TE-SPA codebooks that can achieve a fractional number of bits per channel entry quantization. The TEC and TE-SPA codebooks can be used similarly to the LTE-Advanced dual codebooks, i.e., TEC quantizes long-term/wideband CSI while TE-SPA quantizes short-term/subband CSI. This unified structure for long-term/wideband and short-term/subband CSI quantization is a significant benefit compared to other stand-alone CSI quantization schemes for massive MIMO systems.

4.1 System Model

To simplify explanation, we first consider a block fading MISO channel with M_t transmit antennas at the base station and a single receive antenna at the user. The proposed TEC can be easily extended to a multiple receive antenna case as explained in Section 4.2.3. With the block fading assumption, the received signal for each channel use in the kth fading block, $y[k] \in \mathbb{C}$, is written as

$$y[k] = \sqrt{P}\mathbf{h}^H[k]\mathbf{f}[k]s + z[k],$$

[0]

where P is the transmit power, $\mathbf{h}[k] \in \mathbb{C}^{M_t}$ is the MISO channel vector, $\mathbf{f}[k] \in \mathbb{C}^{M_t}$ is the unit norm beamforming vector, $s \in \mathbb{C}$ is the message signal[1] satisfying $E[s] = 0$ and $E[|s|^2] = 1$, and $z[k] \sim \mathcal{CN}(0, \sigma^2)$ is complex additive white Gaussian noise.

For CSI quantization, we assume that the total number of feedback bits B_{tot} scales linearly with M_t as

$$B_{tot} \triangleq BM_t$$

where B is the number of quantization bits per transmit antenna. The linear relationship of the feedback overhead with the number of antennas is necessary to achieve a certain level of channel quantization error [14] or a full multiplexing gain of MU-MIMO (assuming the conditions on the feedback rate and SNR dependence are satisfied) [95, 96].

If we rely on the conventional approach of using a B_{tot}-bit unstructured vector quantization (VQ) codebook $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_{2^{B_{tot}}}\}$ that consists of unit norm codewords for CSI quantization, the user quantizes its channel by selecting the codeword $\mathbf{c}_{\text{opt}}[k]$ that aligns with the channel most closely as

$$\mathbf{c}_{\text{opt}}[k] = \underset{\mathbf{c} \in \mathcal{C}}{\operatorname{argmax}} |\mathbf{h}^H[k]\mathbf{c}|^2. \tag{4.1}$$

The user then feeds back the binary index of $\mathbf{c}_{\text{opt}}[k]$, i.e., $\mathbf{b}[k] = \text{bin(opt)}$ where $\text{bin}(\cdot)$ converts an integer to its binary representation, to the base station. If the base station adopts maximum ratio transmission (MRT) beamforming, which is popular due to its simplicity for massive MIMO [5], we have $\mathbf{f}[k] = \mathbf{c}_{\text{opt}}[k]$.

Note that the codeword search complexity of using a VQ codebook is $O(M_t 2^{BM_t})$. If B_{tot} or M_t is small as in current cellular systems, the complexity of CSI quantization is not a problem. However, in massive MIMO systems with a very large number of M_t, brute force codeword selection becomes infeasible.

[1]The message signal changes in every channel. We neglect the index for channel use for brevity.

4.2 Trellis-Extended Codebook (TEC)

TEC can exploit and extend preexisting VQ codebooks such as LTE or LTE-Advanced codebooks. Because of its backward compatibility, TEC is an excellent candidate for CSI quantization in future FDD massive MIMO systems. We first explain the concept and the procedure of TEC. We then discuss the codeword-to-branch mapping and codebook design criteria to maximize the performance of TEC. Because we do not consider temporal correlation of channels in this section, we drop the block index k to simplify notations for the remainder of this section.

4.2.1 Concept and procedure of TEC

Similar to [87] and NTCQ, TEC exploits a trellis decoder and a convolutional encoder in channel coding as a CSI quantizer and a CSI reconstructor, respectively. Low-dimensional VQ codewords (e.g., codebooks designed for smaller arrays) are mapped to trellis branches to quantize multiple channel entries simultaneously by the Viterbi algorithm. We first explain the concept of TEC in detail. Then, we summarize the procedure of TEC.

Like NTCQ, TEC is based on the equivalence between the two optimization problems

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{C}^N}{\operatorname{argmin}} \min_{\theta\in[0,2\pi)} \left\| \mathbf{y} - e^{j\theta} \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \right\|_2^2$$

and

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{C}^N}{\operatorname{argmax}} \frac{|\mathbf{y}^H\mathbf{x}|^2}{\|\mathbf{x}\|_2^2}. \tag{4.2}$$

Note that (4.2) is the same as (4.1). Thus, with the constraint of $\|\mathbf{c}\|_2^2 = 1$, we can transform the CSI quantization problem in (4.1) to

$$\mathbf{c}_{\text{opt}} = \underset{\mathbf{c}\in\mathcal{C}}{\operatorname{argmin}} \min_{\theta\in[0,2\pi)} \left\| \mathbf{h} - e^{j\theta}\mathbf{c} \right\|_2^2. \tag{4.3}$$

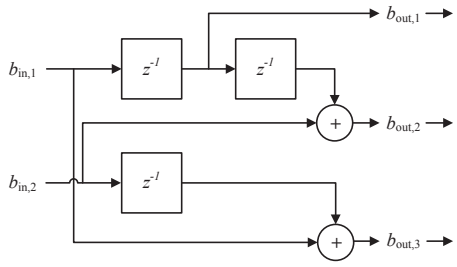


Fig. 4.1.: A rate $\frac{2}{3}$ convolutional encoder that can be used to generate a TEC codebook. In the figure, $b_{\text{in},1}$ and $b_{\text{in},2}$ are the least significant and the most significant input bits, respectively. Same for the output bits.

Instead of optimizing θ over the continuous space $[0, 2\pi)$, we can discretize the search space, i.e., $\theta \in \Theta = \{\theta_1, \ldots, \theta_{K_\theta}\}$, as in noncoherent sequence detection [94]. With a given θ, (4.3) can be efficiently solved by well-known source coding techniques such as TCQ or trellis quantizer [85]. This conversion is successfully exploited in [91] and NTCQ to develop efficient CSI quantizers. TEC also solves (4.3) using trellis quantizers similar to NTCQ. The main difference is that NTCQ handles one channel entry per state transition of the trellis search while TEC processes multiple channel entries simultaneously.

TEC can be implemented using any trellis quantizer. In this work, we adopt the Ungerboeck trellis and convolutional encoder [98] because of their simplicity and good performance. Let B_{in} and B_{out} be the number of input and output bits of a convolutional encoder of interest, respectively. The Ungerboeck convolutional encoder satisfies $B_{out} = B_{in} + 1$. Note that each state in the trellis of the corresponding convolutional encoder has $2^{B_{in}}$ branches; however, the total number of distinctive branch labels is $2^{B_{out}}$. An example of a rate $\frac{2}{3}$ convolutional encoder from [98] (with $B_{in} = 2$ and $B_{out} = 3$) and the corresponding trellis are shown in Fig. 4.1 and 4.2, respectively. As shown in Fig. 4.2, each state has four branches differentiated with inputs and even or odd outputs.

Let L denote the number of simultaneously quantized channel elements in a state transition of a trellis. We assume that L divides the number of transmit antennas

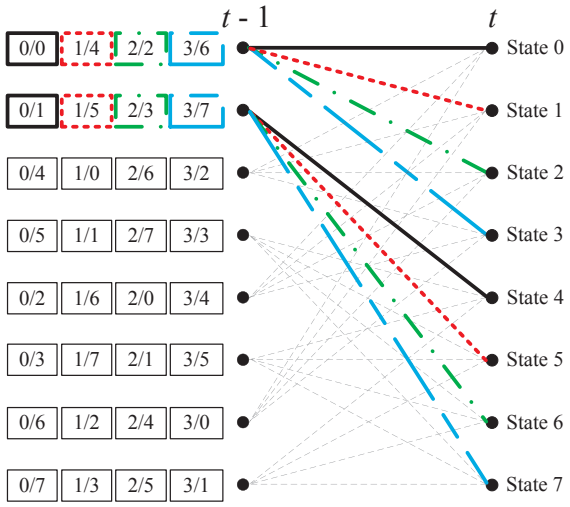


Fig. 4.2.: The trellis representation of the convolutional encoder in Fig. 4.1. Each state transition in the right side is mapped with input/output relation using decimal numbers in each box in the left. For example, 1/4 (in decimal numbers) in the top red-dot box represents the state transition from the state 0 to the state 1 with input=01/output=100 (all in binary numbers).

M_t. Note that TEC supports $B = \frac{B_{in}}{L}$ bits per channel entry quantization, which will become clear later. Thus, if $L > B_{in}$, TEC can achieve a fractional number of bits per channel entry quantization.

To process L channel entries per state transition, TEC maps $L \times 1$ codewords $\mathbf{c}_k^L \in \mathbb{C}^L$ to branches in the trellis. To do this, we need to have a VQ codebook (such as the LTE codebook) with $2^{B_{out}}$ codewords, i.e., $\mathcal{C}_{2^{B_{out}}}^L = \{\mathbf{c}_0^L, \ldots, \mathbf{c}_{2^{B_{out}}-1}^L\}$, to assign all $2^{B_{out}}$ branches of the trellis with different output. We will discuss the codeword-to-branch (or outputs) mapping and the codebook design criteria later. For the time being, we assume that all $2^{B_{out}}$ branches are mapped with some codewords.

To perform the trellis search using the Viterbi algorithm, we need to define a path metric to solve (4.3). Let $\mathbf{p}_t$ be a partial path up to the stage t in the trellis. We also define $\text{in}(\mathbf{p}_t)$ as the binary input sequence corresponding to path $\mathbf{p}_t$ and $\text{out}(\mathbf{p}_t)$ as the sequence of codewords $\mathbf{c}_k^L$'s that are mapped to branches in the path $\mathbf{p}_t$. Note that

$\text{out}(\mathbf{p}_t) \in \mathbb{C}^{Lt}$ where each block of L entries of $\text{out}(\mathbf{p}_t)$ is from a specific codeword $\mathbf{c}_k^L$. Then we can define the path metric based on (4.3) as

$$\begin{aligned} m(\mathbf{p}_t, \theta) &= \left\| \mathbf{h}_{[1:Lt]} - e^{j\theta} \text{out}(\mathbf{p}_t) \right\|_2^2 \\ &= m(\mathbf{p}_{t-1}, \theta) + \left\| \mathbf{h}_{[L(t-1)+1:Lt]} - e^{j\theta} \text{out}([p_{t-1}\ p_t]) \right\|_2^2 \end{aligned} \tag{4.4}$$

where $\mathbf{h}_{[m:n]}$ is the truncated vector of $\mathbf{h}$ from the mth entry to the nth entry. The path metric in (4.4) can be efficiently computed for a given candidate value of θ using the Viterbi algorithm where the total number of stages in the trellis is equal to $T = \frac{M_t}{L}$. The pair $(\mathbf{p}_{\text{best}}, \theta_{\text{best}})$ that minimizes the path metric in (4.4) is given by solving

$$\min_{\theta \in \Theta} \min_{\mathbf{p}_T \in \mathbb{P}_T} m(\mathbf{p}_T, \theta)$$

where $\mathbb{P}_T$ denotes the set of all possible paths up to stage T. The best codeword $\mathbf{c}_{\text{opt}}$ and the binary feedback sequence $\mathbf{b}$ are given as

$$\mathbf{c}_{\text{opt}} = \text{out}(\mathbf{p}_{\text{best}}), \quad \mathbf{b} = \text{in}(\mathbf{p}_{\text{best}}), \tag{4.5}$$

respectively. If we normalize $\mathbf{c}_k^L$ as $\|\mathbf{c}_k^L\|_2^2 = \frac{L}{M_t}$ for all k, then we have $\|\mathbf{c}_{\text{opt}}\|_2^2 = 1$. It is important to point out that $\mathbf{b}$ consists of input bits (not output bits) of the convolutional encoder, which results in $B = \frac{B_{in}}{L}$ bits per channel entry quantization.

The procedure of TEC can be summarized as follows: 1) for a given θ, find the path $\mathbf{p}_T$ that minimizes the path metric defined in (4.4) by running the Viterbi algorithm; 2) among selected candidate paths depending on θ, select θ_{best} (and corresponding $\mathbf{p}_{\text{best}}$) that gives the minimum path metric; 3) $\mathbf{p}_{\text{best}}$ is converted to the binary feedback sequence $\mathbf{b}$ as in (4.5) and $\mathbf{b}$ is fed back to the base station; and 4) the base station reconstructs $\mathbf{c}_{\text{opt}}$ based on $\mathbf{b}$.

Note that searching over θ only increases complexity, not the feedback overhead of TEC. The base station only needs to know the binary feedback sequence $\mathbf{b}$ that represents the best path $\mathbf{p}_{\text{best}}$ to reconstruct $\mathbf{c}_{\text{opt}}$ using the convolutional encoder. We

fix the starting state of the trellis search to the first state. Otherwise, we need an additional feedback overhead to indicate the starting state of the best path.

4.2.2 Codeword-to-branch mapping and codebook design criteria for TEC

To exploit a preexisting VQ codebook in TEC, we need a clever mapping rule between codewords in $\mathcal{C}^L_{2^{B_{tot}}}$ and branches in the trellis. The mapping rule should depend on the structure of the given trellis or convolutional encoder. We propose a mapping rule for the trellis structure in Fig. 4.2 with an arbitrary codebook $\mathcal{C}^L_{2^{B_{tot}}}$. Similar mapping rules can be defined for other trellis structures.

1) Codeword-to-branch mapping rule for Fig. 4.2: Because the branch labels do not vary with θ, we need to separately maximize the minimum Euclidean distance between codeword pairs that are mapped to all even and odd outputs in Fig. 4.2. To further optimize the mapping, we also need to consider the distinctive pairs of paths in Fig. 4.3 because we fix the starting state of the trellis search as the first state in TEC. Considering the red-solid paths and the first state transition of the blue-dot paths, all even outputs are interconnected with each other. For odd outputs, however, we can further maximize the Euclidean distance between the two codewords that are mapped to outputs $\{1, 5\}$ and $\{3, 7\}$.

To realize this, with some abuse of notation, let $\mathcal{C}^L_1$ and $\mathcal{C}^L_2$ denote all possible partitions of $\mathcal{C}^L_{2^{B_{tot}}}$ satisfying

$$\mathcal{C}^L_1 \cup \mathcal{C}^L_2 = \mathcal{C}^L_{2^{B_{tot}}},$$

$$\mathcal{C}^L_1 \cap \mathcal{C}^L_2 = \phi,$$

$$\text{card}(\mathcal{C}^L_1) = \text{card}(\mathcal{C}^L_2) = 2^{B_{tot}-1}$$

where $\text{card}(\cdot)$ is the cardinality of an associated set and ϕ denotes the empty set. Let $\mathbf{c}_{m,k} \in \mathcal{C}^L_k$ for $k = 1, 2$. We denote $\mathcal{C}^L_{odd}$ and $\mathcal{C}^L_{even}$ as the set of codewords mapped to

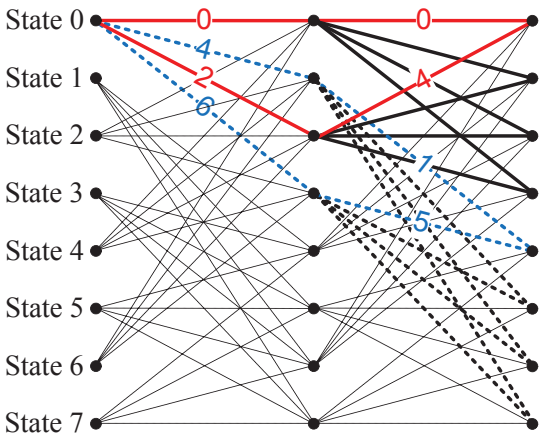


Fig. 4.3.: Distinctive pairs of paths of which the Euclidean distance should be maximized. Two pairs of paths are highlighted with trellis outputs.

the trellis branches of odd and even outputs, respectively. We generate $\mathcal{C}_{odd}^{L}$ and $\mathcal{C}_{even}^{L}$ as

$$\mathcal{C}_{odd}^{L} = \underset{\mathcal{C}_1^L \subset \mathcal{C}^L}{\operatorname{argmax}} \min_{m \neq n} \|\mathbf{c}_{m,1} - \mathbf{c}_{n,1}\|_2^2 ,$$
$$\mathcal{C}_{even}^{L} = \underset{\mathcal{C}_2^L \subset \mathcal{C}^L}{\operatorname{argmax}} \min_{m \neq n} \|\mathbf{c}_{m,2} - \mathbf{c}_{n,2}\|_2^2 , \tag{4.6}$$

respectively. Once we have $\mathcal{C}_{odd}^{L}$ and $\mathcal{C}_{even}^{L}$ as above, we can have arbitrary mappings between the codewords in $\mathcal{C}_{even}^{L}$ and the trellis branches of even outputs. For the trellis branches of odd outputs, however, we need one more step. We divide $\mathcal{C}_{odd}^{L}$ into $\mathcal{C}_{odd,1}^{L}$ and $\mathcal{C}_{odd,2}^{L}$ as we divide $\mathcal{C}_{2^{B_{tot}}}^{L}$ into $\mathcal{C}_{odd}^{L}$ and $\mathcal{C}_{even}^{L}$ in (4.6). Then, we map the codewords in $\mathcal{C}_{odd,k}^{L}$ to the trellis branches with outputs $\{(2k-1),(2k+3)\}$ for $k = 1, 2$.

2) Codebook design criterion: Instead of reusing conventional codebooks, we can also design a codebook that is optimized for TEC. Note that the second term of the path metric in (4.4) is the quantization problem in Euclidean space. Thus, we can generate

a codebook with $2^{B_{out}}$ codewords of dimension $L \times 1$ that maximize the minimum Euclidean distance between all possible codeword pairs as

$$\mathcal{C}^{L}_{\mathrm{ED},2^{B_{out}}} = \underset{\mathcal{C}\in\mathcal{U}_L^{2^{B_{out}}}}{\operatorname{argmax}}\, d^2_{ED,\min}(\mathcal{C}) \tag{4.7}$$

where $\mathcal{U}_L^N \in \mathbb{C}^{L\times N}$ is the set of all $L \times N$ complex matrices with unit norm columns and

$$d^2_{ED,\min}(\mathcal{C}) \triangleq \min_{1\le k<l\le 2^N} \|\mathbf{c}_k - \mathbf{c}_l\|_2^2$$

with $\mathbf{c}_k, \mathbf{c}_l \in \mathcal{C}$.

The proposed codebook design criterion exploits the same concept as the GLP codebook that maximizes the minimum chordal distance between all codeword pairs [43,44]. The difference is that the GLP codebook directly quantizes a channel on the Grassmann manifold while the proposed codebook works in Euclidean space.

Remark: A similar codebook design and codeword-to-branch mapping criteria have been proposed in [87]. However, [87] first generates the $L\times 1$ Euclidean codebook with $2^{B_{in}}$ codewords (not $2^{B_{out}}$ codewords as in the proposed scheme) that are mapped to odd (or even) outputs. With some abuse of notation, denote this Euclidean codebook $\mathcal{C}^L_{odd}$. Then $\mathcal{C}^L_{even}$ is generated by rotating $\mathcal{C}^L_{odd}$ with a unitary matrix $\mathbf{U}$ where $\mathbf{U}$ is designed to maximize the minimum chordal distance between codewords in $\mathcal{C}^L_{odd}\cup\mathcal{C}^L_{even}$. Because $\mathbf{U}$ tries to maximize the minimum chordal distance, not the minimum Euclidean distance, the approach in [87] cannot guarantee to maximize the minimum Euclidean distance between all possible pairs of codewords generated by TEC. Moreover, [87] cannot easily utilize an existing VQ codebook different from TEC.

4.2.3 TEC for multiple receive antennas

We extend the proposed TEC to accommodate MIMO with M_r receive antennas at the user. Assume that $M_t \ge M_r$ and the base station transmits $K \le M_r$ data streams simultaneously. If we rely on a VQ codebook, we select the matrix codeword

(or precoder matrix) $\mathbf{F} \in \mathbb{C}^{M_t \times K}$ that maximizes the instantaneous achievable rate which is defined as

$$R_{ach} = \log_2 \det \left(\mathbf{I}_K + \frac{P}{\sigma^2 K} \mathbf{F}^H \mathbf{H} \mathbf{H}^H \mathbf{F} \right) \tag{4.8}$$

where $\frac{P}{\sigma^2}$ is SNR and $\mathbf{H} \in \mathbb{C}^{M_t \times M_r}$ is the channel matrix. It is not possible to follow this approach with TEC because TEC does not explicitly use a codebook of precoders. Instead, we quantize the first K dominant eigenvectors of $\mathbf{H}\mathbf{H}^H$, which is denoted as $\mathbf{U}(\mathbf{H}) \in \mathbb{C}^{M_t \times K}$. For this case, we need to use matrix codewords $\mathbf{C}_k^{L \times K} \in \mathbb{C}^{L \times K}$ (which has orthogonal columns) to quantize $\mathbf{U}(\mathbf{H})$. We can rewrite the path metric defined in (4.4) as[2]

$$\begin{aligned} m(\mathbf{p}_t, \theta) &= \left\| \mathbf{U}(\mathbf{H})_{[1:Lt]} - e^{j\theta} \mathrm{out}(\mathbf{p}_t) \right\|_F^2 \\ &= m(\mathbf{p}_{t-1}, \theta) + \left\| \mathbf{U}(\mathbf{H})_{[L(t-1)+1:Lt]} - e^{j\theta} \mathrm{out}([p_{t-1}\ p_t]) \right\|_F^2 \end{aligned}$$

where $\mathbf{A}_{[m:n]}$ is the truncated matrix of $\mathbf{A}$ from the mth row to the nth row, and $\|\mathbf{A}\|_F$ denotes the Frobenius norm of a matrix $\mathbf{A}$. We can use the same codebook design and codeword-to-branch mapping criteria to the multiple receive antenna case by changing the 2-norm operation to a Forbenius norm operation. Because we restrict the matrix codewords to have orthogonal columns, the columns of the final selected precoder $\mathbf{F}$ are also orthogonal with each other.

4.3 Trellis-Extended Successive Phase Adjustment (TE-SPA)

In practice, channels are correlated in time and space. There has been much work on differential codebooks that leverage the temporal correlation of channels for better CSI quantization, e.g., [50–57]. However, most of those works focused on a small number of transmit antennas and feedback bits. Thus, we first propose TE-

[2]We can rotate each column of $\mathrm{out}(\mathbf{p}_t)$ separately using different values of θ to optimize the path metric with additional search complexity.

SPA which is a differential codebook version of TEC for temporally correlated massive MIMO systems. Later, we show that TE-SPA can be applied to spatially correlated channels as well.

We consider temporally correlated channels that are modeled by a first order Gauss-Markov process as

$$\mathbf{h}[k] = \eta \mathbf{h}[k-1] + \sqrt{1-\eta^2}\mathbf{g}[k] \tag{4.9}$$

where $0 \leq \eta \leq 1$ is the correlation coefficient, $\mathbf{h}[k]$ is the channel realization at time k, and $\mathbf{g}[k]$ is the innovation process at time k. We assume that $\mathbf{h}[0]$ is independent of $\mathbf{g}[k]$ for all k. Note that the model in (4.9) is also applicable to frequency correlated channels if k denotes the subcarrier or subband index of a wideband channel.

If the channel variation is small in time, i.e., η is close to 1, we can successively reduce quantization error by adjusting the phase of each entry or the block of entries of previous CSI. TE-SPA adjusts phases in a block-wise manner to reduce the feedback overhead. TE-SPA consists of *block-wise phase adjustment matrix generation* and *block shifting*.

4.3.1 Block-wise phase adjustment matrix generation

Let $\hat{\mathbf{h}}_{k-1} = \mathbf{c}_{\text{opt}}[k-1]$ and $\mathbf{h}_k = \mathbf{h}[k]$ represent the previous (quantized) CSI and the current channel vector, respectively, to simplify notations. TE-SPA quantizes the channel at time k by adjusting the phases of $\hat{\mathbf{h}}_{k-1}$ in a block-wise manner. That is, $\hat{\mathbf{h}}_{k-1}$ is rotated with a block-wise phase adjustment matrix $\mathbf{P}_k$ which is given as[3]

$$\mathbf{P}_k = \text{diag}\left(\left[e^{j\varphi_{k,1}}, \ldots, e^{j\varphi_{k,T}}\right] \otimes \mathbf{1}_L\right) \tag{4.10}$$

[3]The block length L with the same phase $\varphi_{k,n}$ in $\mathbf{P}_k$ is a design parameter and does not need to be the same as that of TEC. We assume the length of L is the same as in TEC for simple explanation.

where $T = \frac{M_t}{L}$, $\otimes$ is the Kronecker product, and $\mathbf{1}_L = [1, \ldots, 1]^T$ is the length L all 1 vector. Then, the quantized version of the current CSI becomes

$$\hat{\mathbf{h}}_k = \mathbf{P}_k \hat{\mathbf{h}}_{k-1}.$$

TE-SPA exploits the trellis structure as in TEC to generate $\mathbf{P}_k$, i.e., TE-SPA selects $\{\varphi_{k,n}\}$ from a given set $\Psi = \{\psi_1, \ldots, \psi_{2^{B_{out}}}\}$ as

$$(\varphi_{k,1}, \ldots, \varphi_{k,T}) = \underset{\varphi_{k,n} \in \Psi}{\operatorname{argmin}} \min_{\theta \in \Theta} \left\| \mathbf{h}_k - e^{j\theta} \mathbf{P}_k \hat{\mathbf{h}}_{k-1} \right\|_2^2 \tag{4.11}$$

using the Viterbi algorithm. Note that the convolutional encoders for TEC and TE-SPA can be different, e.g., we could adopt a rate $\frac{2}{3}$ convolutional encoder for TEC while a rate $\frac{1}{2}$ convolutional encoder is used for TE-SPA to reduce successive feedback overhead.

To quantize CSI effectively, we need to appropriately set the values of the elements in Ψ and assign those elements to the trellis branches, which are exactly the same principles as the codebook design and the codeword-to-trellis branch mapping criteria in TEC. Previous works on differential codebook design tried to optimize codebook update methods taking the temporal correlation coefficient η into account. In TE-SPA, this is implicitly handled during the trellis search, i.e., the trellis search selects the best set of phases for $\mathbf{P}_k$ which rotates the previous CSI "close" to the current channel. Therefore, it is better to have values of elements in Ψ such that they are able to generate various rotation matrices as possible. Note that $\mathbf{P}_k$ is determined by the relation among $\{\varphi_{k,n}\}$. If $T = 2$, then $\mathrm{diag}([1, e^{j\frac{\pi}{4}}] \otimes \mathbf{1}_L)$ is the same as $\mathrm{diag}([e^{j\frac{7\pi}{4}}, 1] \otimes \mathbf{1}_L)$ in terms of $\mathbf{P}_k$. Thus, we restrict the search space to $[0, \pi)$ and assign the values in Ψ as[4]

$$\psi_\nu = \frac{\nu - 1}{2^{B_{out}}} \pi, \quad \nu = 1, \ldots, 2^{B_{out}}.$$

[4]For large T, searching over $[0, 2\pi)$ would give different choices of $\mathbf{P}_k$; however, a larger search space gives coarse quantization for ψ_ν resulting in performance degradation.

$$(\varphi_{k,1},\ldots,\varphi_{k,T}) = \underset{\varphi_{k,n}\in\Psi}{\operatorname{argmin}} \min_{\theta\in\Theta} \left\| \mathbf{h}_k \left[\frac{L}{2}(k-1)\right]_c - e^{j\theta}\mathbf{P}_k \hat{\mathbf{h}}_{k-1} \left[\frac{L}{2}(k-1)\right]_c \right\|_2^2, \quad k \geq 1. \tag{4.12}$$

Now, we need a mapping rule between ψ_ν's and trellis outputs. We consider ψ_ν's as PSK constellation points and follow the same mapping rule as in TCM [98]. That is, we maximize the minimum Euclidean distance among ψ_ν's that are mapped to the branches with the same incoming/outgoing states by mapping ψ_ν to the trellis output ν.

Remark: We can further reduce the feedback overhead of TE-SPA. Note that we can rewrite $\mathbf{P}_k$ in (4.10) as

$$\mathbf{P}_k = e^{j\varphi_{k,1}} \operatorname{diag}\left(\left[1,\ldots,e^{j(\varphi_{k,T}-\varphi_{k,1})}\right] \otimes \mathbf{1}_L\right).$$

Let $\breve{\mathbf{P}}_k = \operatorname{diag}\left(\left[1,\ldots,e^{j(\varphi_{k,T}-\varphi_{k,1})}\right] \otimes \mathbf{1}_L\right)$. Then, the objective function in (4.11) can be rewritten as

$$\left\| \mathbf{h}_k - e^{j(\theta+\varphi_{k,1})}\breve{\mathbf{P}}_k \hat{\mathbf{h}}_{k-1} \right\|_2^2.$$

Thus, if we appropriately redesign Θ for the noncoherent search in (4.11), we can always fix the first entry of $\breve{\mathbf{P}}_k$ as 1 and skip (or fix) the first stage of the trellis search which gives a reduced feedback overhead.

4.3.2 Block-shifting

If we fix the block structure of the phase adjustment matrix $\mathbf{P}_k$, then the performance can quickly saturate because we cannot adjust the phase relation of the elements within each block. Moreover, since we fix the starting state of the trellis search, the first state transition suffers from using a restricted number of branches, e.g., only 4 branches with even trellis outputs are exploited for the first state transition in Fig. 4.3. These might not be serious problems for one-shot quantization as in

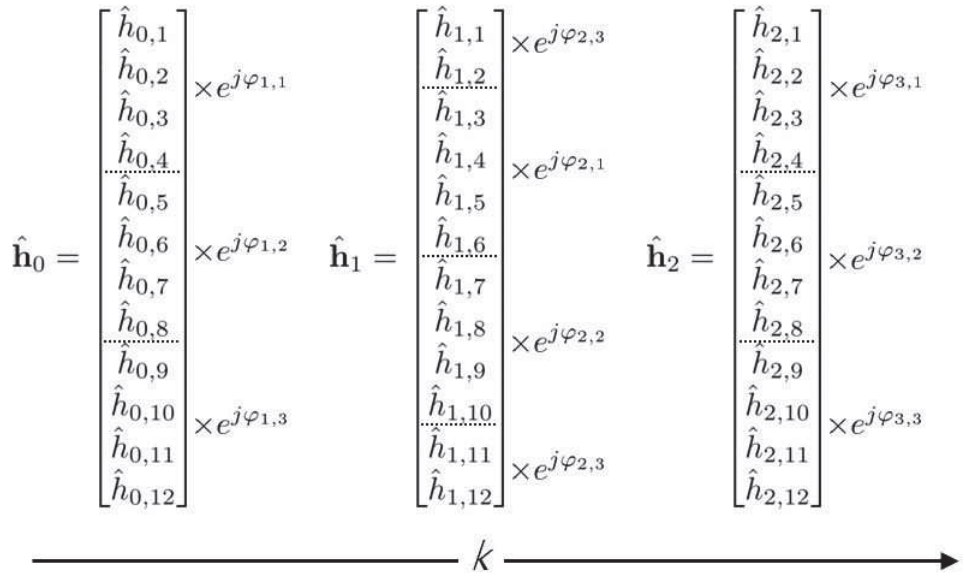


Fig. 4.4.: A conceptual explanation of TE-SPA with block-shifting with $M_t = 12$ and $L = 4$. $\hat{\mathbf{h}}_k$ is the result of multiplying $e^{j\varphi_{k,n}}$'s to $\hat{\mathbf{h}}_{k-1}$ in a block-wise manner.

TEC, but the loss could be accumulated in successive quantizations as in TE-SPA. Therefore, we adopt *block-shifting* to mitigate these problems.

Let $\mathbf{a}[m]_c$ and $\mathbf{A}[m]_c$ denote the left circularly shift of a vector $\mathbf{a}$ and diagonal entries of a matrix $\mathbf{A}$ of m elements, respectively. For example, if $\mathbf{a} = [1, 2, 3, 4, 5]$, then $\mathbf{a}[2]_c = [3, 4, 5, 1, 2]$. Using this notation, we rewrite the optimization problem in (4.11) as in (4.12). We interweave two consecutive blocks by circularly shifting $\frac{L}{2}$ elements in (4.12) to prevent the saturation effect.[5] After generating $\mathbf{P}_k$, the quantized CSI at time k is given as

$$\hat{\mathbf{h}}_k = \mathbf{P}_k \left[-\frac{L}{2}(k-1) \right]_c \hat{\mathbf{h}}_{k-1}.$$

The conceptual explanation of TE-SPA with block-shifting is shown in Fig. 4.4. Note that TEC is used for CSI quantization at $k = 0$. The proposed block shifting can adjust not only the phase relation among blocks but also that of elements within each

[5]To further improve performance, we can dynamically reassign the blocks of $\mathbf{P}_k$ instead of circularly shifting elements in time.

block in time. Moreover, the phase $\varphi_{k,1}$ from the first state transition is multiplied to the different blocks of $\hat{\mathbf{h}}_k$ depending on k, which prevents the accumulation of the loss caused by the first state transition.

4.3.3 Applying TE-SPA to spatially correlated channels

In massive MIMO systems, channels tend to be spatially correlated due to small antenna spacing. We can model spatially correlated channels as

$$\mathbf{h}[k] = \mathbf{R}^{\frac{1}{2}}\mathbf{h}_w[k]$$

where $\mathbf{R} = E[\mathbf{h}[k]\mathbf{h}^H[k]]$ is a spatial correlation matrix and $\mathbf{h}_w[k]$ is uncorrelated channel vector with i.i.d. complex Gaussian entries. Let $\mathbf{u}_1(\mathbf{R})$ denote the dominant eigenvector of $\mathbf{R}$. We assume $\mathbf{R}$ is perfectly known only at the receive side.

If the channels are highly correlated in space, the matrix $\mathbf{R}$ becomes ill conditioned, and $\mathbf{u}_1(\mathbf{R})$ and $\mathbf{h}[k]$ tend to be highly correlated. In this case, we can quantize $\mathbf{u}_1(\mathbf{R})$ using TEC and apply TE-SPA to quantize $\mathbf{h}[k]$ in each fading block of k based on the quantized version of $\mathbf{u}_1(\mathbf{R})$. Because $\mathbf{u}_1(\mathbf{R})$ is a long-term statistic and varies very slowly compared to $\mathbf{h}_w[k]$, the additional feedback overhead for $\mathbf{u}_1(\mathbf{R})$ would be negligible. Although this approach is based on one-step (instead of successive) phase adjustment, we keep the terminology TE-SPA to avoid any confusion.

4.4 Simulations and Discussions

We performed Monte-Carlo simulations using 10000 channel realizations to evaluate the proposed TEC and TE-SPA. We set $K_\theta = 16$ for $\Theta = \{\theta_1, \ldots, \theta_{K_\theta}\}$ to perform the noncoherent search of TEC and TE-SPA.

We first evaluate TEC in i.i.d. Rayleigh fading channels as $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{M_t})$. Because our focus is CSI quantization techniques, we use the average beamforming gain in dB scale that is defined as

$$10 \log_{10} \left(E[|\mathbf{h}^H \mathbf{c}_{\text{opt}}|^2] \right)$$

for a performance metric where the expectation is taken over $\mathbf{h}$. We set $L = 4$ to exploit VQ codebooks with dimension 4×1. Thus, TEC schemes with $B = 3/4$ and $B = 1/2$ bits per entry quantize 4 channel elements using 3 bits and 2 bits, respectively. In Fig. 4.5, we plot the average beamforming gain of TEC with the proposed codeword-to-branch mapping rule using different codebooks, e.g., trellis extended-Euclidean distance (TE-ED) refers to TEC using the Euclidean distance (ED) codebook defined in (4.7), according to the number of transmit antennas M_t. We also plot the average beamforming gain of NTCQ (denoted [101] in the figure) with $B = 1$ and that of RVQ with the same feedback overhead with TEC schemes for comparison purpose. Note that TE-LTE with $B = 1/2$ refers to TEC using only the first 8 among 16 codewords of LTE 4 transmit antennas codebook, which are the same as 8 DFT codewords. The total feedback overhead of each scheme is given as $B_{tot} = BM_t$.

As expected in Section 4.2.2, TE-ED using the ED codebook gives the best performance among the TEC schemes. The gain is more than 1 dB compared to TE-LTE when $B = 3/4$ and M_t is more than 64. TE-LTE suffers from practical constraints[6] on its codewords such as constant modulus (which causes the loss of norm information of channel elements) and finite alphabet properties. The conventional VQ codebook approach using RVQ is better than TEC schemes, but the plot of the RVQ codebook is based on the analytical approximation of $M_t \left(1 - 2^{-\frac{B_{tot}}{M_t - 1}}\right)$ [14] because it is infeasible to simulate the performance of the RVQ codebook with $B_{tot} = 16$ bits (which is

[6]The practical constraints lead to the decreased minimum Euclidean distance among the LTE codewords as well.

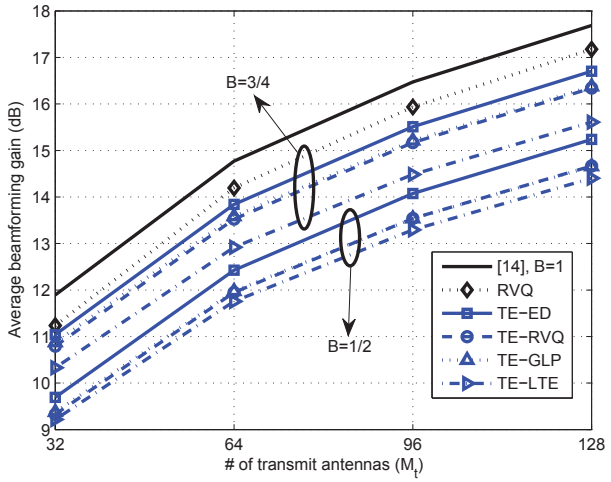


Fig. 4.5.: Average beamforming gain (dB) with M_t in i.i.d. Rayleigh fading channels. TE-'codebook name' refers to TEC using the specific codebook. $B_{tot} = BM_t$.

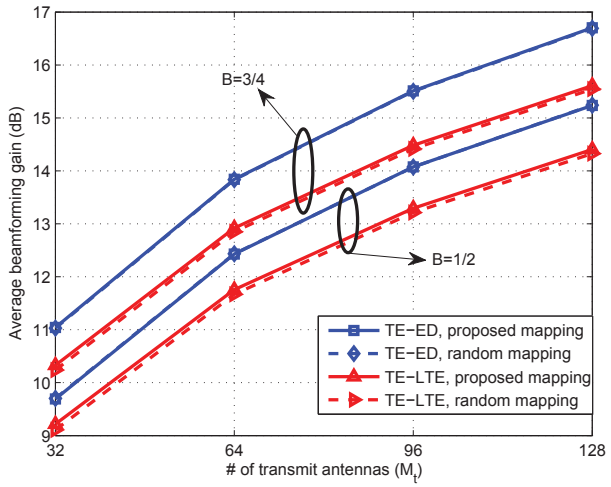


Fig. 4.6.: Average beamforming gain (dB) with M_t in i.i.d. Rayleigh fading channels. TEC schemes with the proposed codeword-to-branch mapping and random mapping are compared.

the case of $M_t = 32$ with $B = 1/2$) or more. NTCQ outperforms TEC with a much larger feedback overhead than the TEC schemes.[7]

[7]We did not compare TEC with [87] because the proposed scheme in [87] cannot even maintain a constant performance gap with the RVQ codebook.

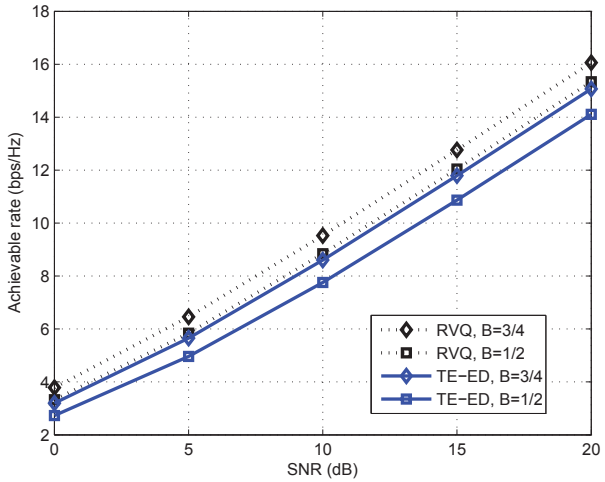


Fig. 4.7.: Achievable rate with SNR in i.i.d. Rayleigh fading channels with $M_t = 16$, $M_r = 2$, and $K = 2$. $B_{tot} = BM_t$.

We also compare the beamforming gains of the proposed codeword-to-branch mapping and a random mapping (per iteration) using TE-ED and TE-LTE in Fig. 4.6. Note that the proposed mapping has negligible impact on the average beamforming gain of TE-ED. The reason is that the Euclidean distance among codewords in the ED codebook is already far apart and the random mapping is also guaranteed to have a good Euclidean distance property. On the other hand, the proposed mapping achieves around 0.1 to 0.2 dB gain compared to the random mapping in TE-LTE. This shows that if we reuse preexisting VQ codebooks that are not optimized in the Euclidean distance, the proposed mapping can achieve additional gain with the same codebook.

Now, we evaluate TEC for a multiple receive antenna case. We set $M_t = 16$, $M_r = 2$, and the transmission rank as $K = 2$. The number of transmit antennas is *not too large* in this case because we want to compare TEC and the RVQ codebook with the same feedback overhead. With $M_t = 16$, TEC with $B = 3/4$ and $B = 1/2$ correspond to $B_{tot} = 12$ and $B_{tot} = 8$ bits, respectively. Denote the average achievable rate as $E[R_{ach}]$ where R_{ach} is defined in (4.8) and the expectation is taken over $\mathbf{H}$. Each entry of $\mathbf{H}$ is distributed with $\mathcal{CN}(0, 1)$. For the RVQ codebooks, the precoder

matrix is selected to maximize R_{ach} while $\mathbf{F}$ is generated as explained in Section 4.2.3 for TEC. We plot the average achievable rates of TE-ED and RVQ in Fig. 4.7 with SNR. The proposed TE-ED maintains a constant gap of around 1 bps/Hz loss compared to the RVQ codebook with the same feedback overhead for all SNR values. Considering the asymptotic optimality of the RVQ codebook in high rank transmission [46], the proposed TEC can achieve a good performance even in multiple receive antenna cases with feasible complexity.

In Fig. 4.8, we evaluate TE-SPA with $M_t = 64$ in temporally correlated Rayleigh fading channels which is shown in (4.9) with $\mathbf{g}[k] \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{M_t})$. We rely on Jakes' model for the temporal correlation coefficient [84] such that $\eta = J_0(2\pi f_D \tau)$ where $J_0(\cdot)$ is the zero-th order Bessel function, f_D is the maximum Doppler frequency, and τ is the channel instantiation interval. With practical system parameters of 2.5GHz carrier frequency, $\tau = 5ms$, and $3km/h$ user velocity, the temporal correlation coefficient is given as $\eta = 0.9881$. We do not consider any feedback delay in this simulation because it has been shown in the previous chapther that the impact of feedback delay is marginal.

At $k = 0$, channels are quantized with TE-ED using $B = 1/2$ bits per channel entry while channels are quantized using TE-SPA using B_{SPA} bits per entry when $k \geq 1$. As shown in the figure, the average beamforming gain increases with k due to reduced quantization error using TE-SPA even with lower feedback overhead of $B_{SPA} = 1/4$. All TE-SPA schemes outperform the RVQ codebook that does not consider temporal correlation of channels in quantization. The gain of using TE-SPA with block shifting is more than 1.6 dB when $B_{SPA} = 1/2$. Note that TE-SPA with block shifting gives far better performance than TE-SPA without block shifting because it can adjust the phase relation of the elements within each block and spread out the loss from the first state transition as explained in Section 4.3.

To evaluate TE-SPA in a more practical scenario, we perform simulations using the spatial channel model (SCM) [102] that is commonly adopted in standards such as 3GPP. In Fig. 4.9, we plot the average beamforming gain using the same simulation

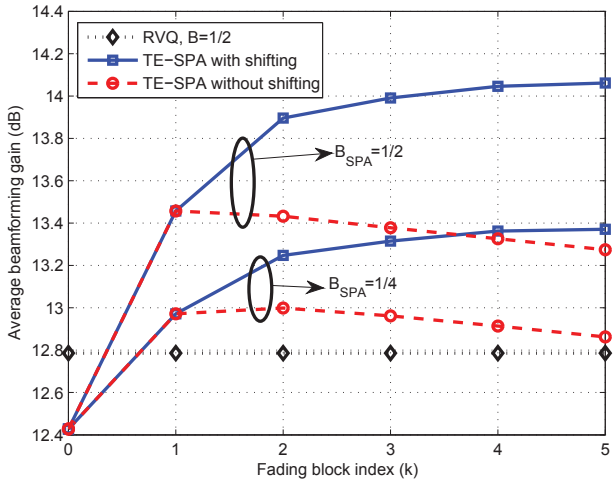


Fig. 4.8.: Average beamforming gain (dB) with k and $M_t = 64$ in temporally correlated channels. Channels are quantized using TE-ED with $B = 1/2$ bits per entry at $k = 0$ for TE-SPA schemes. Total feedback overhead of TE-SPA at $k \geq 1$ is $B_{tot} = B_{SPA} M_t$.

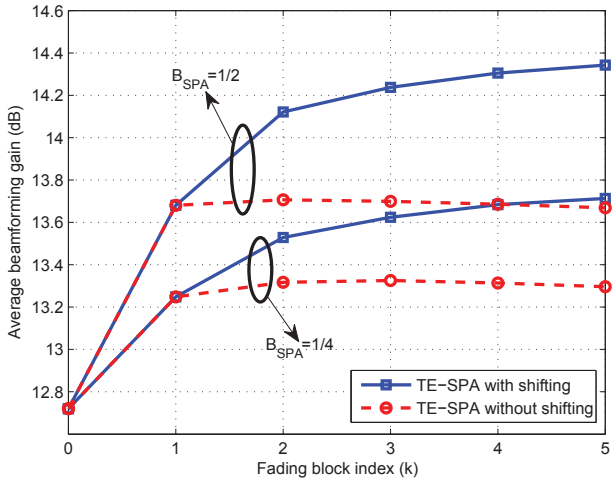


Fig. 4.9.: Average beamforming gain (dB) with k and $M_t = 64$ using an SCM channel model. Simulation setups are the same as in Fig. 4.8 with uniform linear array antennas with 0.5λ antenna spacing and 8 degrees angle spread.

setups as in Fig. 4.8 with uniform linear antenna array with 0.5λ antenna spacing and 8 degrees angle spread. As clearly shown in the figure, TE-SPA also works for the practical scenario.

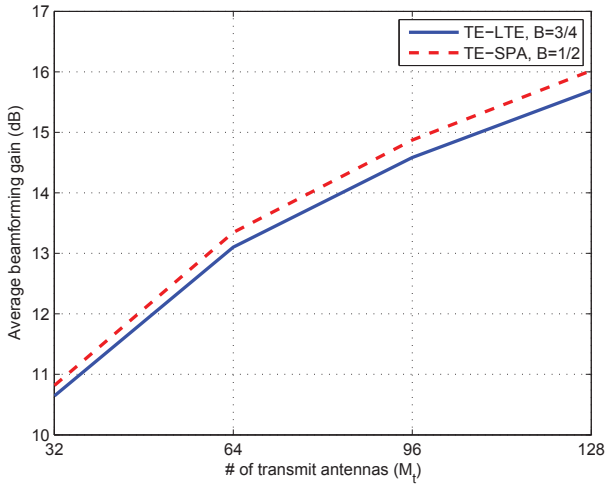


Fig. 4.10.: Average beamforming gain (dB) with M_t in spatially correlated Rayleigh fading channels. TE-LTE with $B = 3/4$ quantizes $\mathbf{h}[k]$ directly while TE-SPA refers to the scheme of which $\mathbf{u}_1(\mathbf{R})$ is quantized with TE-LTE with $B = 1/2$ and $\mathbf{h}[k]$ is quantized by TE-SPA with $B = 1/2$ based on the quantized $\mathbf{u}_1(\mathbf{R})$.

Finally, we evaluate TE-SPA in spatially correlated channels. We adopt the exponential model [103] for the spatial correlation matrix $\mathbf{R}$, which is defined as

$$[\mathbf{R}]_{\ell,r} = \begin{cases} (\alpha e^{j\vartheta})^{r-\ell}, & \ell \leq r \\ [\mathbf{R}]^*_{\ell,r}, & \ell > r \end{cases},$$

where $[\mathbf{R}]_{\ell,r}$ is the (ℓ, r)-th element of $\mathbf{R}$ and α and ϑ are the magnitude and the phase of the correlation coefficient, respectively. We set $\alpha = 0.9$ to mimic a high spatial correlation of a massive MIMO system while $\vartheta \in [0, 2\pi)$ is uniformly randomly generated in each channel realization.

As we can see in Fig. 4.10, TE-SPA is also beneficial for spatially correlated channels even with less feedback overhead than TE-LTE which quantizes $\mathbf{h}[k]$ directly. It is important to point out that TE-SPA for spatially correlated channels has additional feedback overhead, i.e., we adopt TE-LTE with $B = 1/2$ to quantize $\mathbf{u}_1(\mathbf{R})$. However, as stated in Section 4.3.3, $\mathbf{R}$ is a long-term statistic, and the feedback overhead for $\mathbf{u}_1(\mathbf{R})$ would be negligible in long-term sense.

Because TEC and TE-SPA both support various numbers of CSI quantization bits, the proposed techniques can easily allocate different numbers of feedback bits per user based on system requirements or channel conditions [104–106]. This is also a strong benefit for FDD massive MIMO systems of which the feedback overhead needs to be carefully optimized.

5. CODED DISTRIBUTED DIVERSITY: A NOVEL DISTRIBUTED RECEPTION TECHNIQUE FOR WIRELESS COMMUNICATION SYSTEMS

In this chapter, we propose the coded receive diversity technique to minimize the symbol error rate (SER) at the fusion center for distributed reception when the transmitter is equipped with a single transmit antenna. Coded receive diversity fully exploits the connection of the distributed reception problem with coding theory, and we are able to exploit efficient linear block codes such as simplex codes or first-order Reed-Muller codes that achieve the Griesmer bound with equality [107, 108]. We also develop novel shortened concatenated repetition-simplex (SCRS) codes for an arbitrary number of receive nodes and show that the SCRS codes are optimal with respect to the Griesmer bound in many practical scenarios. The SCRS codes are very easy to generate, meaning that we do not need to perform any kind of complex optimization to generate a SCRS code for an arbitrary number of receive nodes. We study the performance of coded receive diversity by analytically deriving the diversity gains attained by maximum likelihood (ML) and minimum Hamming distance detection. It is shown that numerical studies perfectly match with the analytical derivations.

5.1 Motivating Example and System Model

We first show a motivating example of this work and explain a general system model.

[0] ©[2014] IEEE. Reprinted, with permission, from J. Choi, D. J. Love, and p. Bidigare, "Coded Distributed Diversity: A Novel Distributed Reception Technique for Wireless Communication Systems," accepted to *IEEE Transactions on Signal Processing*, Dec. 2014.

5.1.1 Motivating example

Consider a SIMO system with three geographically separated receive nodes. The ith receive node operates with an input-output equation

$$y_i = \sqrt{\rho} h_i s + n_i, \quad i = 1, 2, 3$$

where ρ denotes the transmit SNR, $h_i \in \mathbb{C}$ is the channel from the transmitter to the node i, $s \in \mathcal{S} \subset \mathbb{C}$ is the transmitted signal selected from $\mathcal{S}$ with a uniform distribution, and n_i is complex additive white Gaussian noise (AWGN) distributed as $\mathcal{CN}(0, 1)$. We assume that the noise is spatially independent, i.e., each receive node experiences independent noise. We further assume that the channel collected across the distributed array $\mathbf{h} = [h_1\ h_2\ h_3]$ is a spatially uncorrelated channel with $\mathcal{CN}(0, 1)$ entries and the ith receive node has perfect knowledge of h_i and no knowledge of the other users' channels.

If the fusion center knows $\mathbf{h} = [h_1\ h_2\ h_3]$ and $\mathbf{y} = [y_1\ y_2\ y_3]$ perfectly, then as in a standard, centralized combining system, the fusion center can produce

$$\widetilde{y} = \frac{\mathbf{y}\mathbf{z}^*}{\mathbf{h}\mathbf{z}^*} \tag{5.1}$$

where $\mathbf{z} = \mathbf{h}/\|\mathbf{h}\|$ is the optimal linear combiner. The processed output $\widetilde{y}$ is used to detect the transmitted symbol s. However, the main focus of this work is the case when each receive node only can send a processed (or compressed) version of y_i, which we denote by u_i throughout this work, using a small number of bits per channel use to the fusion center, and the fusion center tries to decode the transmitted symbol based on $\{u_i\}_{i=1}^3$ along with possibly the knowledge of $\mathbf{h}$. We assume that each node can forward u_i without any error to the fusion center.[1] Many receive architectures in both commercial and military systems fall into this distributed reception scenario.

[1]This assumption is reasonable for many scenarios, e.g., 1) the receive nodes are connected with the fusion center through wired lines as in most CoMP, DAS, or radar systems, 2) the receive nodes and the fusion center are closely located with each other in wireless sensor networks.

In this example, we focus on the case when each node can pass only *one bit* for u_i per channel use to the fusion center; however, the transmitted symbol s is uniformly selected from a QPSK constellation

$$\mathcal{S} = \left\{ \sqrt{\frac{1}{2}}(1+j), \sqrt{\frac{1}{2}}(1-j), \sqrt{\frac{1}{2}}(-1+j), \sqrt{\frac{1}{2}}(-1-j) \right\}.$$

Thus, the fusion center needs to detect the transmitted symbol using 3 bits (1 bit per receive node) per channel use.

With a naive approach, this problem can be mapped into a binary hypothesis testing problem at each node. For example, nodes 1 and 3 detect the real component as

$$u_i = \begin{cases} 1 & \text{if} \quad \mathrm{Re}(h_i^* y_i) \geq 0 \\ 0 & \text{if} \quad \mathrm{Re}(h_i^* y_i) < 0 \end{cases}, \quad i = 1, 3$$

while node 2 detects the imaginary component as

$$u_2 = \begin{cases} 1 & \text{if} \quad \mathrm{Im}(h_2^* y_2) \geq 0 \\ 0 & \text{if} \quad \mathrm{Im}(h_2^* y_2) < 0 \end{cases},$$

and all nodes send their decisions $\{u_i\}_{i=1}^3$ to the fusion center. With an assumption that each node i has perfect knowledge of its channel h_i, the probability of incorrectly detecting the desired component at node i is given by

$$P_e^b(h_i, \rho) = Q\left(\sqrt{|h_i|^2 \rho}\right) \tag{5.2}$$

for $i = 1, 2, 3$. With full CSI knowledge at the fusion center, ML detection will give a probability of symbol error as

$$P_{e,unc}(\mathbf{h}, \rho) = 1 - \left(1 - \min_{i \in \{1,3\}} P_e^b(h_i, \rho)\right)\left(1 - P_e^b(h_2, \rho)\right).$$

Because $P_{e,unc}(\mathbf{h},\rho)$ is dominated by $P_e^b(h_2,\rho)$, the diversity order is given as

$$-\lim_{\rho\to\infty}\frac{\log(P_{e,unc}(\rho))}{\log\rho}=1 \tag{5.3}$$

with $P_{e,unc}(\rho)=E\left[P_{e,unc}(\mathbf{h},\rho)\right]$ where the expectation is taken over $\mathbf{h}$. This is a discouraging result because the distributed reception with three nodes has not provided any increase in diversity.

Note that a better solution exists. As in the previous approach, nodes 1 and 2 detect the real and imaginary component, respectively. However, node 3 now detects the product of the real and imaginary components such that

$$u_3=\begin{cases}1 & \text{if} \quad \mathrm{Re}(h_3^*y_3)\,\mathrm{Im}(h_3^*y_3)\geq 0\\ 0 & \text{if} \quad \mathrm{Re}(h_3^*y_3)\,\mathrm{Im}(h_3^*y_3)<0\end{cases}.$$

Nodes 1 and 2 have a probability of incorrect detection as in (5.2) while node 3 has a probability of detecting incorrectly given by

$$P_e^b(h_3,\rho)=2Q\left(\sqrt{|h_3|^2\rho}\right)\left(1-Q\left(\sqrt{|h_3|^2\rho}\right)\right).$$

If we let $P_{e,(i)}^b(h_{(i)},\rho)$ be the i-th largest probability of error among the nodes such that

$$P_{e,(1)}^b(h_{(1)},\rho)\geq P_{e,(2)}^b(h_{(2)},\rho)\geq P_{e,(3)}^b(h_{(3)},\rho),$$

a probability of error at the fusion center is given by

$$P_{e,code}(\mathbf{h},\rho)=1-\left(1-P_{e,(2)}^b(h_{(2)},\rho)\right)\left(1-P_{e,(3)}^b(h_{(3)},\rho)\right)$$

with ML detection. Then, the diversity order becomes

$$-\lim_{\rho\to\infty}\frac{\log(P_{e,code}(\rho))}{\log\rho}=2$$

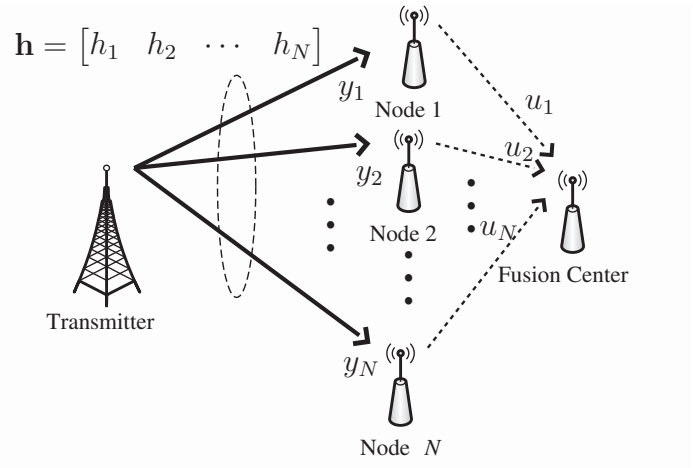


Fig. 5.1.: A conceptual figure of distributed reception.

where $P_{e,code}(\rho) = E\left[P_{e,code}(\mathbf{h}, \rho)\right]$ with the expectation taken over $\mathbf{h}$.

Thus, without increasing the number of bits sent from any of the nodes to the fusion center or changing the channel model, we have increased the diversity order from 1 to 2 by using a smart detection scheme at each node. More generally, this work aims to address the following question:

"*How should each receive node quantize y_i into a small number of bits to be sent to the fusion center when detecting M-ary modulation in distributed reception?*"

As we show later, this problem has intriguing ties to coding theory since the problem of designing the quantization map at the receive nodes can be regarded as encoding of data at the receive nodes.

5.1.2 System model

We consider a network consisting of a transmitter, a fusion center, and N geographically separated receive nodes. The conceptual figure of our system model is shown in Fig. 5.1. The received signal at the i-th node, y_i, is written as

$$y_i = \sqrt{\rho} h_i s + n_i, \quad i = 1, \cdots, N$$

where $s \in \mathcal{S}$ is the transmitted symbol from an M-ary constellation

$$\mathcal{S} = \{s_1, s_2, \ldots, s_M\} \subset \mathbb{C}.$$

We assume h_i and n_i have the same distributions as in the motivating example. We further assume that s is selected uniformly from $\mathcal{S}$ and satisfies $E[s] = 0$ and $E[|s|^2] = 1$. We define the conditional symbol error probability at the i-th receive node as

$$P_e(h_i, \rho) \triangleq E\left[Pr\left(\widehat{s}_i \neq s \mid s \text{ sent}, h_i, \rho\right)\right] \tag{5.4}$$

where the expectation is taken over s and $\widehat{s}_i$ is the estimated symbol at the i-th receive node defined as

$$\widehat{s}_i = \underset{t \in \mathcal{S}}{\operatorname{argmin}} \|y_i - \sqrt{\rho} h_i t\|^2. \tag{5.5}$$

Note that the majority of the distributed reception work has been dedicated for binary modulation schemes, i.e., binary hypothesis testing in AWGN channels without fading. In this case, $s \in \mathcal{S} = \{s_1, s_2\}$, and each node can make a hard decision on the transmitted symbol. We consider generalized distributed reception in this work such that the transmitter can send the symbol from an arbitrary M-ary constellation. To make the system practical, we assume that the bandwidth between the receive node and the fusion center is limited. Thus, each receive node only can forward a quantized version of the estimate $\widehat{s}_i$ (which is represented using multiple bits) to the fusion center.

5.2 Coded Receive Diversity and Diversity Order

We first explain the general concept of coded receive diversity and then discuss the symbol detection schemes using the quantized node information. We finish this section with diversity order analyses with respect to the decoding schemes.

5.2.1 General concept of coded receive diversity technique

Note that the M-ary constellation $\mathcal{S}$ can be represented with a $\log_2(M)$-bit message that we denote as $\mathbf{b} = [b_1\ b_2 \cdots b_{\log_2(M)}]$. Each node quantizes its received signal y_i into a B-bit vector[2] $u_i \in GF(2^B)$. We assume

$$B \leq \log_2(M)$$

to limit the overhead needed for the distributed decisions. This gives rise to the concept of a *compression ratio* that is defined as

$$K \triangleq \frac{\log_2(M)}{B}$$

which satisfies $K \geq 1$. We assume K is an integer value throughout this work. We let $\mathbf{a} = [a_1\ a_2\ \cdots a_K]$ be the vectorized version of $\mathbf{b}$ with entries in $GF(2^B)$. There are multiple ways of converting $\mathbf{b}$ into $\mathbf{a}$ using different primitive polynomials of $GF(2^B)$; however, using a specific primitive polynomial does not affect average performance.

An example of the system parameters is $M = 16$ (e.g., 16-quadrature amplitude modulation (QAM)) which gives

$$\mathbf{b} \in \left\{ \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}, \cdots, \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} \right\},$$

and QPSK signaling from each receive node to the fusion center, resulting in $B = 2$. The primitive polynomial with coefficients in $GF(2)$ used to describe $GF(4)$ is $x^2 + x + 1$.

Commonly, detectors for M-ary constellations are designed using non-overlapping decision regions. Denote the decision regions by $\{\mathcal{W}_1, \ldots, \mathcal{W}_M\}$ such that

$$\mathcal{W}_1 \bigcup \cdots \bigcup \mathcal{W}_M = \mathbb{C}.$$

[2]We let $GF(q)^m$ denote the m-dimensional vector of elements in $GF(q)$. This is different representation than $GF(q^m)$ which denotes the finite field of order q^m.

Using the decision regions, the detection problem at receive node i can be formulated as $\widehat{s}_i = s_{m_0}$ with

$$m_0 = \underset{1 \le m \le M}{\operatorname{argmax}} \mathbb{1}\left(y_i \in \mathcal{W}_m\right)$$

with $\mathbb{1}(\cdot)$ denoting the indicator function which returns 1 if the argument is true.

In our problem, however, we assume that the number of decision regions at each receive node is smaller than M, i.e., the compression ratio is constrained as $K \ge 1$. Let the non-overlapping decision regions at node i be $\left\{\mathcal{D}_{i,1}, \ldots, \mathcal{D}_{i,2^B}\right\}$ such that the union of the regions spans the complex plane. The distributed reception problems can be succinctly stated as determining the sets of decision regions $\left\{\mathcal{D}_{i,1}, \ldots, \mathcal{D}_{i,2^B}\right\}$ for $i = 1, \ldots, N$ to minimize the probability of symbol detection error at the fusion center. As shown in the motivating example of Section 5.1, this problem is nontrivial. We show how well-developed coding techniques can be used to design the sets of decision regions.

Because of the constraint on the compression ratio $K \ge 1$, we assume the decision regions $\left\{\mathcal{D}_{i,1}, \ldots, \mathcal{D}_{i,2^B}\right\}$ of node i are constructed by certain unions of the linear combinations of the constellation decision regions $\{\mathcal{W}_1, \ldots, \mathcal{W}_M\}$. To do this, we formulate the problem using finite field notations.

To simplify the notation, let $\mathbf{b}$ denote the bit representation of the transmitted symbol s. Suppose that the node i first estimates the transmitted symbol from the received signal y_i as in (5.5) to generate a $\log_2(M)$-bit vector $\widehat{\mathbf{b}}_i$. If the node i detects s correctly, then we have $\widehat{\mathbf{b}}_i = \mathbf{b}$. Note that $\widehat{\mathbf{b}}_i$ can be represented with a K-entry vector $\widehat{\mathbf{a}}_i$ with entries in $GF(2^B)$. The node i then generates $u_i = f_i\left(\widehat{\mathbf{a}}_i\right)$ using a function

$$f_i : \ GF(2^B)^K \to GF(2^B)$$

and sends u_i to the fusion center.

If each node received $\mathbf{a}$ (or equivalently $\mathbf{b}$) without any error, this problem can be formulated as a coding problem. The K-dimensional *message* $\mathbf{a}$ is transformed to an N-dimensional vector codeword $\mathbf{u} = [u_1 \ \cdots \ u_N]$ with entries in $GF(2^B)$. In

the distributed reception case, each node is coding on noisy data, and the vector $\mathbf{u}$ is corrupted with *noise* corresponding to reception error at each node. Despite this, the goal in distributed reception is very similar to code design in coding theory. We must find a coding technique that minimizes the decoding error of the transmitted symbol at the fusion center.

Similar to the coding problem, we focus on the creation of an M vector codeword set $\{\mathbf{u}[1], \ldots, \mathbf{u}[M]\}$ where each codeword $\mathbf{u}[k]$ corresponds to a constellation point $s_k \in \mathcal{S}$. Further, we focus on linear block codes to enable efficient encoding. This means that the function f_i is explicitly given as

$$u_i = f_i\left(\widehat{\mathbf{a}}_i\right) = \widehat{\mathbf{a}}_i \mathbf{g}_i^T \tag{5.6}$$

where $\mathbf{g}_i \in GF(2^B)^K$. We can collect everything together in vector form such that

$$\mathbf{u} = \mathbf{c}(s) + \mathbf{v}$$

where $\mathbf{c}(s) \in GF(2^B)^N$ denotes the distributed detection bits if all nodes make the correct bit decisions when s is transmitted and $\mathbf{v} \in GF(2^B)^N$ represents *noise* caused by reception error at each node. Due to the linear structure, a generator matrix $\mathbf{G} \in GF(2^B)^{K \times N}$ can be given as

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_N \end{bmatrix}^T$$

and

$$\mathbf{c}(s) = \mathbf{a}\mathbf{G}.$$

This generates a code for the constellation points $\mathcal{S}$ as

$$\mathcal{C} = \{\mathbf{c}(s) : s \in \mathcal{S}\}.$$

We define a matrix $\mathbf{C}$ from the set $\mathcal{C}$ as

$$\mathbf{C} = \Big[\mathbf{c}(s_1) \quad \mathbf{c}(s_2) \quad \cdots \quad \mathbf{c}(s_M)\Big]^T \tag{5.7}$$

which will be useful for explaining the difference between the proposed coded receive diversity technique and the scheme from [69] in Section 5.4. We also define the minimum Hamming distance of the code

$$d_{\min}(\mathcal{C}) \triangleq \min_{s,s' \in \mathcal{S}: s \neq s'} d_H(\mathbf{c}(s), \mathbf{c}(s'))$$

where $d_H(\cdot, \cdot)$ denotes the Hamming distance metric.

To explain the procedure of the coded receive diversity technique in words, each receive node i processes $\widehat{\mathbf{a}}_i$ (which is nothing but a hard-detected version of y_i) with the i-th column of $\mathbf{G}$ as in (5.6). With $\mathbf{u} = [u_1 \ \cdots \ u_N]$ from all N receive nodes, the fusion center detects the transmitted symbol by using decoding schemes explained next.

5.2.2 Decoding schemes at fusion center

For practical reasons, we assume that the i-th node has knowledge *only* of $\mathcal{S}$, y_i, and h_i. Each node passes u_i to the fusion center, and the fusion center tries to detect the transmitted symbol $s \in \mathcal{S}$ using $\mathbf{u} = [u_1 \ \cdots \ u_N]$. We consider the cases when the fusion center has knowledge of $\mathbf{h}$ and lacks knowledge of $\mathbf{h}$. Note that the fusion center does not have full access to $\mathbf{y}$ in our scenario, which prevents the use of the linear combiner $\mathbf{z} = \mathbf{h}/\|\mathbf{h}\|$ to estimate $\widetilde{y}$ as in (5.1) even with full knowledge of $\mathbf{h}$. We discuss three different decoding schemes for distributed reception over fading channels:

1) ML decoding with full CSI: If the fusion center has full access to $\mathbf{h}$, it computes

$$\begin{aligned}\widehat{s} &= \underset{t\in\mathcal{S}}{\operatorname{argmax}}\, Pr\left(u_1,\ldots,u_N \mid t,\mathbf{h},\rho\right)\\ &= \underset{t\in\mathcal{S}}{\operatorname{argmax}} \prod_{i=1}^{N} Pr\left(u_i \mid t,h_i,\rho\right)\end{aligned}$$

where $Pr(\cdot)$ denotes the probability that is computed as

$$Pr\left(u_i \mid t,h_i,\rho\right) = \frac{1}{\pi}\int_{\mathcal{D}_{i,u_i}} e^{-\left|y_i-\sqrt{\rho}h_i t\right|^2} dy_i$$

with $\mathcal{D}_{i,u_i}$ denoting the decision region of node i corresponding to u_i.[3]

2) Selected subset ML decoding with full CSI: If the number of receive nodes N is very large, the complexity of ML decoding can be excessive. With our coded receive diversity framework, however, we can reduce decoding complexity significantly while obtaining comparable performance with ML decoding.

First, we assume that every L-th receive node shares the same processing rule, i.e.,

$$\mathbf{g}_\ell = \mathbf{g}_{\ell+L} = \cdots = \mathbf{g}_{\ell+\lfloor\frac{N-\ell}{L}\rfloor L}, \quad \ell\in\{1,\ldots,L\}$$

and $\mathbf{g}_\ell \neq \mathbf{g}_k$ if $\ell\neq k$ for $\ell,k\in\{1,\ldots,L\}$. We define the equivalence class of the node ℓ as

$$[\ell]_L = \{i\in\{1,\ldots,N\} :\ i \bmod L = \ell \bmod L\} \tag{5.8}$$

which denotes the set of nodes that share the same processing rule with node $\ell\in\{1,\ldots,L\}$. Because the fusion center has full access to $\mathbf{h}$, it can select the node ℓ_0 among $[\ell]_L$ as

$$\ell_0 = \underset{k\in[\ell]_L}{\operatorname{argmax}} |h_k|^2 \tag{5.9}$$

[3]In practice, we can generate empirical probabilities of $Pr\left(u_i \mid s,h_i\right)$ in advance to perform ML decoding with full CSI.

to perform ML decoding using bits from the L selected receive nodes. This selected subset ML decoding is appropriate for one of our code designs explained in Section 5.3.2.

3) Minimum Hamming distance decoding without CSI: If the fusion center does not have any knowledge of $\mathbf{h}$, it needs to rely on the simple Hamming distance decoding. The fusion center then tries to detect the transmitted symbol as

$$\widehat{s} = \underset{t \in \mathcal{S}}{\operatorname{argmin}}\, d_H(\mathbf{c}(t), \mathbf{u})$$

where $\mathbf{c}(t)$ is the vector that would be sent from the all N receive nodes if the transmitted symbol was perfectly decoded at each node and $\mathbf{u} = [u_1 \ \cdots \ u_N]$.

5.2.3 Diversity analysis

We present the diversity analyses of the proposed coded receive diversity technique with three different decoding schemes explained in the previous section in the following.

Lemma 5.2.1 *Using a codeword set $\mathcal{C} = \{\mathbf{c}(s_1), \ldots, \mathbf{c}(s_M)\}$, a coded receive diversity system achieves a diversity order of $d_{\min}(\mathcal{C})$ using ML decoding.*

Proof The diversity order of the probability of error $P_{e,code,ML}(\rho)$ can be obtained by analyzing the worst case pairwise error probability $\max\limits_{s \neq s'} Pr(s \rightarrow s')$. Instead of working with the optimal decoder directly, we first upper bound the pairwise error probability with a suboptimal decoder. The considered suboptimal detector chooses

$$\widehat{s}_{\text{sub}} = \underset{t \in \mathcal{S}}{\operatorname{argmax}} \prod_{i=1}^{N} Pr\left(u_i \mid c_i(t), h_i, \rho\right) \tag{5.10}$$

$$Pr\left(\sum_{i:c_i(s)\neq c_i(s')}\left(\log\frac{Pr\left(u_i \mid c_i(s'),h_i,\rho\right)}{Pr\left(u_i \mid c_i(s),h_i,\rho\right)}\right)\geq 0\middle| \, s \text{ sent},\mathbf{h},\rho\right)\leq Pr_{\text{sub}}(s\to s' \mid h_{i_0},\rho). \tag{5.12}$$

where $c_i(t)$ is the i-th entry of $\mathbf{c}(t)$. In the event of a tie, it is broken arbitrarily. If the transmitted symbol is s, using the coding framework with the decoder in (5.10), the pairwise error $s \to s'$ occurs when

$$\prod_{i=1}^{N} Pr\left(u_i \mid c_i(s'),h_i,\rho\right)-\prod_{i=1}^{N} Pr\left(u_i \mid c_i(s),h_i,\rho\right)\geq 0$$

which can be converted to

$$\sum_{i:c_i(s)\neq c_i(s')}\left(\log\frac{Pr\left(u_i \mid c_i(s'),h_i,\rho\right)}{Pr\left(u_i \mid c_i(s),h_i,\rho\right)}\right)\geq 0. \tag{5.11}$$

Define

$$Pr_{\text{sub}}(s\to s' \mid h_i,\rho)=Pr\left(\log\frac{Pr\left(u_i \mid c_i(s'),h_i,\rho\right)}{Pr\left(u_i \mid c_i(s),h_i,\rho\right)}\geq 0\right)$$

as a pairwise error probability at node i with given h_i and ρ using the decoder in (5.10), and let the index i_0 be

$$i_0=\underset{i:c_i(s)\neq c_i(s')}{\operatorname{argmax}}\ |h_i|^2.$$

Then, we can bound the probability of (5.11) as in (5.12) by only considering the pairwise error probability of the i_0-th receive node which has the strongest channel gain among $\{i : c_i(s) \neq c_i(s')\}$. The right-hand side of (5.12) is a pairwise error probability of single-input single-output communication between the transmitter and

$$
\begin{aligned}
&Pr(s \to s' \mid \mathbf{h}, \rho) \\
&= Pr\left(\sum_{i=1}^{N} \log\left(\frac{Pr(u_i \mid s', h_i, \rho)}{Pr(u_i \mid s, h_i, \rho)} \right) \geq 0 \middle| \, s \text{ sent}, \mathbf{h}, \rho \right) \\
&\overset{(a)}{\geq} Pr\left(\sum_{i:c_i(s)\neq c_i(s')} \log\left(\frac{Pr(y_i \mid s', h_i, \rho)}{Pr(y_i \mid s, h_i, \rho)} \right) \right. \\
&\qquad\qquad \left. \geq \sum_{i:c_i(s)=c_i(s')} \log\left(\frac{Pr(u_i \mid s, h_i, \rho)}{Pr(u_i \mid s', h_i, \rho)} \right) \middle| \, s \text{ sent}, \mathbf{h}, \rho \right) \qquad (5.13) \\
&\overset{(b)}{\geq} Pr\left(\sum_{i:c_i(s)\neq c_i(s')} \log\left(\frac{Pr(y_i \mid s', h_i, \rho)}{Pr(y_i \mid s, h_i, \rho)} \right) \geq \alpha_0 \middle| \, s \text{ sent}, \mathbf{h}, \rho \right) \\
&\overset{(c)}{=} Pr\left(\sum_{i:c_i(s)\neq c_i(s')} \left(|y_i - \sqrt{\rho} h_i s|^2 - |y_i - \sqrt{\rho} h_i s'|^2 \right) \geq \alpha_0 \middle| \, s \text{ sent}, \mathbf{h}, \rho \right) \qquad (5.14)
\end{aligned}
$$

---

the i_0-th receive node. Taking the expectation over h_{i_0}, the pairwise error probability is bounded as

$$Pr(s \to s') \leq E\left[(Pr_{\text{sub}}(s \to s' \mid h_{i_0}, \rho) \right].$$

Note that selecting the i_0-th receive node is similar to antenna selection. Using the antenna selection diversity result in [109, 110], we have

$$-\lim_{\rho\to\infty} \frac{\log\left(E\left[Pr_{\text{sub}}(s \to s' \mid h_{i_0}, \rho)\right]\right)}{\log \rho} = d_H(\mathbf{c}(s), \mathbf{c}(s')),$$

which results in

$$-\lim_{\rho\to\infty} \frac{\log\left(\max\limits_{s\neq s'} Pr(s \to s') \right)}{\log \rho} \geq d_{\min}(\mathcal{C}).$$

To obtain a lower bound on $Pr(s \to s')$ (or an upper bound of the diversity order), we can use (5.14) where (a) is due to the fact that u_i is a degraded version of y_i, (b) is from the maximization over $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \dots u_N \end{bmatrix}$ with[4]

$$\alpha_0 = \max_{\mathbf{u}} \sum_{i:c_i(s)=c_i(s')} \log\left(\frac{Pr(u_i \mid s, h_i, \rho)}{Pr(u_i \mid s', h_i, \rho)}\right),$$

and (c) comes from the variable substitution

$$Pr(y_i \mid s, h_i, \rho) = \frac{1}{\pi} e^{-|y_i - \sqrt{\rho} h_i s|^2}.$$

Note that (5.14) is a decoding method using maximum ratio combining (MRC) over $d_H(\mathbf{c}(s), \mathbf{c}(s'))$ receive nodes with the threshold α_0 when the fusion center has perfect knowledge of $\{y_i\}_{i:c_i(s)\neq c_i(s')}$. Because the diversity order is derived in the high SNR regime, the threshold α_0 has no impact on the diversity order of MRC since $\alpha_0 \to 0$ as $\rho \to \infty$. Thus, taking the maximum over any pair $s \neq s'$, the diversity order of ML decoding is upper bounded by $d_{\min}(\mathcal{C})$, which finishes the proof. ■

Lemma 5.2.2 *If L, the number of distinctive processing rules, divides the total number of receive nodes N, a coded receive diversity system using a codeword set $\mathcal{C} = \{\mathbf{c}(s_1), \dots, \mathbf{c}(s_M)\}$ achieves a diversity order of $d_{\min}(\mathcal{C})$ using selected subset ML decoding.*

Proof The diversity order of selected subset ML is upper bounded by that of ML decoding, i.e., $d_{\min}(\mathcal{C})$. To obtain the lower bound, we again rely on the pairwise error probability. First, we let $\mathbf{G}_{[1:L]}$ be the generating matrix consists of the first L columns of $\mathbf{G}$ and $\mathcal{C}_L = \{\mathbf{c}_L(s_1), \dots, \mathbf{c}_L(s_M)\}$ be the resulting code from $\mathbf{G}_{[1:L]}$. Denote $d_{\min}(\mathcal{C}_L)$ the minimum Hamming distance of $\mathcal{C}_L$. The pairwise error probability

[4] α_0 is a function of given variables h_i, s, and ρ. We intentionally neglect this when defining α_0 for brevity.

at a group $[\ell]_L$ (defined in (5.8)) that shares the same processing rule with the ℓ-th node is given as

$$Pr(s \to s' \mid \mathbf{h}_{[\ell]_L}, \rho) = Pr(s \to s' \mid h_{\ell_0}, \rho)$$

where $\mathbf{h}_{[\ell]_L}$ is the channel vector consists of channel elements of $[\ell]_L$ and ℓ_0 is the selected node index defined in (5.9). Using the same step as the proof of Lemma 5.2.1, the diversity order lower bound is given as

$$\frac{N}{L} d_{\min}(\mathcal{C}_L) = d_{\min}(\mathcal{C}),$$

where N/L is due to the selection diversity from $[\ell]_L$. ■

Remark: In the general N case, the diversity order of selected subset ML would lie between $\left\lfloor \frac{N}{L} \right\rfloor d_{\min}(\mathcal{C}_L)$ and $\frac{N}{L} d_{\min}(\mathcal{C}_L)$.

Lemma 5.2.3 *Using a codeword set $\mathcal{C} = \{\mathbf{c}(s_1), \ldots, \mathbf{c}(s_M)\}$, a coded receive diversity system achieves a diversity order of $\lceil d_{\min}(\mathcal{C})/2 \rceil$ using minimum Hamming distance decoding.*

Proof First, we let $p = E\left[P_e(h_i, \rho)\right]$ to simplify notation where $P_e(h_i, \rho)$ is the conditional symbol error probability defined in (5.4). Note that $0 \le p \le 1$. Now, consider again the pairwise error probability. If the number of nodes with incorrect receptions is $\lceil d_{\min}(\mathcal{C})/2 \rceil$ or more, the error pattern will fall outside of the Hamming sphere of radius $\lfloor (d_{\min}(\mathcal{C}) - 1)/2 \rfloor$ centered at the correct codeword. Thus, we have

$$\begin{aligned}
P_{e,code,H}(\rho) &\le \sum_{i=\lceil d_{\min}(\mathcal{C})/2 \rceil}^{N} \binom{N}{i} p^i (1-p)^{N-i} \\
&\le N! \sum_{i=\lceil d_{\min}(\mathcal{C})/2 \rceil}^{N} p^i (1-p)^{N-i} \\
&\le N! \sum_{i=\lceil d_{\min}(\mathcal{C})/2 \rceil}^{N} p^i \\
&\le (N+1)!\, p^{\lceil d_{\min}(\mathcal{C})/2 \rceil}
\end{aligned}$$

and

$$P_{e,code,H}(\rho) \geq p^{\lceil d_{\min}(\mathcal{C})/2 \rceil}(1-p)^{N-\lceil d_{\min}(\mathcal{C})/2 \rceil}.$$

Using the fact that $p \to 0$ as $\rho \to \infty$, the diversity order is bounded as

$$\begin{aligned} -\lim_{\rho\to\infty} \frac{\log(P_{e,code,H}(\rho))}{\log \rho} &\geq -\lim_{\rho\to\infty} \frac{\log\left((N+1)!\, p^{\lceil d_{\min}(\mathcal{C})/2 \rceil}\right)}{\log \rho} \\ &= \lceil d_{\min}(\mathcal{C})/2 \rceil \end{aligned}$$

and

$$\begin{aligned} -\lim_{\rho\to\infty} \frac{\log(P_{e,code,H}(\rho))}{\log \rho} &\leq -\lim_{\rho\to\infty} \frac{\log\left(p^{\lceil d_{\min}(\mathcal{C})/2 \rceil}(1-p)^{N-\lceil d_{\min}(\mathcal{C})/2 \rceil}\right)}{\log \rho} \\ &= \lceil d_{\min}(\mathcal{C})/2 \rceil \end{aligned}$$

which finishes the proof. ■

Note that the diversity order is closely related to the symbol error probability. Although it is hard to derive the symbol error probability of the proposed coded receive diversity technique in general, using the upper bound of the symbol error probability of any linear block code derived in [84], we can upper bound the symbol error probability of the proposed technique as

$$P_e(\rho) \leq (M-1)\triangle^{d_{\min}(\mathcal{C})}$$

where $\triangle$ is a constant that is a function of $\mathcal{S}$. This upper bound would be loose in general; however, the numerical studies in Section 5.4 show that the symbol error probability of the proposed technique has the slope of $d_{\min}(\mathcal{C})$.

5.3 Code Design and Performance Implications

Lemmas 5.2.1, 5.2.2, and 5.2.3 all show that the coding structure across the receive nodes dictates system performance, and it is better to have as large of a minimum

Hamming distance as possible for a code $\mathcal{C}$ (or $\mathcal{C}_L$ for selected subset ML). The coding structure is heavily dependent on the number of nodes N and the compression ratio K. In this section we aim to clarify this relationship and look at some simple codes that can be employed.

5.3.1 Code bounds

Most common technique used to understand codes in coding theory employs metric ball bounds, e.g., the sphere packing bound and Gilbert-Varshamov bound which are most useful in understanding code properties when K grows with N, particularly when K/N converges to a fixed value as $N \to \infty$. However, we are more concerned with the case where K is fixed and does not scale with N. Moreover, we are interested in the case when K is relatively small and N is not extremely large.

The most applicable bound to this situation is the Griesmer bound [107,108]. The Griesmer bound shows that the smallest N of a code $\mathcal{C}$ that can achieve a minimum Hamming distance of $d_{\min}(\mathcal{C})$ must satisfy

$$N \geq \sum_{i=0}^{K-1} \left\lceil \frac{d_{\min}(\mathcal{C})}{2^{iB}} \right\rceil$$

By removing the ceiling function and rearranging terms, we have an upper bound of $d_{\min}(\mathcal{C})$ as

$$\frac{N2^{(K-1)B}}{1 + 2^B + \cdots + 2^{(K-1)B}} \geq d_{\min}(\mathcal{C}).$$

We are interested in codes that can achieve this upper bound.

5.3.2 Code selection

The Griesmer bound gives us insight into code choice for many different scenarios (e.g., see [108]). The following codes are a few cases that achieve the Griesmer bound with equality.

1) Simplex Codes

The *simplex code*, which is the dual code of the Hamming code, can achieve this minimum Hamming distance. If we denote $GF(2^B)$ as $\{0, 1, 2, \cdots, q-1\}$, a generator matrix of the 2^B-ary simplex code is given as

$$\mathbf{G}_{simplex} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \cdots & 1 & 1 \\ 0 & 0 & \cdots & 0 & & q-1 & q-1 \\ \vdots & \vdots & & \vdots & \iddots & \vdots & \vdots \\ 0 & 1 & \cdots & 1 & & q-1 & q-1 \\ 1 & 0 & \cdots & q-1 & \cdots & q-2 & q-1 \end{bmatrix}. \tag{5.15}$$

In words, the generator matrix $\mathbf{G}_{simplex}$ is the $K \times (2^{KB}-1)/(2^B-1)$ matrix with columns chosen to correspond to all non-zero vectors in $GF(2^B)^K$ with first non-zero entry fixed to one. Note that if $d_{\min}(\mathcal{C}) = 2^{(K-1)B}$, then the number of receive nodes becomes

$$\begin{aligned} N &= \sum_{i=0}^{K-1} \frac{2^{(K-1)B}}{2^{iB}} \\ &= 1 + 2^B + \cdots + 2^{(K-1)B} \\ &= \frac{2^{KB}-1}{2^B-1}. \end{aligned}$$

2) First-Order Reed-Muller Codes

A first-order Reed-Muller code exists for $N = 2^{(K-1)B}$ with minimum distance $d_{\min}(\mathcal{C}) = 2^{(K-2)B}(2^B-1)$. It also achieves the Griesmer bound with equality [108].

3) Shortened Concatenated Repetition-Simplex (SCRS) Codes

A simple approach to code design when $N \neq 2^{(K-1)B}$ and $N \neq (2^{KB}-1)/(2^B-1)$ is to shorten a concatenated code consisting of a shorter simplex code and a repetition code. In this case, the outer code is the simplex code and the inner code is a repetition code.

To construct our code, we first define two variables

$$N_{out} = \frac{(2^{KB}-1)}{(2^{B}-1)}, \quad N_{in} = \left\lceil \frac{N(2^{B}-1)}{(2^{KB}-1)} \right\rceil,$$

and construct the $K \times N_{out}N_{in}$, generator matrix

$$\begin{aligned} \mathbf{G}_{concat} &= \mathbf{1}_{1\times N_{in}} \otimes \mathbf{G}_{simplex} \\ &= [\mathbf{G}_{simplex}\ \mathbf{G}_{simplex}\ \cdots\ \mathbf{G}_{simplex}] \end{aligned}$$

where $\mathbf{G}_{simplex}$ is the $K \times N_{out}$ simplex code's generator matrix given in (5.15), $\mathbf{1}_{1\times N_{in}}$ is the N_{in} row vector of all ones (i.e., $\mathbf{1}_{1\times N_{in}} = [1\ 1 \cdots 1]$), and $\otimes$ represents the Kronecker product. If we let

$$N^{'} = N_{out}N_{in} - N,$$

then the extended code uses the shortened generator matrix given by

$$\mathbf{G}_{extend} = \mathbf{G}_{concat} \begin{bmatrix} \mathbf{I}_N \\ \mathbf{0}_{N^{'}\times N} \end{bmatrix}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix and $\mathbf{0}_{N^{'}\times N}$ is the $N^{'} \times N$ all zero matrix.

5.3.3 SCRS codes analyses

A SCRS code achieves a minimum distance of

$$d_{\min}(\mathcal{C}) \geq \left\lfloor \frac{N(2^{B}-1)}{(2^{KB}-1)} \right\rfloor 2^{(K-1)B}. \tag{5.16}$$

When $N = K(2^{KB}-1)/(2^B-1)$, the code is optimal with respect to the Griesmer bound because

$$\begin{aligned}\sum_{i=0}^{K-1}\left\lceil\frac{N(2^B-1)2^{(K-1)B}}{2^{iB}(2^{KB}-1)}\right\rceil &= \sum_{i=0}^{K-1}K\frac{2^{(K-1)B}}{2^{iB}}\\ &= K\sum_{i=0}^{K-1}2^{iB}\\ &= K\left(\frac{2^{KB}-1}{2^B-1}\right)\\ &= N.\end{aligned}$$

For arbitrary N, the following lemma states that the SCRS codes are optimal in terms of the Griesmer bound when $K = 2$.

Lemma 5.3.1 *The length N SCRS code with $K = 2$ formed from concatenating the 2^B-ary simplex code and repetition code has the following properties:*
1) The minimum Hamming distance becomes

$$d_{\min}(\mathcal{C}) = \alpha 2^B + r - 1$$

where $\alpha = \lfloor N/(2^B+1)\rfloor$, $N = \alpha(2^B+1)+r$, and r is the remainder when N is divided by $N_{out} = (2^B+1)$.
2) The code achieves the Griesmer bound with equality.

Proof For any length $N = \alpha N_{out} + r$, the generator matrix can be written as

$$\mathbf{G} = [\underbrace{\mathbf{G}_{simplex}\ \mathbf{G}_{simplex}\ \cdots\ \mathbf{G}_{simplex}}_{\alpha}\ \mathbf{G}_{K\times r}]$$

where $\mathbf{G}_{simplex}$ is $K \times N_{out}$ matrix given as

$$\mathbf{G}_{simplex} = \begin{bmatrix} 0 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & 2 & \cdots & q-1 \end{bmatrix},$$

and the matrix $\mathbf{G}_{K\times r}$ consists of the first r columns of $\mathbf{G}_{simplex}$. The minimum distance of the SCRS code is

$$d_{\min}(\mathcal{C}) = \alpha 2^B + d_{K\times r}$$

where $d_{K\times r}$ is the minimum distance of the code with generator matrix $\mathbf{G}_{K\times r}$.

It is obvious that $d_{K\times r} = 0$ if $r = 0, 1$ and $d_{K\times r} = 1$ if $r = 2$. For more general r, the $K \times r$ code has a $(r - K) \times r$ parity check matrix

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 1 & 1 \\ 0 & 1 & \ddots & 0 & 1 & 2 \\ \vdots & & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & r-3 \\ 0 & 0 & \cdots & 1 & 1 & r-2 \end{bmatrix}.$$

By checking the minimum number of columns in the parity check matrix for which a nontrivial combination gives the all zero out, we can see that $d_{K\times r} = r - 1$ for $r > 0$. Therefore, $d_{\min}(\mathcal{C}) = \alpha 2^B + r - 1$.

The Griesmer bound for $K = 2$ tells us that to provide a minimum distance of d requires a code of length at least $d + 1$ when $d = 1, 2, \ldots, 2^B$. This means that our $K \times r$ code is optimal in the sense of achieving equality in the Griesmer bound. For the entire code, note that

$$\begin{aligned} \alpha 2^B + r - 1 + \left\lceil \frac{\alpha 2^B + r - 1}{2^B} \right\rceil &= \alpha 2^B + r - 1 + \alpha + 1 \\ &= \alpha(2^B + 1) + r \\ &= N. \end{aligned}$$

This shows that the SCRS codes are optimal in terms of the Griesmer bound. ■

Note that the case when $K = 2$ is a very practical scenario in distributed reception. For example, the scenario corresponds to the case when the transmitter sends a 16QAM (or QPSK) symbol and each receive node forwards a QPSK (or BPSK)

symbol to the fusion center. It would be very unlikely for the receive nodes (which might consist of cheap sensors) to send a high-order modulation symbol than QPSK to the fusion center in many distributed reception applications.

Remark: The proposed SCRS codes are suitable for selected subset ML decoding explained in Section 5.2.2 because every N_{out}-th receive node shares a common processing rule in the SCRS codes.

It is important to point out that maximum distance separable (MDS) codes that achieve the Singleton bound are not suitable to our system setup. For example, Reed-Solomon codes, which are the most popular MDS codes, must satisfy $N \leq 2^B - 1$ [108] where N is the number of receive nodes and B is the number of bits used to quantize the received signal at each receive node in our system setup. As explained in Section 5.2.1, we focus on the scenario of $B \leq \log_2(M)$ where M is the size of constellation $\mathcal{S}$. Thus, it is unlikely to satisfy the constraint of Reed-Solomon codes with our system setup.

5.3.4 Achievable rate

In most communication systems, the source can be modeled well as uniformly distributed over the constellation $\mathcal{S}$, which gives the achievable rate (or the mutual information) with channel realizations contained in the vector $\mathbf{h}$ as

$$I_u(\mathbf{h}) = \frac{1}{M} \sum_{s \in \mathcal{S}} \sum_{\mathbf{u} \in \mathcal{U}} \left\{ Pr(\mathbf{u} \mid s, \mathbf{h}) \log_2 \left(\frac{Pr(\mathbf{u} \mid s, \mathbf{h})}{\frac{1}{M} \sum_{s' \in \mathcal{S}} Pr(\mathbf{u} \mid s', \mathbf{h})} \right) \right\}$$

where $\mathcal{U}$ denotes the set of all 2^{NB} possible outputs from the receive nodes. Given this, the average achievable rate is given by

$$R_{avg} = E\left[I_u(\mathbf{h})\right] \tag{5.17}$$

with the expectation taken with respect to $\mathbf{h}$.

Note that all of these achievable rate expressions are dependent on the quantization structure used at each receive node. This is implicit because the transition probabilities between the input symbols and output symbols are dependent on this quantization structure. However, in general, it is hard to derive transition probabilities analytically, which prevents to have a closed-form expression of the achievable rate of the proposed coded receive diversity technique. Thus, we numerically study the achievable rate of the proposed coded receive diversity technique in Section 5.4 and show that the proposed scheme can provide benefits even with respect to the achievable rate in some scenarios.

5.4 Numerical Studies and Discussions

We perform Monte-Carlo simulations to evaluate the proposed coded receive diversity technique in this section. We assume all channel entries are independent, Rayleigh distributed, i.e., $h_i \sim \mathcal{CN}(0,1)$ for all i, during simulation; however, the proposed techniques can be applied to any kind of channel models of interest. The proposed scheme is based on the SCRS codes to simulate different numbers of the receive node N.

We first compare the proposed coded receive diversity technique to the scheme from [69]. In [69], the optimized codeword set matrix for local decision and decoding rules using simulated annealing for QPSK constellation data symbols, $B = 1$ processing at each receive node, and $N = 10$ nodes is given as

$$\text{Codeword Set Matrix: } (6, 12, 4, 9, 12, 9, 12, 6, 1, 3)$$

using the notation in [69]. Each integer in the matrix represents a binary column vector of the matrix, e.g., the integer 12 in column 2 represents $[0\ 0\ 1\ 1]^T$. Each row and column of the matrix represents one of the QPSK constellation points and

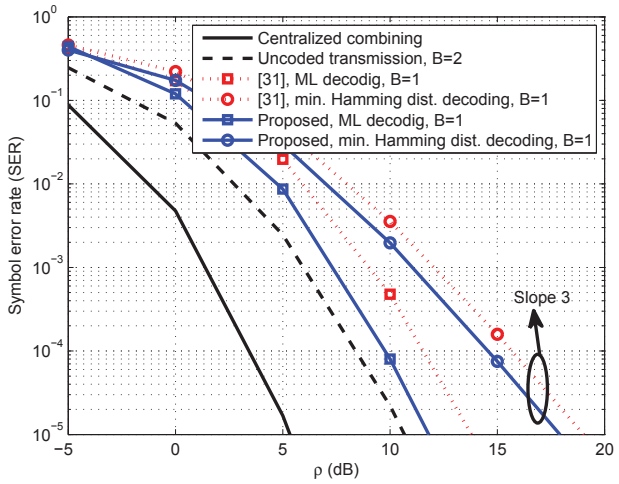


Fig. 5.2.: SER vs. SNR in dB scale with $M = 4$ and $N = 10$. Each receive node of the proposed scheme and the scheme from [69] forwards $B = 1$ bit per channel use to the fusion center while uncoded transmission relies on $B = \log_2 M$ forwarded bits per channel use from each node.

the decision rule of each receive node, respectively.[5] For example, if node 2 (which corresponds to column 2 in the codeword set matrix) detects the transmitted symbol as the first or second (third or fourth) QPSK constellation point, it forwards 0 (1) to the fusion center. If node 3, which corresponds to $[0\ 0\ 1\ 0]^T$ in the codeword set matrix, detects the transmitted symbol as the third QPSK constellation point, it forwards 1 to the fusion center. Otherwise, 0 is forwarded to the fusion center. With N binary bits forwarded from all the receive nodes, the fusion center adopts the same decoding rule with the proposed coded receive diversity technique for the fair comparison.

The concept of the codeword set matrix is similar to the matrix $\mathbf{C}$ in (5.7). Rows of both matrices represent the constellation points $\mathcal{S}$. However, local decision rules are completely different, i.e., the local decision rules in [69] are based on the codeword set matrix explained above while the proposed scheme relies on the generator matrix

[5]We rely on a hard decision for the local decision rule of the scheme from [69]. The local decision rule developed in [69] needs global channel knowledge (or the output distribution function of other nodes assuming a certain input symbol) at each node, which is unrealistic.

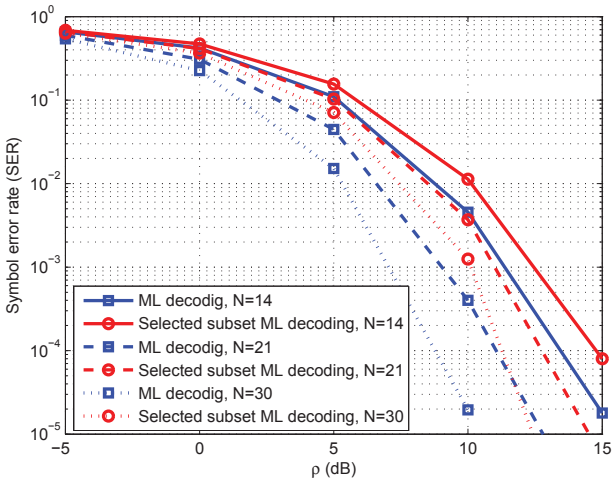
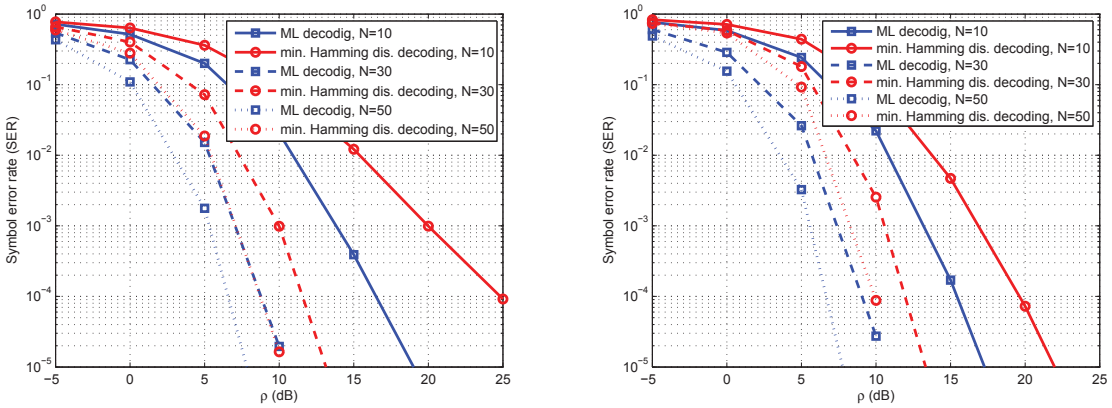


Fig. 5.3.: SER of the proposed coded receive diversity technique with ML and selected subset ML decoding schemes according to SNR in dB scale with $M = 8$ and $B = 1$.



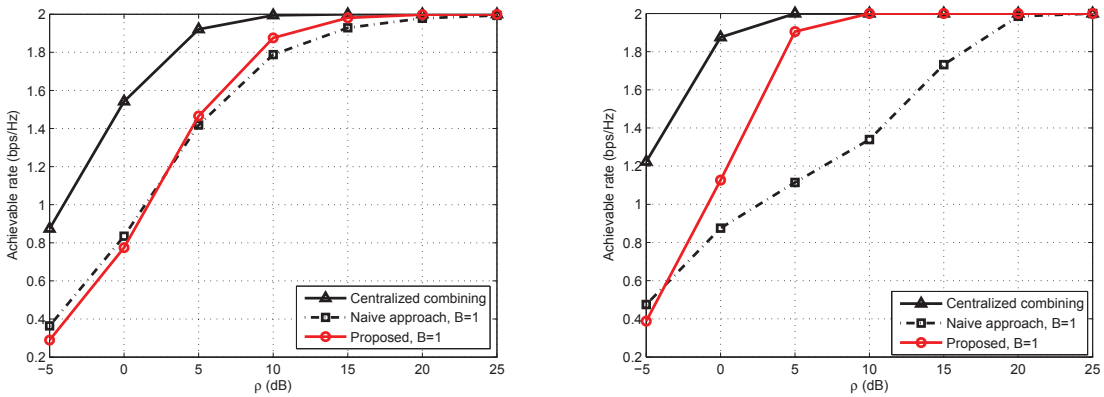
(a) $M = 8$ (8PSK) and $B = 1$.

(b) $M = 16$ (16QAM) and $B = 2$.

Fig. 5.4.: SER vs. SNR in dB scale with different values of M, B, and N.

G. Due to the dependency of the local decision rules, the optimized codeword set matrix in [69] is not able to maximize the minimum Hamming distance between codewords, resulting in performance degradation compared to the proposed coded receive diversity technique, which is shown in Fig. 5.2.

Fig. 5.2 compares the SER of the proposed scheme (using the SCRS codes) and the scheme in [69] for QPSK symbols according to the transmit SNR ρ with $N = 10$



(a) Rayleigh fading for all channels.

(b) $|h_1| = |h_3| = 1.5$ and $|h_2| = 0.3$.

Fig. 5.5.: Achievable rate vs. SNR in dB scale with $M = 4$, $N = 3$, and $B = 1$. The naive approach is explained in the motivating example in Section 5.1.1.

receive nodes. We also plot the results of centralized combining with $\mathbf{z} = \mathbf{h}/\|\mathbf{h}\|$ in (5.1) and uncoded $B = \log_2 M$ bits transmission from each receive node to the fusion center for comparison purpose. In uncoded transmission, the fusion center performs majority decoding based on the forwarded N estimated symbols from the receive nodes.

In both ML decoding and minimum Hamming distance decoding, the proposed scheme outperforms the scheme in [69].[6] These results are expected because the corresponding SCRS code has the minimum Hamming distance of 6 while the scheme in [69] has the minimum Hamming distance of 5. According to Lemma 5.2.1 and 5.2.3, the diversity orders of the SCRS code are 6 and 3 for ML and minimum Hamming distance decoding, respectively, while those of the scheme in [69] are 5 and 3, respectively. The slopes of the probability of errors for minimum Hamming distance decoding of the SCRS code and the scheme from [69] in Fig. 5.2 perfectly match with the diversity order analysis derived in Lemma 5.2.3. We also expect that the slopes of ML decoding would match with the derivation in Lemma 5.2.1 once the SNR becomes larger. Because we have an explicit expression of the SCRS code for

[6]We do not consider selected subset ML in this case because selected subset ML is not suitable to the scheme in [69].

an arbitrary number of the receive nodes N, the proposed coded receive diversity technique is very practical and easy to implement.

We compare the proposed diversity technique with ML and selected subset ML decoding schemes in Fig. 5.3. We set $M = 8$ and $B = 1$ (which gives $N_{out} = 7$ for the SCRS code) with different numbers of receive nodes N. When N_{out} (or L with the notation in the selected subset ML decoding section) divides N, it is clear that selected subset ML has the same diversity order as ML decoding although selected subset ML suffers from an SNR loss. Note that even when N_{out} does not divide N (the case when $N = 30$ in Fig. 5.3), selected subset ML gives comparable diversity gain with ML decoding with much less complexity.

In Figs. 5.4a and 5.4b, we plot the SER of the proposed coded receive diversity technique according to ρ with different values of M, B, and N. We can see from the figures that as the number of the receive nodes increases, we attain a better SER with the same ρ. Moreover, the number of the receive nodes N does not need to be large to achieve a practical SER of 10^{-2} or 10^{-3} with moderate ρ for all cases, which clearly shows the practicality of the proposed coded receive diversity technique.

Finally, we perform Monte-Carlo simulations with 10,000 channel realizations to verify the average achievable rate of the proposed scheme which is explained in Section 5.3.4. We compare R_{avg} in (5.17) of the proposed coded receive diversity technique, centralized combining, and the naive approach which is explained in the motivating example in Section 5.1.1. To simplify simulations, we set $\mathcal{S}$ with QPSK constellation, $B = 1$, and $N = 3$, which is the same setup as the motivating example in Section 5.1. We consider two different scenarios: 1) Rayleigh fading channels for all channels between the transmitter and the receive nodes and 2) normalized fading channels such that channel amplitudes are normalized as $|h_1| = |h_3| = 1.5$ and $|h_2| = 0.3$ for all channel realizations. The second scenario would be the case when the second node is in a deep fade while two other nodes are in stably good channel conditions.

We plot the results of the scenarios 1 and 2 in Figs. 5.5a and 5.5b, respectively. In the first scenario, the proposed scheme and the naive approach are comparable with

each other. This is reasonable because the proposed coding structure is not intended to increase the achievable rate. However, the proposed scheme outperforms the naive approach in the second scenario. This is because the second node that processes the imaginary component of the transmitted symbol is in a deep fade in the naive approach, resulting in significant achievable rate degradation. On the contrary, the proposed coded receive diversity is even better in the second scenario than the first since the fusion center can obtain much of mutual information only from nodes 1 and 3 which have good channel conditions.

There can be several different extensions of the proposed coded receive diversity technique: 1) supporting an arbitrary compression ratio; 2) extending to distributed space-time code designs with simultaneous transmission from the receive nodes; 3) accommodating multiple transmit antennas at the transmitter. In the next chapter, we will discuss the scenario of multiple transmit antennas in distributed reception.

6. QUANTIZED DISTRIBUTED RECEPTION FOR MIMO WIRELESS SYSTEMS USING SPATIAL MULTIPLEXING

In this chapter, we consider distributed reception with multiple transmit antennas. With the minimal quantization overhead at the receive nodes, i.e., one bit for each of the real and imaginary parts of the received signal, we develop an optimal ML receiver and a low-complexity ZF-type receiver at the fusion center. Despite its suboptimality, the ZF-type receiver is simple to implement and shows comparable performance with the ML receiver in the low SNR regime but experiences an error rate floor at high SNR. It is shown that this error floor can be overcome by increasing the number of receive nodes.

6.1 System Model

We consider a network consisting of a transmitter with N_t antennas, communicating with a receive fusion center that is connected to K geographically separated, single antenna receive nodes. The transmitter tries to send N_t independent data symbols simultaneously by spatial multiplexing[1] to the fusion center via the help of the receive nodes. The received signal at the k-th receive node is given as

$$y_k = \sqrt{\frac{\rho}{N_t}}\mathbf{h}_k^H \mathbf{x} + n_k, \quad k = 1, \cdots, K \tag{6.1}$$

[0]J. Choi, D. J. Love, D. R. Brown III, and M. Boutin, "Quantized Distributed Reception for MIMO Wireless Systems Using Spatial Multiplexing," submitted to *IEEE Transactions on Signal Processing*, Dec. 2014.

[1]The transmitter also can send a number of symbols smaller than N_t by adopting precoding or antenna selection, which is outside the scope of this paper.

where ρ is the transmit SNR, $\mathbf{h}_k \in \mathbb{C}^{N_t}$ is the independent and identically distributed (i.i.d.) Rayleigh fading channel vector between the transmitter and the k-th receive node, n_k is complex additive white Gaussian noise (AWGN) distributed as $\mathcal{CN}(0,1)$ at the k-th node, and $\mathbf{x} = [x_1, \cdots, x_{N_t}]^T$ is the transmitted signal vector. We assume $x_i \in \mathcal{S}$ is from a standard M-ary constellation

$$\mathcal{S} = \{s_1, \cdots, s_M\} \subset \mathbb{C}$$

which satisfies $\mathrm{E}[\|\mathbf{x}\|^2] = N_t$ and $\mathrm{E}[\mathbf{x}] = \mathbf{0}_{N_t}$. The input-output relation in (6.1) can be also written as

$$\mathbf{y} = \sqrt{\frac{\rho}{N_t}}\mathbf{H}\mathbf{x} + \mathbf{n}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_K \end{bmatrix}^T,$$
$$\mathbf{n} = \begin{bmatrix} n_1 & n_2 & \cdots & n_K \end{bmatrix}^T,$$
$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \cdots & \mathbf{h}_K \end{bmatrix}^H.$$

We further assume that the fusion center can access the full knowledge[2] of $\mathbf{h}_k$ for all k.

If the fusion center has full knowledge of y_k for all k, then the optimal receiver is given as

$$\hat{\mathbf{x}}_{\text{opt}} = \underset{\mathbf{x}' \in \mathcal{S}^{N_t}}{\operatorname{argmin}} \left\| \mathbf{y} - \sqrt{\frac{\rho}{N_t}}\mathbf{H}\mathbf{x}' \right\|^2$$

where $\mathcal{S}^n$ is the cartesian product of $\mathcal{S}$ of order n. However, we are interested in the scenario when each receive node *quantizes* its received signal and conveys the quantized received signal, $\hat{y}_k$, to the fusion center. Therefore, the fusion center needs to use other approaches to decode the transmitted symbols in our problem.

[2]Recently, we developed channel estimation techniques for the scenario of this paper in [111].

We assume $\hat{y}_k$ can be forwarded from the k-th receive node to the fusion center without any error. This assumption would be reasonable because the receive nodes and the fusion center are usually connected by a very high-rate link or located near each other in practice. We further assume that the forward link transmission and the LAN are operated on different time or frequency resources to prevent interference between the two.

To make the problem practical, we assume that the receive nodes only can perform very simple operation, i.e., they do not decode the transmitted vector $\mathbf{x}$ but instead simply quantize y_k directly. Moreover, to minimize the data transmission overhead from the receive nodes to the fusion center, we assume each receive node quantizes y_k using two bits, i.e., one bit for each of the real and imaginary parts of y_k. Thus, the quantized received signal $\hat{y}_k$ can be written as

$$\hat{y}_k = \mathrm{sgn}(\mathrm{Re}(y_k) - \tau_{\mathrm{Re},k}) + j\,(\mathrm{sgn}(\mathrm{Im}(y_k) - \tau_{\mathrm{Im},k}))$$

where $\mathrm{sgn}(\cdot)$ is the sign function defined as

$$\mathrm{sgn}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases},$$

and $\tau_{\mathrm{Re},k}$ and $\tau_{\mathrm{Im},k}$ are quantization thresholds of the real and imaginary parts of y_k at user k, respectively.

With a given realization of $\mathbf{h}_k$, we consider the simple, yet effective, thresholds

$$\tau_{\mathrm{Re},k} = \mathrm{E}\left[\mathrm{Re}(y_k)\right] = \mathrm{E}\left[\sqrt{\frac{\rho}{N_t}}\mathrm{Re}(\mathbf{h}_k^H\mathbf{x}) + \mathrm{Re}(n_k)\right] = 0,$$
$$\tau_{\mathrm{Im},k} = \mathrm{E}\left[\mathrm{Im}(y_k)\right] = \mathrm{E}\left[\sqrt{\frac{\rho}{N_t}}\mathrm{Im}(\mathbf{h}_k^H\mathbf{x}) + \mathrm{Im}(n_k)\right] = 0,$$

where equalities are based on the assumption that n_k is distributed as $\mathcal{CN}(0,1)$, or equivalently $\mathrm{Re}(n_k)$ and $\mathrm{Im}(n_k)$ are independent and both distributed as $\mathcal{N}(0,\frac{1}{2})$, and the entries of $\mathbf{x}$ are independently drawn from $\mathcal{S}^{N_t}$ with equal probabilities, which

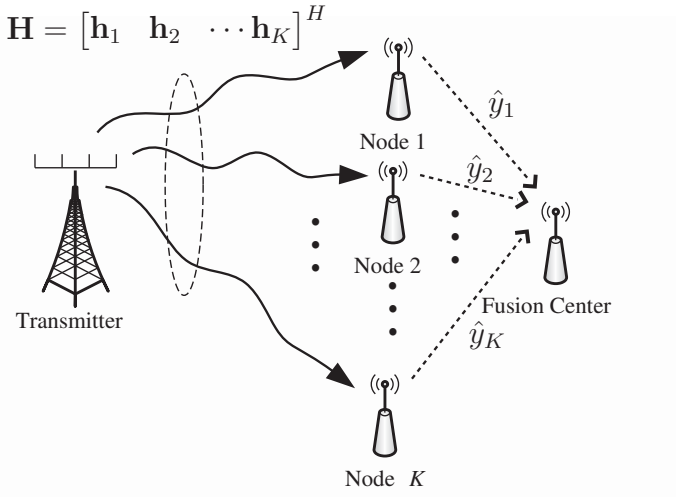


Fig. 6.1.: The conceptual figure of distributed reception with multiple antennas at the transmitter. Each receive node is equipped with a single receive antenna.

gives $\mathrm{E}[\mathrm{Re}(\mathbf{c}^T\mathbf{x})] = 0$ and $\mathrm{E}[\mathrm{Im}(\mathbf{c}^T\mathbf{x})] = 0$ for an arbitrary combining vector $\mathbf{c} \in \mathbb{C}^{N_t}$. Although simple, these thresholds are consistent with the optimal threshold design studied in [72] in an average sense. We assume the quantization thresholds $\tau_{\mathrm{Re},k} = 0$ and $\tau_{\mathrm{Im},k} = 0$ for the remainder of this paper.

Once the fusion center receives $\hat{y}_k$ from all receive nodes, it attempts to decoded the transmitted data symbols $\mathbf{x}$ using the forwarded information and channel knowledge. We define

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \cdots & \hat{y}_K \end{bmatrix}^T$$

which is useful in Section 6.2.2. The conceptual explanation of the scenario is depicted in Fig. 6.1.

6.2 Quantized Distributed Reception Techniques

With the knowledge of $\mathbf{H}$ and $\hat{\mathbf{y}}$ at the fusion center, we can implement different kinds of receivers considering complexity and performance. We first develop an optimal ML receiver and low-complexity ZF-type receiver. Then, we discuss the per-

formance of receivers regarding system parameters such as ρ and K. We also explain a possible modification of the ZF-type receiver and analyze the achievable rate of quantized distributed reception.

6.2.1 ML receiver

We convert the problem of interest to the real domain to facilitate analysis. This can be done by defining $\mathbf{H}_{\mathrm{R},k} \in \mathbb{R}^{2\times 2N_t}$, $\mathbf{x}_{\mathrm{R}} \in \mathbb{R}^{2N_t}$ and $\mathbf{n}_{\mathrm{R},k} \in \mathbb{R}^2$ as

$$\mathbf{H}_{\mathrm{R},k} = \begin{bmatrix} \mathrm{Re}(\mathbf{h}_k^T) & \mathrm{Im}(\mathbf{h}_k^T) \\ -\mathrm{Im}(\mathbf{h}_k^T) & \mathrm{Re}(\mathbf{h}_k^T) \end{bmatrix} = \begin{bmatrix} \mathbf{h}_{\mathrm{R},k,1}^T \\ \mathbf{h}_{\mathrm{R},k,2}^T \end{bmatrix},$$
$$\mathbf{x}_{\mathrm{R}} = \begin{bmatrix} \mathrm{Re}(\mathbf{x}) \\ \mathrm{Im}(\mathbf{x}) \end{bmatrix}, \quad \mathbf{n}_{\mathrm{R},k} = \begin{bmatrix} \mathrm{Re}(n_k) \\ \mathrm{Im}(n_k) \end{bmatrix}$$

where

$$\mathbf{h}_{\mathrm{R},k,1} = \begin{bmatrix} \mathrm{Re}(\mathbf{h}_k) \\ \mathrm{Im}(\mathbf{h}_k) \end{bmatrix}, \quad \mathbf{h}_{\mathrm{R},k,2} = \begin{bmatrix} -\mathrm{Im}(\mathbf{h}_k) \\ \mathrm{Re}(\mathbf{h}_k) \end{bmatrix}.$$

Then, the received signal y_k also can be rewritten in the real domain as

$$\mathbf{y}_{\mathrm{R},k} = \begin{bmatrix} y_{\mathrm{R},k,1} \\ y_{\mathrm{R},k,2} \end{bmatrix} = \begin{bmatrix} \mathrm{Re}(y_k) \\ \mathrm{Im}(y_k) \end{bmatrix} = \sqrt{\frac{\rho}{N_t}}\mathbf{H}_{\mathrm{R},k}\mathbf{x}_{\mathrm{R}} + \mathbf{n}_{\mathrm{R},k},$$

and the vectorized version of the quantized $\hat{y}_k$ in the real domain is given as

$$\hat{\mathbf{y}}_{\mathrm{R},k} = \begin{bmatrix} \hat{y}_{\mathrm{R},k,1} \\ \hat{y}_{\mathrm{R},k,2} \end{bmatrix} = \begin{bmatrix} \mathrm{sgn}(\mathrm{Re}(y_k)) \\ \mathrm{sgn}(\mathrm{Im}(y_k)) \end{bmatrix}. \tag{6.2}$$

Once the fusion center receives $\hat{\mathbf{y}}_{\mathrm{R},k}$ from all receive nodes, it generates the *sign-refined* channel matrix $\widetilde{\mathbf{H}}_{\mathrm{R},k}$ according to

$$\widetilde{\mathbf{H}}_{\mathrm{R},k} = \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathrm{R},k,1}^T \\ \widetilde{\mathbf{h}}_{\mathrm{R},k,2}^T \end{bmatrix}$$

where $\widetilde{\mathbf{h}}_{\mathrm{R},k,i}$ is defined as

$$\widetilde{\mathbf{h}}_{\mathrm{R},k,i} = \hat{y}_{\mathrm{R},k,i}\mathbf{h}_{\mathrm{R},k,i}. \tag{6.3}$$

Because $\hat{y}_{\mathrm{R},k,i}$ is ± 1, (6.3) can be considered as a sign refinement of $\mathbf{h}_{\mathrm{R},k,i}$. We let $\mathcal{S}_{\mathrm{R}}$ be

$$\mathcal{S}_{\mathrm{R}} = \left\{ \begin{bmatrix} \mathrm{Re}(s_1) \\ \mathrm{Im}(s_1) \end{bmatrix}, \cdots, \begin{bmatrix} \mathrm{Re}(s_M) \\ \mathrm{Im}(s_M) \end{bmatrix} \right\}$$

where M is the size of the constellation $\mathcal{S}$. We also define two sets $\mathcal{P}$ and $\mathcal{N}$ using $\{y_{\mathrm{R},k,i}\}$ for $k \in \{1, \cdots, K\}$ and $i \in \{1, 2\}$ as

$$\mathcal{P} = \{(k,i) : \ y_{\mathrm{R},k,i} \geq 0\}, \quad \mathcal{N} = \{(k,i) : \ y_{\mathrm{R},k,i} < 0\}.$$

With these definitions, we can define a likelihood function as

$$\begin{aligned}
L(\mathbf{x}'_{\mathrm{R}}) &= \Pr\left(\sqrt{\frac{\rho}{N_t}} \mathbf{h}_{\mathrm{R},k,i}^T \mathbf{x}'_{\mathrm{R}} + n_{\mathrm{R},k,i} \geq 0 \middle| \forall (k,i) \in \mathcal{P} \right) \\
&\quad \cdot \Pr\left(\sqrt{\frac{\rho}{N_t}} \mathbf{h}_{\mathrm{R},k,i}^T \mathbf{x}'_{\mathrm{R}} + n_{\mathrm{R},k,i} < 0 \middle| \forall (k,i) \in \mathcal{N} \right) \\
&\overset{(a)}{=} \Pr\left(\sqrt{\frac{\rho}{N_t}} \widetilde{\mathbf{h}}_{\mathrm{R},k,i}^T \mathbf{x}'_{\mathrm{R}} \geq -n_{\mathrm{R},k,i} \middle| \forall (k,i) \in \mathcal{P} \right) \\
&\quad \cdot \Pr\left(\sqrt{\frac{\rho}{N_t}} \widetilde{\mathbf{h}}_{\mathrm{R},k,i}^T \mathbf{x}'_{\mathrm{R}} \geq n_{\mathrm{R},k,i} \middle| \forall (k,i) \in \mathcal{N} \right) \\
&\overset{(b)}{=} \Pr\left(\sqrt{\frac{\rho}{N_t}} \widetilde{\mathbf{h}}_{\mathrm{R},k,i}^T \mathbf{x}'_{\mathrm{R}} \geq -n_{\mathrm{R},k,i} \middle| \forall (k,i) \in \mathcal{P} \right) \\
&\quad \cdot \Pr\left(\sqrt{\frac{\rho}{N_t}} \widetilde{\mathbf{h}}_{\mathrm{R},k,i}^T \mathbf{x}'_{\mathrm{R}} \geq -n_{\mathrm{R},k,i} \middle| \forall (k,i) \in \mathcal{N} \right) \\
&\overset{(c)}{=} \prod_{i=1}^{2} \prod_{k=1}^{K} \Phi\left(\sqrt{\frac{2\rho}{N_t}} \widetilde{\mathbf{h}}_{\mathrm{R},k,i}^T \mathbf{x}'_{\mathrm{R}} \right)
\end{aligned}$$

where $\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{\tau^2}{2}} d\tau$, (a) is based on the sign refinement in (6.3), (b) is because $n_{\mathrm{R},k,i}$ and $-n_{\mathrm{R},k,i}$ have the same distribution (or the same probability density function) such that $\Pr(c \geq n_{\mathrm{R},k,i}) = \Pr(c \geq -n_{\mathrm{R},k,i})$ for an arbitrary constant c, and

(c) comes from the fact that $n_{\mathrm{R},k,i}$ is independent for all k and i and from distribution $\mathcal{N}\left(0, \frac{1}{2}\right)$. Then, the ML receiver is given as[3]

$$\hat{\mathbf{x}}_{\mathrm{R,ML}} = \underset{\mathbf{x}'_{\mathrm{R}} \in \mathcal{S}_{\mathrm{R}}^{N_t}}{\operatorname{argmax}} \prod_{i=1}^{2} \prod_{k=1}^{K} \Phi\left(\sqrt{\frac{2\rho}{N_t}} \widetilde{\mathbf{h}}_{\mathrm{R},k,i}^{T} \mathbf{x}'_{\mathrm{R}}\right). \tag{6.4}$$

The complexity of the exhaustive search of the ML receiver increases exponentially with the number of transmit symbols in spatial multiplexing, i.e., we need to search over M^{N_t} elements. Therefore, in practice, it is desired to implement a low complexity receiver for large numbers of transmit antennas.

Remark 1: If the number of receive nodes K is less than the number of transmit antennas N_t, then the decoding performance at the fusion center would be very poor. This situation will likely not hold for our problem setting because we can easily have $K \gg N_t$ based on the IoT environment.

Remark 2: Instead of quantizing both the real and imaginary parts of the received signal at each node, we can have the same performance on average by quantizing and forwarding only the real *or* imaginary part of the received signal with twice the number of receive nodes. This is based on the assumption that the real and imaginary parts of the noise n_k are i.i.d. for all k.

6.2.2 Low-complexity zero-forcing-type receiver

The ML receiver defined in (6.4) has no constraint on the norm of the transmit vector $\mathbf{x}$. To develop our ZF-type receiver, however, we assume $\|\mathbf{x}\|^2 = N_t$. If $\mathcal{S}$ is a phase shift keying (PSK) constellation with $|s_m|^2 = 1$ for all m, this constraint is trivially satisfied. Even for a quadrature amplitude modulation (QAM) constellation (that is properly normalized), the constraint can be approximately satisfied because

$$\sum_{i=1}^{N_t} |x_i|^2 \approx N_t$$

[3]A similar ML receiver is also derived in [72].

when N_t is large and x_i is drawn from $\mathcal{S}$ with equal probabilities. Simulation results in Section 6.3 show that our ZF-type receiver works even with not-so-large N_t, e.g., $N_t = 10$.

Before proposing our ZF-type receiver, we first state the following lemma which establishes the theoretical foundation of our receiver.

Lemma 6.2.1 *Define a matrix* $\widetilde{\mathbf{H}}_{\mathrm{R,S}} \in \mathbb{R}^{2K \times 2N_t}$ *by stacking* $\widetilde{\mathbf{H}}_{\mathrm{R},k}$ *as*

$$\widetilde{\mathbf{H}}_{\mathrm{R,S}} = \begin{bmatrix} \widetilde{\mathbf{H}}_{\mathrm{R},1}^T & \widetilde{\mathbf{H}}_{\mathrm{R},2}^T \cdots & \widetilde{\mathbf{H}}_{\mathrm{R},K}^T \end{bmatrix}^T, \tag{6.5}$$

and let $\mathbf{t}(\mathbf{x}'_{\mathrm{R}})$ *be*

$$\begin{aligned} \mathbf{t}(\mathbf{x}'_{\mathrm{R}}) &= \begin{bmatrix} t_1(\mathbf{x}'_{\mathrm{R}}) & t_2(\mathbf{x}'_{\mathrm{R}}) & \cdots & t_{2K}(\mathbf{x}'_{\mathrm{R}}) \end{bmatrix}^T, \\ t_\ell(\mathbf{x}'_{\mathrm{R}}) &= \widetilde{\mathbf{h}}_{\mathrm{R},k,i}^T \mathbf{x}'_{\mathrm{R}} \end{aligned}$$

where $\ell = 2(k-1)+i$ *for* $k = 1, \cdots, K$ *and* $i = 1, 2$. *Note that* $\|\mathbf{x}'_{\mathrm{R}}\|^2 = N_t$ *based on the assumption. Then the likelihood function* $L(\mathbf{x}'_{\mathrm{R}})$ *is upper bounded as*

$$\begin{aligned} L(\mathbf{x}'_{\mathrm{R}}) &= \prod_{i=1}^{2} \prod_{k=1}^{K} \Phi\left(\sqrt{\frac{2\rho}{N_t}} \widetilde{\mathbf{h}}_{\mathrm{R},k,i}^T \mathbf{x}'_{\mathrm{R}} \right) \\ &= \prod_{\ell=1}^{2K} \Phi\left(\sqrt{\frac{2\rho}{N_t}} t_\ell(\mathbf{x}'_{\mathrm{R}}) \right) \\ &\leq \prod_{\ell=1}^{2K} \Phi\left(\sqrt{\frac{\rho}{K}} \|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}} \right) \end{aligned}$$

when $t_\ell(\mathbf{x}'_{\mathrm{R}}) = \sqrt{\frac{N_t}{2K}} \|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}}$ *for all* ℓ *where* $\|\cdot\|_{\mathrm{A}}$ *is an arbitrary matrix norm that is consistent with the vector two-norm.*

Proof To prove Lemma 6.2.1, we derive an upper bound of the maximum of $L(\mathbf{x}'_{\mathrm{R}})$ with the relaxed constraint $\mathbf{x}'_{\mathrm{R}} \in \mathbb{R}^{2N_t}$ instead of $\mathbf{x}'_{\mathrm{R}} \in \mathcal{S}_{\mathrm{R}}^{N_t}$. Note that the norm constraint $\|\mathbf{x}'_{\mathrm{R}}\|^2 = N_t$ still holds. With the definitions of $t_\ell(\mathbf{x}'_{\mathrm{R}})$ and $\mathbf{t}(\mathbf{x}'_{\mathrm{R}})$, we have

$$\begin{aligned}
\max_{\substack{\mathbf{x}'_{\mathrm{R}}\in\mathbb{R}^{2N_t},\\ \|\mathbf{x}'_{\mathrm{R}}\|^2=N_t}} L(\mathbf{x}'_{\mathrm{R}}) &= \max_{\substack{\mathbf{x}'_{\mathrm{R}}\in\mathbb{R}^{2N_t},\\ \|\mathbf{x}'_{\mathrm{R}}\|^2=N_t}} \prod_{i=1}^{2}\prod_{k=1}^{K} \Phi\left(\sqrt{\frac{2\rho}{N_t}}\widetilde{\mathbf{h}}_{\mathrm{R},k,i}^T \mathbf{x}'_{\mathrm{R}}\right) \\
&\le \max_{\substack{\mathbf{t}(\mathbf{x}'_{\mathrm{R}})\in\mathbb{R}^{2K},\\ \|\mathbf{t}(\mathbf{x}'_{\mathrm{R}})\|^2\le N_t\|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}}^2}} \prod_{\ell=1}^{2K} \Phi\left(\sqrt{\frac{2\rho}{N_t}} t_\ell(\mathbf{x}'_{\mathrm{R}})\right) \qquad (6.6) \\
&= \max_{\substack{\mathbf{t}(\mathbf{x}'_{\mathrm{R}})\in\mathbb{R}^{2K},\\ \|\mathbf{t}(\mathbf{x}'_{\mathrm{R}})\|^2\le N_t\|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}}^2,\\ t_\ell(\mathbf{x}'_{\mathrm{R}})>0,\forall\ell}} \prod_{\ell=1}^{2K} \Phi\left(\sqrt{\frac{2\rho}{N_t}} t_\ell(\mathbf{x}'_{\mathrm{R}})\right) \qquad (6.7) \\
&= \max_{\substack{\mathbf{t}(\mathbf{x}'_{\mathrm{R}})\in\mathbb{R}^{2K},\\ \|\mathbf{t}(\mathbf{x}'_{\mathrm{R}})\|^2= N_t\|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}}^2,\\ t_\ell(\mathbf{x}'_{\mathrm{R}})>0,\forall\ell}} \prod_{\ell=1}^{2K} \Phi\left(\sqrt{\frac{2\rho}{N_t}} t_\ell(\mathbf{x}'_{\mathrm{R}})\right) \qquad (6.8)
\end{aligned}$$

where (6.6) is based on the fact that

$$\|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\mathbf{x}'_{\mathrm{R}}\|^2 \le \|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}}^2 \|\mathbf{x}'_{\mathrm{R}}\|^2 = N_t \|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}}^2,$$

and (6.7) is because $\Phi(a) > \Phi(0)$ for $a > 0$. Note that the inequality constraint on $\|\mathbf{t}(\mathbf{x}'_{\mathrm{R}})\|^2$ in (6.7) is changed to the equality constraint in (6.8).

The objective function in (6.8) is trivially bounded by one; however, there is a certain maximum point in our problem because of the norm constraint of $\|\mathbf{t}(\mathbf{x}'_{\mathrm{R}})\|^2 = N_t\|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}}^2$. Let $g_\ell = \sqrt{\frac{2\rho}{N_t}} t_\ell(\mathbf{x}'_{\mathrm{R}})$ and $\mathbf{g} = \begin{bmatrix} g_1 & g_2 & \cdots & g_{2K} \end{bmatrix}^T$. Instead of finding the solution for (6.8) directly, we first find a local extrema of

$$\begin{aligned}
\log\left[\prod_{\ell=1}^{2K} \Phi\left(\sqrt{\frac{2\rho}{N_t}} t_\ell(\mathbf{x}'_{\mathrm{R}})\right)\right] &= \sum_{\ell=1}^{2K} \log\left[\Phi\left(\sqrt{\frac{2\rho}{N_t}} t_\ell(\mathbf{x}'_{\mathrm{R}})\right)\right] \\
&= \sum_{\ell=1}^{2K} \log \Phi\left(g_\ell\right) \qquad (6.9)
\end{aligned}$$

by looking at the point at which the tangential derivatives to the circle $\|\mathbf{g}\|^2 = 2\rho\|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}}^2$ are equal to zero.[4] The tangential derivatives of (6.9) are given by

$$\left(g_n\frac{\partial}{\partial g_m} - g_m\frac{\partial}{\partial g_n}\right)\sum_{\ell=1}^{2K}\log\Phi\left(g_\ell\right) = g_n\frac{\Phi'\left(g_m\right)}{\Phi\left(g_m\right)} - g_m\frac{\Phi'\left(g_n\right)}{\Phi\left(g_n\right)}$$

for $n, m = 1, 2, \cdots, 2K$ and $n \neq m$. Setting the tangential derivatives equal to zero, we obtain the equations

$$g_n\frac{\Phi'\left(g_m\right)}{\Phi\left(g_m\right)} = g_m\frac{\Phi'\left(g_n\right)}{\Phi\left(g_n\right)}$$

or equivalently,

$$\frac{1}{g_m}\frac{\Phi'\left(g_m\right)}{\Phi\left(g_m\right)} = \frac{1}{g_n}\frac{\Phi'\left(g_n\right)}{\Phi\left(g_n\right)} \tag{6.10}$$

for $n, m = 1, 2, \cdots, 2K$ and $n \neq m$ because $g_\ell > 0$ for all ℓ. Clearly, this system of equations is satisfied when $g_n = g_m$ for all $n, m = 1, 2, \cdots, 2K$. Under the constraint $\|\mathbf{g}\|^2 = 2\rho\|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}}^2$, one possible solution point is given as

$$g_\ell = \sqrt{\frac{\rho}{K}}\|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}} \tag{6.11}$$

for all ℓ. Note that the point in (6.11) is the only solution for (6.10) because

$$G(s) = \frac{1}{s}\Phi'(s)\frac{1}{\Phi(s)}$$

is a product of three functions that are strictly monotonically decreasing with $s \in (0, \infty)$, and thus $G(s)$ is also strictly monotonically decreasing with s.

[4]Because our searching space is restricted to the circle $\|\mathbf{g}\|^2 = 2\rho\|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}}^2$, the point where the tangential derivatives equal to zero is a local extrema of the objective function.

Because $t_\ell(\mathbf{x}'_{\mathrm{R}}) = \sqrt{\frac{N_t}{2\rho}} g_\ell$, the point

$$t_\ell(\mathbf{x}'_{\mathrm{R}}) = \sqrt{\frac{N_t}{2K}} \|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}}$$

for $\ell = 1, \cdots, 2K$ is the only extreme point of the objective function in (6.8). We can show that the extreme point is indeed the maximum point of (6.8) by using the lemma in Appendix A.5. ■

Lemma 6.2.1 states that when $\mathbf{t}(\mathbf{x}'_{\mathrm{R}}) = \sqrt{\frac{N_t}{2K}} \|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}} \mathbf{1}_{2K}$, it maximizes the likelihood function with the norm constraint $\|\mathbf{t}(\mathbf{x}'_{\mathrm{R}})\|^2 = N_t \|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}}^2$. From the fact that

$$\mathbf{t}(\mathbf{x}'_{\mathrm{R}}) = \widetilde{\mathbf{H}}_{\mathrm{R,S}} \mathbf{x}'_{\mathrm{R}},$$

the vector $\check{\mathbf{x}}_{\mathrm{R}}$, which is given as

$$\check{\mathbf{x}}_{\mathrm{R}} = \widetilde{\mathbf{H}}_{\mathrm{R,S}}^{\dagger} \mathbf{t}(\mathbf{x}'_{\mathrm{R}}) = \sqrt{\frac{N_t}{2K}} \|\widetilde{\mathbf{H}}_{\mathrm{R,S}}\|_{\mathrm{A}} \widetilde{\mathbf{H}}_{\mathrm{R,S}}^{\dagger} \mathbf{1}_{2K},$$

would be a reasonable estimate for the transmitted vector.

To implement this receiver in terms of the quantized received signals, let $\hat{\mathbf{y}}_{\mathrm{R}}$ be

$$\hat{\mathbf{y}}_{\mathrm{R}} = \begin{bmatrix} \hat{\mathbf{y}}_{\mathrm{R},1}^T & \hat{\mathbf{y}}_{\mathrm{R},2}^T & \cdots & \hat{\mathbf{y}}_{\mathrm{R},K}^T \end{bmatrix}^T$$

where $\hat{\mathbf{y}}_{\mathrm{R},k}$ is defined in (6.2). It is easy to show by using the relation between $\mathbf{H}_{\mathrm{R,S}}$ and $\widetilde{\mathbf{H}}_{\mathrm{R,S}}$ (or between their rows given in (6.3)) that

$$\widetilde{\mathbf{H}}_{\mathrm{R,S}}^{\dagger} \mathbf{1}_{2K} = \mathbf{H}_{\mathrm{R,S}}^{\dagger} \hat{\mathbf{y}}_{\mathrm{R}}$$

because the i-th row of $\widetilde{\mathbf{H}}_{\mathrm{R,S}}$ is the same as that of $\mathbf{H}_{\mathrm{R,S}}$ with the sign adjustment by the sign of the i-th element of $\hat{\mathbf{y}}_{\mathrm{R}}$. Based on these observations, we propose a ZF-type receiver at the fusion center, i.e., the fusion center generates $\check{\mathbf{x}}_{\mathrm{R,ZF}} \in \mathbb{R}^{2N_t}$ as

$$\check{\mathbf{x}}_{\mathrm{R,ZF}} = \mathbf{H}_{\mathrm{R,S}}^{\dagger} \hat{\mathbf{y}}_{\mathrm{R}}. \tag{6.12}$$

With $\mathbf{H}$ and $\hat{\mathbf{y}}$ which are defined in Section 6.1, the same receiver with (6.12) can be implemented in the complex domain as

$$\check{\mathbf{x}}_{\mathrm{ZF}} = \mathbf{H}^{\dagger} \hat{\mathbf{y}} \tag{6.13}$$

where $\check{\mathbf{x}}_{\mathrm{ZF}} \in \mathbb{C}^{N_t}$.

Note that the squared norm of $\check{\mathbf{x}}_{\mathrm{ZF}}$ may not be N_t anymore; however, the normalization term does not have any impact on PSK symbol decisions. If x_i is from a QAM constellation, the normalization term is important because the amplitude of each entry of $\check{\mathbf{x}}_{\mathrm{ZF}}$ does matter in the decoding process. Because the fusion center does not have any knowledge of the squared norm of $\mathbf{x}$, the best way to normalize $\check{\mathbf{x}}_{\mathrm{ZF}}$ is

$$\|\check{\mathbf{x}}_{\mathrm{ZF}}\|^2 = N_t$$

assuming the elements of $\mathbf{x}$ are uniformly distributed from $\mathcal{S}$.

Finally, the fusion center needs to detect $\hat{\mathbf{x}}_{\mathrm{ZF}} = \begin{bmatrix} \hat{x}_{\mathrm{ZF},1} & \hat{x}_{\mathrm{ZF},2} & \cdots & \hat{x}_{\mathrm{ZF},N_t} \end{bmatrix}^T$ by selecting the closest constellation point from $\check{\mathbf{x}}_{\mathrm{ZF}} = \begin{bmatrix} \check{x}_{\mathrm{ZF},1} & \check{x}_{\mathrm{ZF},2} & \cdots & \check{x}_{\mathrm{ZF},N_t} \end{bmatrix}^T$ as

$$\hat{x}_{\mathrm{ZF},n} = \underset{s' \in \mathcal{S}}{\operatorname{argmin}} \, |\check{x}_{\mathrm{ZF},n} - s'|^2 \tag{6.14}$$

for $n = 1, \cdots, N_t$. The complexity of the ZF-type receiver is much lower than that of the ML receiver because each of these minimizations is over a set of M elements.

6.2.3 Receiver performance

In this subsection, we analyze the performance of ML and ZF-type estimators where the entries of the estimates can be arbitrary complex numbers. We assume $\|\mathbf{x}\|^2 = N_t$ in this subsection. The following lemma shows the behavior of the ML estimator in the asymptotic regime of K for arbitrary $\rho > 0$.

Lemma 6.2.2 *Let* $\check{\mathbf{x}}_{\mathrm{ML}}$ *be the outcome of the ML estimator*

$$\check{\mathbf{x}}_{\mathrm{ML}} = \underset{\substack{\mathbf{x}' \in \mathbb{C}^{N_t}, \\ \|\mathbf{x}'\|^2 = N_t}}{\operatorname{argmax}} L(\mathbf{x}'). \tag{6.15}$$

For arbitrary $\rho > 0$, $\check{\mathbf{x}}_{\mathrm{ML}}$ *converges to the true transmitted vector* $\mathbf{x}$ *in probability, i.e.,*

$$\check{\mathbf{x}}_{\mathrm{ML}} \xrightarrow{p} \mathbf{x}$$

as $K \to \infty$.

Proof We consider the real domain in the proof to simplify notation. The lemma can be proved by showing the inequality

$$L(\mathbf{x}_{\mathrm{R}}) > L(\mathbf{u}_{\mathrm{R}})$$

in probability for any $\mathbf{u}_{\mathrm{R}} \in \mathbb{R}^{2N_t} \setminus \{\mathbf{x}_{\mathrm{R}}\}$ with the constraint $\|\mathbf{u}_{\mathrm{R}}\|^2 = N_t$ when $K \to \infty$ for arbitrary $\rho > 0$. We take logarithm of the likelihood function and have

$$\begin{aligned} \log L(\mathbf{x}^{\ddagger}) &= \log \left(\prod_{i=1}^{2} \prod_{k=1}^{K} \Phi \left(\sqrt{\frac{2\rho}{N_t}} \widetilde{\mathbf{h}}_{\mathrm{R},k,i}^{T} \mathbf{x}^{\ddagger} \right) \right) \\ &= \sum_{i=1}^{2} \sum_{k=1}^{K} \log \Phi \left(\sqrt{\frac{2\rho}{N_t}} \widetilde{\mathbf{h}}_{\mathrm{R},k,i}^{T} \mathbf{x}^{\ddagger} \right). \end{aligned}$$

Because the $\widetilde{\mathbf{h}}_{\mathrm{R},k,i}$'s are independent for all k,

$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}\log\Phi\left(\sqrt{\frac{2\rho}{N_t}}\widetilde{\mathbf{h}}_{\mathrm{R},k,i}^{T}\mathbf{x}^{\ddagger}\right)\xrightarrow{p}\mathrm{E}\left[\log\Phi\left(\sqrt{\frac{2\rho}{N_t}}\widetilde{\mathbf{h}}_{\mathrm{R},k,i}^{T}\mathbf{x}^{\ddagger}\right)\right]$$

by the weak law of large numbers, and we have

$$\frac{1}{K}\log L(\mathbf{x}^{\ddagger})\xrightarrow{p}2\mathrm{E}\left[\log\Phi\left(\sqrt{\frac{2\rho}{N_t}}\widetilde{\mathbf{h}}_{\mathrm{R},k,i}^{T}\mathbf{x}^{\ddagger}\right)\right]$$

as $K\to\infty$ where the expectation is taken over the channel.

Then, we need to show that

$$\mathrm{E}\left[\log\Phi\left(\sqrt{\frac{\rho}{N_t}}\widetilde{\mathbf{h}}_{\mathrm{R},k,i}^{T}\mathbf{x}_{\mathrm{R}}\right)\right]>\mathrm{E}\left[\log\Phi\left(\sqrt{\frac{\rho}{N_t}}\widetilde{\mathbf{h}}_{\mathrm{R},k,i}^{T}\mathbf{u}_{\mathrm{R}}\right)\right]$$

where the expectations are taken over the channel. Because $\log\Phi(\cdot)$ is a strictly monotonically increasing concave function, the above inequality is true if $\widetilde{\mathbf{h}}_{\mathrm{R},k,i}^{T}\mathbf{x}_{\mathrm{R}}$ first-order stochastically dominates $\widetilde{\mathbf{h}}_{\mathrm{R},k,i}^{T}\mathbf{u}_{\mathrm{R}}$ [112]. In Appendix A.6, we show

$$\widetilde{\mathbf{h}}_{\mathrm{R},k,i}^{T}\mathbf{x}_{\mathrm{R}}\overset{d}{>}\widetilde{\mathbf{h}}_{\mathrm{R},k,i}^{T}\mathbf{u}_{\mathrm{R}} \tag{6.16}$$

conditioned on the received signal $y_{\mathrm{R},k,i}$ where $\overset{d}{>}$ denotes strict first-order stochastic dominance. ■

We define the MSE between $\mathbf{x}$ and $\check{\mathbf{x}}$ as

$$\mathrm{MSE}=\frac{1}{N_t}\mathrm{E}\left[\|\mathbf{x}-\check{\mathbf{x}}\|^{2}\right]$$

where the expectation is taken over the realizations of channel and noise. The following corollary shows the MSE performance of the ML estimator in the asymptotic regime of K for arbitrary $\rho>0$.

Corollary 6.2.1 *The MSE of the ML estimator converges to zero, i.e.,*

$$\lim_{K\to\infty} \mathrm{MSE}_{\mathrm{ML}} \to 0$$

for arbitrary $\rho > 0$.

Proof Note that the norm of $\mathbf{x}$ is bounded, i.e.,

$$\|\mathbf{x}\|^2 = N_t < \infty.$$

The convergence in probability of a random variable with a bounded norm implies the convergence in mean-square sense [113]. Thus, we have

$$\lim_{K\to\infty} \mathrm{E}\left[\left\|\mathbf{x} - \check{\mathbf{x}}_{\mathrm{ML}\|}\right\|^2\right]$$

which finishes the proof. ■

This analytical derivation for the ML estimator shows that the proposed ML receiver can perfectly decode the transmitted vector in the limit as K grows large with fixed ρ. Moreover, numerical studies in Section 6.3 show that increasing ρ would be sufficient for the ML receiver to decode the transmitted vector correctly with fixed, but sufficiently large, K.

We now analyze the MSE of the ZF-type estimator. Although it is difficult to derive the MSE of the ZF-type estimator in general, we are able to have a closed-form expression for $\mathrm{MSE}_{\mathrm{ZF}}$ by approximating quantization loss as additional Gaussian noise where the approximation is frequently adopted in many frame expansion works, e.g., [75, 76, 114].

Lemma 6.2.3 *If we approximate the quantization error using an additional Gaussian noise* $\mathbf{w}$ *as*

$$\hat{\mathbf{y}} = \sqrt{\frac{\rho}{N_t}}\mathbf{H}\mathbf{x} + \mathbf{n} + \mathbf{w} \tag{6.17}$$

with[5] $\mathbf{w} \sim \mathcal{CN}(\mathbf{0}_K, \sigma_q^2 \frac{\rho}{N_t}\mathbf{I}_K)$ *and assume* $\mathbf{H}^H\mathbf{H} = K\mathbf{I}_{N_t}$*, the MSE of the ZF-type estimator is given as*[6]

$$\mathrm{MSE}_{\mathrm{ZF}} = \mathrm{E}\left[\left\|\mathbf{x} - \sqrt{N_t}\frac{\check{\mathbf{x}}_{\mathrm{ZF}}}{\|\check{\mathbf{x}}_{\mathrm{ZF}}\|}\right\|^2\right] = \frac{N_t\rho^{-1} + \sigma_q^2}{K}$$

where $\check{\mathbf{x}}_{\mathrm{ZF}}$ *is defined in* (6.13).

Proof Because $\mathbf{n}$ and $\mathbf{w}$ are independent, (6.17) can be rewritten as

$$\hat{\mathbf{y}} = \sqrt{\frac{\rho}{N_t}}\mathbf{H}\mathbf{x} + \mathbf{n}'$$

with $\mathbf{n}' \sim \mathcal{CN}\left(\mathbf{0}_K, \left(1 + \sigma_q^2\frac{\rho}{N_t}\right)\mathbf{I}_K\right)$. Applying Proposition 1 in [75] with appropriate normalization, $\mathrm{MSE}_{\mathrm{ZF}}$ can be bounded as

$$\frac{K\left(N_t\rho^{-1} + \sigma_q^2\right)}{B^2} \leq \mathrm{MSE}_{\mathrm{ZF}} \leq \frac{K\left(N_t\rho^{-1} + \sigma_q^2\right)}{A^2} \tag{6.18}$$

where A and B are fixed constants that satisfy

$$A\mathbf{I}_{N_t} \leq \mathbf{H}^H\mathbf{H} \leq B\mathbf{I}_{N_t}.$$

The matrix inequality $A\mathbf{I}_{N_t} \leq \mathbf{H}^H\mathbf{H}$ means that the matrix $\mathbf{H}^H\mathbf{H} - A\mathbf{I}_{N_t}$ is a positive semidefinite matrix. Due to the assumption on the channel matrix, we have

$$A = B = K,$$

and the lower and upper bounds in (6.18) both become $\frac{N_t\rho^{-1}+\sigma_q^2}{K}$, which finishes the proof. ■

[5]The constant $\frac{\rho}{N_t}$ in the variance of $\mathbf{w}$ is to reflect the effect of SNR in the quantization error.
[6]$\check{\mathbf{x}}_{\mathrm{ZF}}$ is normalized to have the same norm as $\mathbf{x}$.

Remark 3: Note that the assumption $\mathbf{H}^H\mathbf{H} = K\mathbf{I}_{N_t}$ in Lemma 6.2.3 can be satisfied in the limit as the number of receive nodes grows large because

$$\mathbf{H}^H\mathbf{H} \xrightarrow{p} K\mathbf{I}_{N_t}$$

as $K \to \infty$ under our $\mathcal{CN}(0,1)$ i.i.d. channel assumption [2].

Remark 4: It is well known that the Gaussian approximation is the worst case additive noise and gives a lower bound on the mutual information [13]. Due to the inversely proportional (although implicit) relation between the mutual information and the MSE, we expect that the derivation in Lemma 6.2.3 would give an upper bound on the MSE of the ZF-type estimator. This is verified in Section 6.3 with numerically obtained σ_q^2.

We also have the following corollary when ρ becomes large.

Corollary 6.2.2 *With the same assumptions used in Lemma 6.2.3, the MSE of the ZF-type estimator is given as*

$$\mathrm{MSE}_{\mathrm{ZF}} = \frac{\sigma_q^2}{K}$$

when ρ goes to infinity.

Proof The proof of Corollary 6.2.2 is a direct consequence of taking the limit $\rho \to \infty$ on the result of Lemma 6.2.3. ■

Lemma 6.2.3 and Corollary 6.2.2 show that we can make $\mathrm{MSE}_{\mathrm{ZF}}$ arbitrarily small by increasing K regardless of the effect of noise or quantization error. However, due to the quantization process at each receive node, we have $\sigma_q^2 > 0$, and $\mathrm{MSE}_{\mathrm{ZF}}$ never goes to zero with fixed K even when $\rho \to \infty$, which gives an error rate floor in the high SNR regime. These MSE analyses are based on the ZF-type estimator and the approximation of the quantization process in (6.17); however, the numerical results in Section 6.3 show that the analyses also hold for the SER case with actual quantization process using the proposed ZF-type receiver.

6.2.4 Modified zero-forcing-type receiver

As mentioned in the previous subsection, the ZF-type receiver suffers from an error rate floor when ρ goes to infinity with fixed K. Although the error rate floor is indeed inevitable with the ZF-type receiver, we can improve the SER of the ZF-type receiver in the high SNR regime by performing post-processing for $\hat{\mathbf{x}}_{\mathrm{ZF}}$ given in (6.14).

When $\rho \to \infty$, the effect of noise disappears, and we have

$$\widetilde{\mathbf{H}}_{\mathrm{R,S}}\mathbf{x}_{\mathrm{R}} \succeq \mathbf{0}_{2K}$$

by the sign adjustment, where $\widetilde{\mathbf{H}}_{\mathrm{R,S}}$ is defined in (6.5), $\mathbf{x}_{\mathrm{R}}$ is the transmitted vector in the real domain, and $\succeq$ represents element-wise inequality. This fact also recalls the positive constraint on $t_\ell(\mathbf{x}'_{\mathrm{R}})$ in (6.7) used to upper bound the maximum of $L(\mathbf{x}'_{\mathrm{R}})$. Even in the high SNR regime, however, the $\hat{\mathbf{x}}_{\mathrm{ZF}}$ that is estimated from the ZF-type receiver may not satisfy the inequality constraints, which would cause an error rate floor. Thus, we formulate a linear program as

$$\begin{aligned} &\max_{\hat{\mathbf{x}}_{\mathrm{R}} \in \mathbb{R}^{2N_t}} \hat{\mathbf{x}}_{\mathrm{ZF}}^T \hat{\mathbf{x}}_{\mathrm{R}} \\ &\text{s.t.} \quad \widetilde{\mathbf{H}}_{\mathrm{R,S}}\hat{\mathbf{x}}_{\mathrm{R}} \succeq \mathbf{0}_{2K} \end{aligned}$$

to force the estimate $\hat{\mathbf{x}}_{\mathrm{R}}$ to satisfy the inequality constraints. The estimate $\hat{\mathbf{x}}_{\mathrm{R}}$ should be mapped to $\mathcal{S}$ as in (6.14) before decoding.

It was shown in [75] that in the context of frame expansion without any noise, the reconstruction method by linear programming can give a MSE proportional to $\frac{1}{K^2}$, which is much better than the ZF-type receiver which results in a MSE proportional to $\frac{1}{K}$. However, if ρ is not large enough, this post-processing by linear programming can cause performance degradation because the sign refinement may not be perfect, resulting in incorrect inequality constraints for the linear programming. Moreover, in this case, having more receive nodes may cause more errors due to the higher chance of having wrong inequality constraints. Note that more receive nodes corresponds to

more rows in $\widetilde{\mathbf{H}}_{\mathrm{R,S}}$ that force more inequality constraints. We numerically evaluate the modified ZF-type receiver in Section 6.3.

6.2.5 Achievable rate analysis

We can obtain the achievable rate of quantized distributed reception by evaluating the mutual information between the transmitted vector $\mathbf{x}$ and the quantized received signal $\hat{\mathbf{y}}$ given channel realization $\mathbf{H}$. If we assume $\mathbf{x}$ is uniformly distributed, i.e., $\Pr(\mathbf{x}) = \frac{1}{M^{N_t}}$, then the mutual information can be written as[7]

$$I(\mathbf{H}) = \frac{1}{M^{N_t}} \sum_{\mathbf{x}\in\mathcal{S}^{N_t}} \sum_{\hat{\mathbf{y}}\in\mathcal{Y}} \Pr\left(\hat{\mathbf{y}} \mid \mathbf{H}, \mathbf{x}\right) \log_2 \left(\frac{\Pr\left(\hat{\mathbf{y}} \mid \mathbf{H}, \mathbf{x}\right)}{\frac{1}{M^{N_t}} \sum_{\mathbf{x}'\in\mathcal{S}^{N_t}} \Pr\left(\hat{\mathbf{y}} \mid \mathbf{H}, \mathbf{x}'\right)} \right)$$

where $\mathcal{Y}$ is the set of all $(2^2)^K$ possible outcomes of the quantized received signal. Then, the average achievable rate is given as

$$R_{\mathrm{ach}} = \mathrm{E}\left[I(\mathbf{H})\right] \tag{6.19}$$

where the expectation is taken over $\mathbf{H}$.

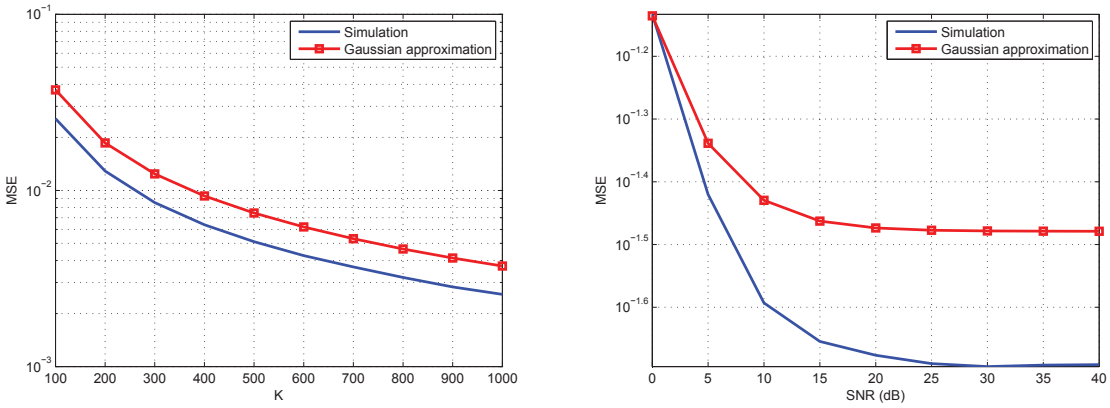
In general, it is difficult to obtain $I(\mathbf{H})$ analytically [115]. However, we are able to calculate $I(\mathbf{H})$ using the quantization structure in our problem. With the real domain notation in Section 6.2.1, we have

$$\Pr\left(\hat{\mathbf{y}} \mid \mathbf{H}, \mathbf{x}\right) = \prod_{i=1}^{2} \prod_{k=1}^{K} \Pr\left(\hat{y}_{\mathrm{R},k,i} \mid \mathbf{h}_{\mathrm{R},k,i}, \mathbf{x}_{\mathrm{R}}\right).$$

Moreover, the probability of $\hat{y}_{\mathrm{R},k,i} = 1$ can be derived as

$$\begin{aligned} \Pr\left(\hat{y}_{\mathrm{R},k,i} = 1 \mid \mathbf{h}_{\mathrm{R},k,i}, \mathbf{x}_{\mathrm{R}}\right) &= \Pr\left(\sqrt{\frac{\rho}{N_t}} \mathbf{h}_{\mathrm{R},k,i}^T \mathbf{x}_{\mathrm{R}} \geq -n_{\mathrm{R},k,i}\right) \\ &= \Phi\left(\sqrt{\frac{2\rho}{N_t}} \mathbf{h}_{\mathrm{R},k,i}^T \mathbf{x}_{\mathrm{R}}\right). \end{aligned}$$

[7] Although the mutual information also depends on ρ and N_t, we omit them for brevity.



(a) $\rho = 10$. (b) $K = 100$.

Fig. 6.2.: The MSE of the ZF-type estimator and its approximation in Lemma 6.2.3 with increasing either of K or ρ. We set $M = 8$ (8PSK) and $N_t = 4$ for both figures.

Similarly, we have

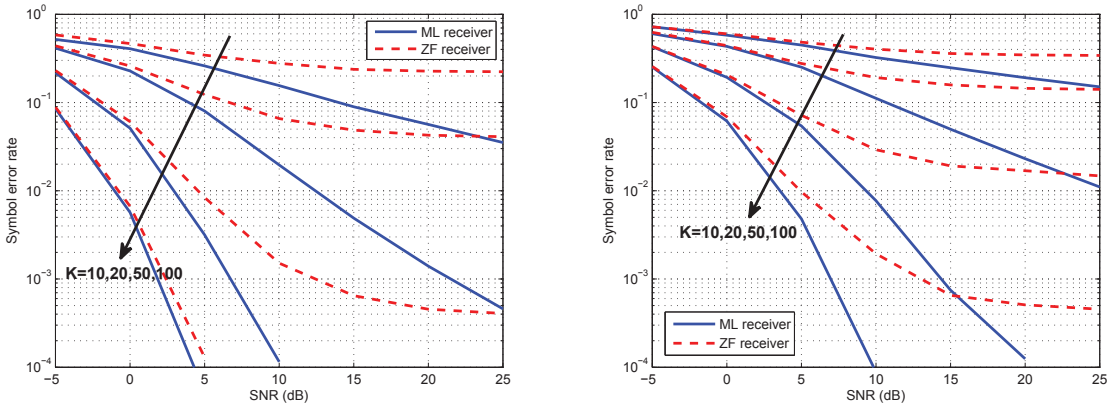
$$\Pr\left(\hat{y}_{\mathrm{R},k,i} = -1 \mid \mathbf{h}_{\mathrm{R},k,i}, \mathbf{x}_{\mathrm{R}}\right) = 1 - \Phi\left(\sqrt{\frac{2\rho}{N_t}}\mathbf{h}_{\mathrm{R},k,i}^{T}\mathbf{x}_{\mathrm{R}}\right).$$

Using these probabilities, we can calculate $I(\mathbf{H})$ analytically for given $\mathbf{H}$ and $\mathbf{x}$.

6.3 Numerical Results

In this section, we evaluate the proposed techniques with Monte Carlo simulations. We first evaluate the MSE of the ZF-type estimator and its analytical approximation derived in Lemma 6.2.3 where σ_q^2 in Lemma 6.2.3 is obtained numerically by averaging the empirical variance of the distribution[8] $(\mathbf{y} - \hat{\mathbf{y}})$ with different values of SNRs. The ML estimator is not considered in this simulation because it is computationally impractical to search over the (norm-constrained) N_t-dimensional complex space for the ML estimator. In Fig. 6.2a, we increase K with fixed $\rho = 10$ (i.e., an SNR of 10 dB[9]) while we increase ρ with fixed $K = 100$ in Fig. 6.2b. We set $N_t = 4$ and

[8] $\hat{\mathbf{y}}$ is the actual quantized received signal, not the approximation in (6.17).
[9] Recall that ρ is related to total transmit power, not per antenna transmit power, in our system setup.



(a) $M = 4$ (QPSK) and $N_t = 6$.

(b) $M = 8$ (8PSK) and $N_t = 4$.

Fig. 6.3.: SER vs. SNR in dB scale with different values of N_t and M for the constellation $\mathcal{S}$. Both figures are the case of 12 bits transmission per channel use.

$M = 8$ (8PSK) in both figures. It is clear that the MSE of the ZF-type estimator is certainly bounded with fixed K as ρ becomes larger. However, if K becomes larger, the MSE of the ZF-type estimator decreases without bound. As mentioned in Remark 4, the additive Gaussian noise approximation for the quantization error gives an upper bound for the MSE of the ZF-type estimator.

To see the diversity gain of each receiver, we consider the average SER which is defined as

$$\mathrm{SER} = \frac{1}{N_t} \sum_{n=1}^{N_t} \mathrm{E}\left[\Pr\left(\hat{x}_n \neq x_n \mid \mathbf{x} \text{ sent}, \mathbf{H}, \mathbf{n}, \rho, N_t, K, \mathcal{S}\right)\right]$$

where the expectation is taken over $\mathbf{x}$, $\mathbf{H}$, and $\mathbf{n}$. We compare the SERs of ML and ZF-type (without the modification by the linear programming) receivers regarding the transmit SNR ρ in dB scale with different values of N_t and M in Fig. 6.3. Note that both figures are for the case of 12 bits transmission per channel use because the total number of bits transmitted per channel use is given as

$$B_{\mathrm{tot}} = N_t \log_2 M.$$

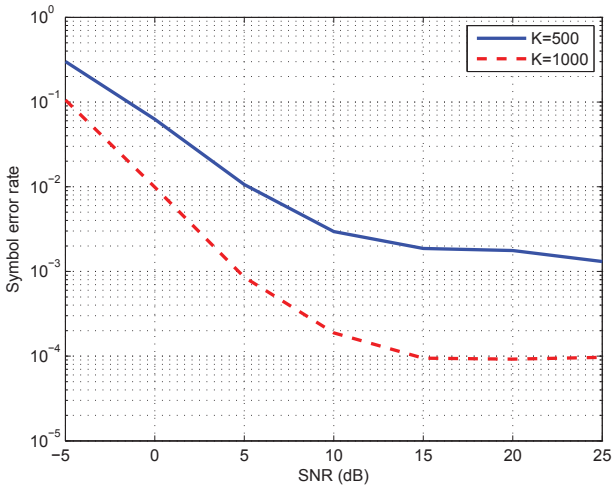


Fig. 6.4.: SER vs. SNR in dB scale using the ZF-type receiver with $M = 16$ (16QAM) for the constellation $\mathcal{S}$ and $N_t = 10$.

It is clear from the figures that as ρ or K increase, the SER of the ML receiver becomes smaller without any bound while that of the ZF-type receiver is certainly bounded in the high SNR regime. However, the SER of the ZF-type receiver can be improved by increasing K, which is the same as the MSE results. The results show that the ZF-type receiver would be a good option for quantized distributed reception with a large number of receive nodes in the IoT environment.

Comparing these figures, if the number of transmit antennas N_t at the transmitter is large, it is desirable to simultaneously transmit more symbols chosen from a smaller sized constellation to get better SER results when the number of transmitted bits per channel use, B_{tot}, is fixed for both the ML and ZF-type receivers. This result is suitable to massive MIMO systems where the transmitter is equipped with a large number of antennas.

We also plot the SERs of the ZF-type receiver[10] with a 16QAM constellation for $\mathcal{S}$ and $N_t = 10$ in Fig. 6.4. The figure shows that the proposed receivers also work for a non-PSK constellation even with not-so-large N_t. Thus, the norm constraint $\|\mathbf{x}\|^2 = N_t$ is not critical for the ZF-type receiver.

[10]We do not consider the ML receiver due to its excessive complexity.

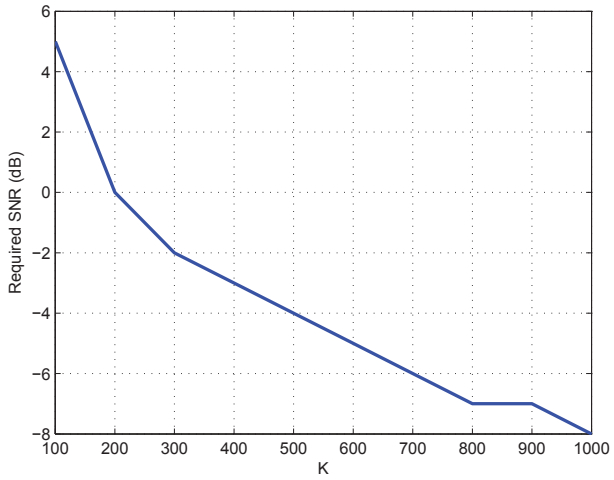


Fig. 6.5.: Required SNR vs. K for the ZF-type receiver to achieve the target SER of 0.01 with $N_t = 4$ and $M = 8$ (8PSK) for the constellation $\mathcal{S}$.

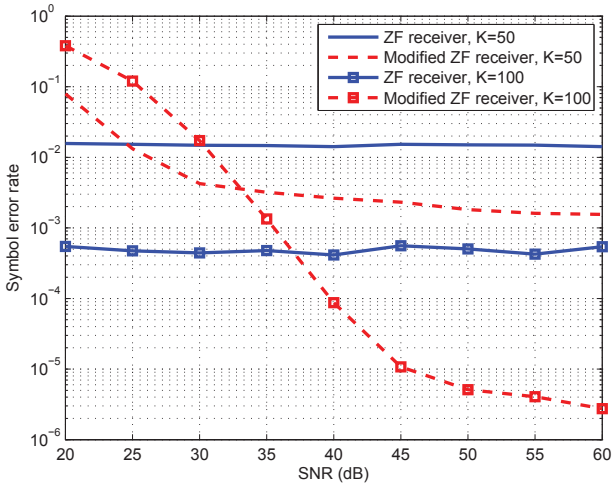


Fig. 6.6.: SER vs. SNR in dB scale for the ZF-type and modified ZF-type receivers with $N_t = 4$ and $M = 8$ (8PSK) for the constellation $\mathcal{S}$.

To numerically evaluate the array gain, we plot the required SNR (in dB) for the ZF-type receiver to achieve the target SER of 0.01 against K in Fig. 6.5. As the number of receive nodes increases, the required SNR to achieve the target SER decreases. Therefore, if we can exploit a large number of receive nodes, the transmitter may be able to rely on cost efficient power amplifiers with small transmit power.

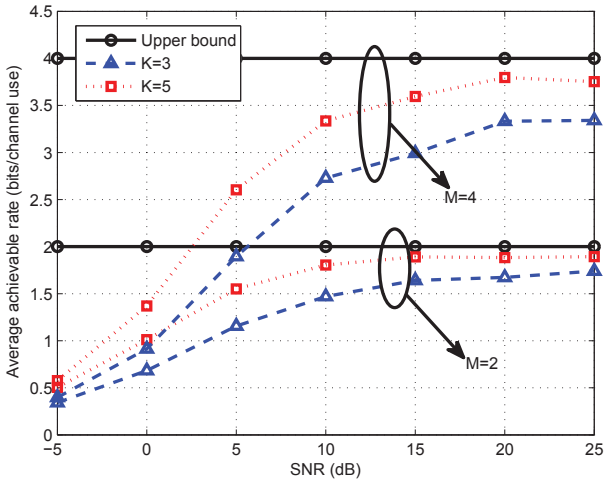


Fig. 6.7.: Average achievable rates of quantized distributed reception vs. SNR with $N_t = 2$ and different values of K and M.

In Fig. 6.6, we plot the SERs of the ZF-type receiver and a ZF-type receiver modified to use linear programming explained in Section 6.2.4. We only consider the high SNR regime because the modified ZF-type receiver is aimed to increase the performance of the ZF-type receiver when the SNR is high. The figure clearly shows that the modified ZF-type receiver performs much better than the ZF-type receiver when the effect of noise becomes negligible; however, it performs worse than the ZF-type receiver when the SNR is not sufficiently high. Moreover, having more receive nodes deteriorates the performance of the modified ZF-type receiver in this case, which is explained in Section 6.2.4.

In Fig. 6.7, we plot the average achievable rate defined in (6.19) to evaluate the benefit of spatial multiplexing in distributed reception. Due to the computational complexity, we only consider $N_t = 2$ with $K = 3$ and 5 receive nodes.[11] Moreover, we limit the minimum value of $\Pr\left(\hat{y}_{\mathrm{R},k,i} \mid \mathbf{h}_{\mathrm{R},k,i}, \mathbf{x}_{\mathrm{R}}\right)$ to 0.0001 to avoid numerical errors on $I(\mathbf{H})$. The figure shows that the achievable rates of quantized distributed reception for both $M = 2$ (BPSK) and 4 (QPSK) cases become close to its maximum value as ρ increases even with small numbers of K. It is expected that we can achieve the

[11] Because the size of $\mathcal{Y}$ is $(2^2)^K$, it quickly becomes computationally impractical as K increases.

maximum achievable rate with a *not-so-large* number of receive nodes, e.g., $K = 10$, with moderate SNR values.

To make the scenario more practical, the fusion center may decode the transmitted symbols with limited or no global channel knowledge, which is an interesting future research topic.

7. CONCLUSIONS

In this dissertation, we proposed efficient downlink training and CSI quantization techniques to design practical FDD massive MIMO systems. We also studied distributed reception scenarios and proposed superior quantization and decoding methods for the cases of single and multiple transmit antennas.

In Chapter 2, we proposed open and closed-loop training frameworks using successive channel prediction/estimation at the user for FDD massive MIMO systems. By exploiting prior channel information such as the long-term channel statistics and previous received training signals at the user, channel estimation performance can be significantly improved with only small length of training signals in each fading block compared to open-loop/single-shot training. Moreover, with a small amount of feedback, which indicates the best training signal to be sent for the next fading block, from the user to the base station, the downlink training overhead can be further reduced even when the transmitter lacks any kind of side information, e.g., statistics of the channel.

In Chapter 3, we proposed an efficient channel quantization method, dubbed NTCQ, for massive MIMO systems employing limited feedback beamforming. While the quantization criterion (maximization of beamforming gain or minimization of chordal distance) is associated with the Grassmann manifold, the key to the proposed NTCQ approach is to leverage efficient encoding (via the Viterbi algorithm) and codebook design (via TCQ) in Euclidean space. Efficient encoding relies on the mapping of quantization on the Grassmann manifold to noncoherent sequence detection and the near-optimal implementation of noncoherent detection using a bank of coherent detectors (i.e., Euclidean space quantizers). Standard rate-distortion theory and asymptotic results for RVQ tell us that good Euclidean codebooks should work

well in Grassmannian space. Our numerical results show that the NTCQ provides better performance than uncoded schemes such as those considered in [91].

The advantages of NTCQ include flexibility and scalability in the number of channel coefficients: additional coefficients can be accommodated simply by increasing the blocklength, and the encoding complexity is linear in the number of transmit antennas. It can also be easily modified to take advantage of channel conditions such as temporal and spatial correlations. Our numerical results show that these advanced schemes can improve the performance significantly or reduce feedback overhead considerably depending on the system requirement.

In Chapter 4, we proposed TEC and TE-SPA which are efficient channel quantization techniques for FDD massive MIMO systems. The proposed TEC exploits a trellis quantizer combined with VQ codebooks to achieve a practical feedback overhead and complexity. TEC can easily satisfy backward compatibility by exploting standardized codebooks such as LTE or LTE-Advanced codebooks. We proposed a codeword-to-branch mapping and codebook design criteria to maximize the performance of TEC. TEC also can support multiple receive antennas making a unified CSI quantization framework possible. It has been shown using simulations that TEC can maintain a constant performance gap with RVQ which is known to be asymptotically optimal.

For TE-SPA, we incorporated a trellis structure to quantize temporally correlated channels in a successive manner. TE-SPA also can be adapted to spatially correlated channels without any difficulty. TEC and TE-SPA can be thought of as an evolution of the LTE-Advanced dual codebooks for long-term/wideband and short-term/subband CSI quantization. The numerical results confirmed that the proposed TE-SPA can reduce quantization loss even with reduced feedback overhead.

In Chapter 5, we proposed a unified framework for coded receive diversity distributed reception. We consider distributed reception for the case when a transmitter broadcasts a signal to multiple geographically separated receive nodes through fading channels, and each receive node processes and forwards the received signal to a fusion

center. The fusion center then tries to detect the transmitted signal exploiting the forwarded data from all the receive nodes and channel state information if available. The proposed coded receive diversity technique is based on the strong connection between the distributed reception problem and coding theory. By leveraging this connection, we are able to adopt appropriate linear block codes, e.g., simplex and first-order Reed-Muller codes that achieve the Griesmer bound with equality, to design processing rules at the receive nodes and maximize the diversity gain. We also developed novel shortened concatenated repetition-simplex (SCRS) codes to support an arbitrary number of the receive nodes. We analytically proved that the SCRS codes are optimal with respect to the Griesmer bound in many practical scenarios. We evaluated the proposed coded receive diversity technique by numerical studies. Because of its simple and flexible structure, the proposed technique can be applied to various scenarios including cellular systems, wireless sensor networks, and radar systems.

In the last chapter, we studied a distributed reception scenario where the transmitter is equipped with multiple transmit antennas and broadcasts multiple independent data symbols by spatial multiplexing to a set of geographically separated receive nodes through fading channels. Each receive node then processes its received signal and forwards it to the fusion center, and the fusion center tries to decode the transmitted data symbols by exploiting the forwarded information and global channel knowledge. We implemented an optimal ML receiver and a low-complexity ZF-type receiver for this scenario. The SER of the ML receiver can be made arbitrarily small by increasing SNR and the number of receive nodes. The ZF-type receiver suffers from an error rate floor as the SNR increases. This floor can be lowered by increasing the number of receive nodes.

REFERENCES

REFERENCES

[1] T. L. Marzetta, "How much training is required for multiuser MIMO?" *Proceedings of IEEE Asilomar Conference on Signals, Systems, and Computers*, Oct. 2006.

[2] ——, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[3] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[4] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[5] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[6] D. J. Love, R. W. Heath Jr., V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.

[7] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2640–2651, Aug. 2011.

[8] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 264–273, Feb. 2013.

[9] R. Muller, M. Vehkapera, and L. Cottatellucci, "Blind pilot decontamination," *International ITG Workshop on Smart Antennas*, Mar. 2013.

[10] J. Guey and L. D. Larsson, "Modeling and evaluation of MIMO systems exploiting channel reciprocity in TDD mode," *Proceedings of IEEE Vehicular Technology Conference*, Sep. 2004.

[11] A. Pitarokoilis, S. Mohammed, and E. G. Larsson, "Uplink performance of time-reversal MRC in massive MIMO systems subject to phase noise," *IEEE Transactions on Wireless Communications*, to appear. [Online]. Available: http://arxiv.org/abs/1308.2747

[12] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 7112–7139, Nov. 2014.

[13] B. Hassibi and B. Hochwald, "How much training is needed in multipleantenna wireless links?" *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[14] C. K. Au-Yeung and D. J. Love, "On the performance of random vector quantization limited feedback beamforming in a MISO system," *IEEE Transactions on Wireless Communications*, vol. 6, no. 2, pp. 458–462, Feb. 2007.

[15] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Elsevier Computer Networks*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[16] "Coordinated multi-point operation for lte physical layer aspects (release 11)," *3GPP TR 36.819*, 2011.

[17] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Communications Magazine*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[18] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana, "Coordinated multipoint transmission and reception in lte-advanced: Deployment scenarios and operational challenges," *IEEE Communications Magazine*, vol. 50, no. 2, pp. 148–155, Feb. 2012.

[19] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, and M. Tanno, "Coordinated multipoint transmission/reception techniques for LTE-Advanced," *IEEE Wireless Communications*, vol. 17, no. 3, pp. 26–34, Jun. 2010.

[20] S. Brueck, L. Zhao, J. Giese, and M. A. Amin, "Centralized scheduling for joint transmission coordinated multi-point in LTE-Advanced," *International ITG Workshop on Smart Antennas*, Feb. 2010.

[21] L. Venturino, N. Prasad, and X. Wang, "Coordinated linear beamforming in downlink multi-cell wireless networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 4, pp. 1451–1461, Apr. 2010.

[22] W. Yu, T. Kwon, and C. Shin, "Multicell coordination via joint scheduling, beamforming and power spectrum adaptation," *Proceedings of IEEE INFOCOM*, Apr. 2011.

[23] W. Roh and A. Paulraj, "Outage performance of the distributed antenna systems in a composite fading channel," *Proceedings of IEEE Vehicular Technology Conference*, Sep. 2002.

[24] W. Choi and J. G. Andrews, "Downlink performance and capacity of distributed antenna systems in a multicell environment," *IEEE Transactions on Wireless Communications*, vol. 6, no. 1, pp. 69–73, Jan. 2007.

[25] J. Zhang and J. G. Andrews, "Distributed antenna systems with randomness," *twireless*, vol. 7, no. 9, pp. 3636–3646, Sep. 2008.

[26] A. M. Haimovich, R. S. Blum, and L. J. Cimini, "MIMO radar with widely separated antennas," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 116–129, Jan. 2008.

[27] Q. He, R. S. Blum, H. Godrich, and A. M. Haimovich, "Target velocity estimation and antenna placement for MIMO radar with widely separated antennas," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 1, pp. 79–100, Feb. 2010.

[28] M. Dianat, M. R. Taban, J. Dianat, and V. Sedighi, "Target localization using least squares estimation for MIMO radars with widely separated antennas," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 49, no. 4, pp. 2730–2741, Oct. 2013.

[29] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Computer Networks*, vol. 38, no. 4, pp. 393–422, Mar. 2002.

[30] ——, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, Aug. 2002.

[31] V. Mhatre and C. Rosenberg, "Design guidelines for wireless sensor networks: Communication, clustering and aggregation," *Ad Hoc Networks*, vol. 2, no. 1, pp. 45–63, Jan. 2004.

[32] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Computer Networks*, vol. 52, no. 12, pp. 2292–2330, Aug. 2008.

[33] A. Hammons, J. Hampton, N. Merheb, and M. Cruz, "Cooperative MIMO field measurements for military UHF band in low-rise urban environment," *in 5th IEEE Sensor Array and Multichannel Signal Processing Workshop*, Jul. 2008.

[34] S. H. Lee, S. Lee, H. Song, and H. S. Lee, "Wireless sensor network design for tactical military applications: Remote large-scale environments," *Proceedings of IEEE Military Communications Conference*, Oct. 2009.

[35] J. H. Kotecha and A. M. Sayeed, "Transmit signal design for optimal estimation of correlated MIMO channels," *IEEE Transaction on Signal Processing*, vol. 52, pp. 546–557, Feb. 2004.

[36] F. Rubio, D. Guo, M. L. Honig, and X. Mestre, "On optimal training and beamforming in uncorrelated MIMO systems with feedback," *Proceedings of Conference on Information Sciences and Systems*, Mar. 2008.

[37] W. Santipach and M. L. Honig, "Optimization of training and feedback overhead for beamforming over block fading channels," *IEEE Transactions on Information Theory*, vol. 56, no. 12, pp. 6103–6115, Dec. 2010.

[38] E. Björnson and B. Ottersten, "A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician distrubance," *IEEE Transaction on Signal Processing*, vol. 58, no. 3, pp. 1807–1820, Mar. 2010.

[39] C. Komninakis, C. Fragouli, A. H. Sayed, and R. D. Wesel, "Multi-input multi-output fading channel tracking and equalization using Kalman estimation," *IEEE Transaction on Signal Processing*, vol. 50, pp. 1065–1075, May 2002.

[40] K. Huber and S. Haykin, "Improved Bayesian MIMO channel tracking for wireless communications: Incorporating a dynamical model," *IEEE Transactions on Wireless Communications*, vol. 5, pp. 2468–2476, Sep. 2006.

[41] S. Noh, M. D. Zoltowski, Y. Sung, and D. J. Love, "Pilot beam pattern design for channel estimation in massive MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 787–801, Oct. 2014.

[42] B. Chalise, L. Haering, and A. Czylwik, "Robust uplink to downlink spatial covariance matrix transformation for downlink beamforming," *Proceedings of IEEE International Conference on Communications*, Jun. 2004.

[43] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2562–2579, Oct. 2003.

[44] D. J. Love, R. W. Heath Jr., and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2735–2747, Oct. 2003.

[45] S. Zhou, Z. Wang, and G. B. Giannakis, "Quantifying the power-loss when transmit-beamforming relies on finite rate feedback," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1948–1957, Jul. 2005.

[46] W. Santipach and M. L. Honig, "Capacity of multiple-antenna fading channel with quantized precoding matrix," *IEEE Transactions on Information Theory*, vol. 55, no. 3, pp. 1218–1234, Mar. 2009.

[47] D. J. Love and R. W. Heath Jr., "Grassmannian beamforming on correlated MIMO channels," *Proceedings of IEEE Global Telecommunications Conference*, Dec. 2004.

[48] ——, "Limited feedback diversity techniques for correlated channels," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 2, pp. 718–722, Mar. 2006.

[49] P. Xia and G. B. Giannakis, "Design and analysis of transmit-beamforming based on limited-rate feedback," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1853–1863, Mar. 2006.

[50] B. Banister and J. Zeidler, "Feedback assisted transmission subspace tracking for MIMO systems," *IEEE Journal on Selected Areas in Communications*, vol. 21, pp. 452–463, May 2003.

[51] J. Yang and D. Williams, "Transmission subspace tracking for MIMO systems with low-rate feedback," *IEEE Transactions on Communications*, vol. 55, no. 8, pp. 1629–1639, Aug. 2007.

[52] R. W. Heath Jr., T. Wu, and A. C. K. Soong, "Progressive refinement of beamforming vectors for high-resolution limited feedback," *EURASIP Journal Advances in Signal Processing*, vol. 2009, no. 6, Feb. 2009.

[53] K. Huang, R. W. Heath Jr., and J. G. Andrews, "Limited feedback beamforming over temporally-correlated channels," *IEEE Transaction on Signal Processing*, vol. 57, no. 5, pp. 1959–1975, May 2009.

[54] D. Sacristan and A. Pascual-Iserte, "Differential feedback of MIMO channel gram matrices based on geodesic curves," *IEEE Transactions on Wireless Communications*, vol. 9, no. 12, pp. 3714–3727, Dec. 2010.

[55] T. Kim, D. J. Love, and B. Clerckx, "MIMO system with limited rate differential feedback in slow varying channel," *IEEE Transactions on Communications*, vol. 59, no. 4, pp. 1175–1180, Apr. 2010.

[56] J. Choi, B. Clerckx, N. Lee, and G. Kim, "A new design of polar-cap differential codebook for temporally/spatially correlated MISO channels," *IEEE Transactions on Wireless Communications*, vol. 11, no. 2, pp. 703–711, Feb. 2012.

[57] J. Choi, B. Clerckx, and D. J. Love, "Differential codebook for general rotated dual-polarized MISO channels," *Proceedings of IEEE Global Telecommunications Conference*, Dec. 2012.

[58] W. Santipach and M. L. Honig, "Asymptotic performance of MIMO wireless channels with limited feedback," *Proceedings of IEEE Military Communications Conference*, Oct. 2003.

[59] S. Lin and D. J. Costello, Jr., *Error Control Coding*. New Jersey: Prentice Hall, 2004.

[60] J. C. Roh and B. D. Rao, "Transmit beamforming in multiple-antenna systems with finite rate feedback: A VQ-based approach," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1101–11 112, Mar. 2006.

[61] R. S. Blum, "Distributed detection for diversity reception of fading signals in noise," *IEEE Transactions on Information Theory*, vol. 45, no. 1, pp. 158–164, Jan. 1999.

[62] D. R. Brown III, U. Madhow, M. Ni, M. Rebholz, and P. Bidigare, "Distributed reception with hard decision exchanges," *IEEE Transactions on Wireless Communications*, vol. 13, no. 6, pp. 3406–3418, Jun. 2014.

[63] M. Gastpar, G. Kramer, and P. Gupta, "The multiple-relay channel: Coding and antenna-clustering capacity," *Proceedings of IEEE International Symposium on Information Theory*, Jun./Jul. 2002.

[64] A. Høst-Madsen and J. Zhang, "Capacity bounds and power allocation for the wireless relay channel," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2020–2040, Jun. 2005.

[65] M. R. Souryal and H. You, "Quantize-and-forward relaying with M-ary phase shift keying," *Proceedings of IEEE Wireless Communications and Networking Conference*, Apr. 2008.

[66] S. Simoens, O. M. noz Medina, J. Vidal, and A. D. Coso, "Compress-and-forward cooperative MIMO relaying with full channel state information," *IEEE Transaction on Signal Processing*, vol. 58, no. 2, pp. 781–791, Feb. 2010.

[67] I. Avram, N. Aerts, H. Bruneel, and M. Moeneclaey, "Quantize and forward cooperative communication: Channel parameter estimation," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1167–1179, Mar. 2012.

[68] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.

[69] T.-Y. Wang, Y. S. Han, P. K. Varshney, and P.-N. Chen, "Distributed fault-tolerant classification in wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 724–734, Apr. 2005.

[70] T.-Y. Wang, Y. S. Han, B. Chen, and P. K. Varshney, "A combined decision fusion and channel coding scheme for distributed fault-tolerant classification in wireless sensor networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 7, pp. 1695–1705, Jul. 2006.

[71] C. Yao, P.-N. Chen, T.-Y. Wang, Y. S. Han, and P. K. Varshney, "Performance analysis and code design for minimum Hamming distance fusion in wireless sensor networks," *IEEE Transactions on Information Theory*, vol. 53, no. 5, pp. 1716–1734, May 2007.

[72] J. Fang and H. Li, "Hyperplane-based vector quantization for distributed estimation in wireless sensor networks," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5682–5699, Dec. 2009.

[73] ——, "Optimal/near-optimal dimensionality reduction for distributed estimation in homogeneous and certain inhomogeneous scenarios," *IEEE Transaction on Signal Processing*, vol. 58, no. 8, pp. 4339–4353, Aug. 2010.

[74] R. R. Tenney and N. R. Sandell Jr., "Detection with distributed sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-17, no. 4, pp. 501–510, Jul. 1981.

[75] V. K. Goyal, M. Vetterli, and N. T. Thao, "Quantized overcomplete expansions in $\mathbb{R}^n$: Analysis, synthesis, and algorithms," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 16–30, Jan. 1998.

[76] V. K. Goyal and J. Kovačević, "Quantized frame expansions with erasures," *Applied and Computational Harmonic Analysis*, vol. 10, no. 3, pp. 203–233, May 2001.

[77] D. J. Love, J. Choi, and P. Bidigare, "A closed-loop training approach for massive MIMO beamforming systems," *Proceedings of Conference on Information Sciences and Systems*, Mar. 2012.

[78] A. J. Duly, T. Kim, D. J. Love, and J. V. Krogmeier, "Closed-loop beam alignment for massive MIMO channel estimation," *IEEE Communications Letters*, vol. 18, pp. 1439–1442, Apr. 2014.

[79] "Evolved universal terrestrial radio access (E-UTRA); physical channels and modulation," *3GPP TS 36.211 v11.0.0*, Sep. 2012.

[80] L. Dai, Z. Wang, and Z. Yang, "Spectrally efficient time-frequency training OFDM for mobile large-scale MIMO systems," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 251–263, Feb. 2013.

[81] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, 1st ed. New Jersey: Prentice Hall, 2000.

[82] J. N. Pierce and S. Stein, "Multiple diversity with nonindependent fading," *Proceedings of the IRE*, vol. 48, no. 1, pp. 89–104, Jan. 1960.

[83] M. S. Paolella, "Computing moments of ratios of quadratic forms in normal variables," *Computational Statistics and Data Analysis*, vol. 42, no. 3, pp. 313–331, 2003.

[84] J. G. Proakis, *Digital Communication*, 4th ed. New York: McGraw-Hill, 2000.

[85] M. W. Marcellin and T. R. Fischer, "Trellis coded quantization of memoryless and Gauss-Markov sources," *IEEE Transactions on Communications*, vol. 38, no. 1, pp. 82–93, Jan. 1990.

[86] Y. Nam, B. L. Ng, K. Sayana, Y. Li, J. Zhang, Y. Kim, and J. Lee, "Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 172–179, Jun. 2013.

[87] C. K. Au-Yeung, D. J. Love, and S. Sanayei, "Trellis coded line packing: Large dimensional beamforming vector quantization and feedback transmission," *IEEE Transactions on Wireless Communications*, vol. 10, no. 6, pp. 1844–1853, Jun. 2011.

[88] C. K. Au-Yeung, A. Jalali, and D. J. Love, "Insights into feedback and feedback signaling for beamformer design," *UCSD Information Theory and Applications Workshop*, Feb. 2009.

[89] C. K. Au-Yeung and S. Sanayei, "Enhanced trellis based vector quantization for coordinated beamforming," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2010.

[90] M. Xu, D. Guo, and M. K. Honig, "MIMO precoding with limited rate feedback: Simple quantizers work well," *Proceedings of IEEE Global Telecommunications Conference*, Dec. 2009.

[91] D. J. Ryan, I. V. L. Clarkson, I. B. Collings, D. Guo, and M. L. Honig, "QAM and PSK codebooks for limited feedback MIMO beamforming," *IEEE Transactions on Communications*, vol. 57, no. 4, pp. 1184–1196, Apr. 2009.

[92] D. J. Ryan, I. B. Collings, and I. V. L. Clarkson, "GLRT-optimal noncoherent lattice decoding," *IEEE Transaction on Signal Processing*, vol. 55, no. 7, pp. 3773–3786, Jul. 2007.

[93] W. Sweldens, "Fast block noncoherent decoding," *IEEE Communications Letters*, vol. 5, no. 4, pp. 132–134, Apr. 2001.

[94] D. Warrier and U. Madhow, "Spectrally efficient noncoherent communication," *IEEE Transactions on Information Theory*, vol. 48, no. 3, pp. 652–668, Mar. 2002.

[95] N. Jindal, "MIMO broadcast channels with finite rate feedback," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5045–5059, Nov. 2006.

[96] P. Ding, D. J. Love, and M. D. Zoltowski, "Multiple antenna broadcast channels with shape feedback and limited feedback," *IEEE Transaction on Signal Processing*, vol. 55, no. 7, pp. 3417–3428, Jul. 2007.

[97] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[98] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Transactions on Information Theory*, vol. 28, no. 1, pp. 55–67, Jan. 1982.

[99] R. H. Etkin and D. N. C. Tse, "Degree of freedom in some underspread MIMO fading channel," *IEEE Transactions on Information Theory*, vol. 52, pp. 1576–1608, Apr. 2006.

[100] D. J. Love and R. W. Heath Jr., "Equal gain transmission in multiple-input multiple-output wireless systems," *IEEE Transactions on Communications*, vol. 51, no. 7, pp. 1102–1110, Jul. 2003.

[101] J. Choi, Z. Chance, D. J. Love, and U. Madhow, "Noncoherent trellis coded quantization: A practical limited feedback technique for massive MIMO systems," *IEEE Transactions on Communications*, vol. 61, no. 12, pp. 5016–5029, Dec. 2013.

[102] "Spatial channel model for multiple input multiple output (mimo) simulations," *3GPP TR 25.996 V6.1.0*, Sep. 2003. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/25996.htm

[103] J. Choi and D. J. Love, "Bounds on eigenvalues of a spatial correlation matrix," *IEEE Communications Letters*, vol. 18, no. 8, pp. 1391–1394, Aug. 2014.

[104] B. Clerckx, G. Kim, J. Choi, and S. Kim, "Allocation of feedback bits among users in broadcast MIMO channels," *Proceedings of IEEE Global Telecommunications Conference*, Dec. 2008.

[105] J. Lee and W. Choi, "Optimal feedback rate sharing strategy in zero-forcing MIMO broadcast channels," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 3000–3011, Jun. 2013.

[106] A. H. Nguyen, Y. Huang, and B. D. Rao, "Optimized quantized feedback in a multiuser system employing CDF based scheduling," *Proceedings of IEEE Vehicular Technology Conference*, Sep. 2013.

[107] J. H. Griesmer, "A bound for error-correcting codes," *IBM Journal of Research and Development*, vol. 4, no. 5, pp. 532–542, Nov. 1960.

[108] R. M. Roth, *Introduction to Coding Theory*. New York: Cambridge University Press, 2006.

[109] S. Thoen, L. Van der Perre, B. Gyselinckx, and M. Engels, "Performance analysis of combined transmit-SC/receive-MRC," *IEEE Transactions on Communications*, vol. 49, no. 1, pp. 5–8, Jan. 2001.

[110] D. J. Love and R. W. Heath Jr., "Necessary and sufficient conditions for full diversity order in correlated Rayleigh fading beamforming and combining systems," *IEEE Transactions on Wireless Communications*, vol. 4, no. 1, pp. 20–23, Jan. 2005.

[111] J. Choi, D. J. Love, and D. R. Brown III, "Channel estimation techniques for quantized distributed reception in MIMO systems," *Proceedings of IEEE Asilomar Conference on Signals, Systems, and Computers*, Nov. 2014.

[112] E. Wolfstetter, *Topics in Microeconomics: Industrial Organization, Auctions, and Incentives*. Cambridge University Press, 1999.

[113] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. Tata McGraw-Hill, 2002.

[114] J. J. Benedetto, A. M. Powell, and Özgür Yılmaz, "Sigma-delta ($\Sigma\Delta$) quantization and finite frames," *IEEE Transaction on Signal Processing*, vol. 52, no. 5, pp. 1990–2005, May 2006.

[115] B. M. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Transactions on Communications*, vol. 51, no. 3, pp. 389–399, Mar. 2008.

[116] C. G. Baker, "Riemannian manifold trust-region methods with applications to eigenproblems," Ph.D. dissertation, Florida State University, 2008.

[117] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*. New York: Academic Press, 1979.

APPENDIX

A. APPENDIX

A.1 Proof of Lemma 2.2.1

Because $\mathbf{R}$ is fixed, minimizing the MSE problem can be converted to

$$\begin{aligned}\underset{\mathbf{X}\in\mathcal{X}}{\operatorname{argmin}}\,\mathrm{MSE}\left(\mathbf{X}\right) &= \underset{\mathbf{X}\in\mathcal{X}}{\operatorname{argmax}}\,\mathrm{tr}\left(\mathbf{R}\mathbf{X}\left(\mathbf{I}_T+\mathbf{X}^H\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^H\mathbf{R}\right)\\ &\overset{(a)}{=} \underset{\mathbf{X}\in\mathcal{X}}{\operatorname{argmax}}\,\mathrm{tr}\left(\left(\mathbf{I}_T+\mathbf{X}^H\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^H\mathbf{R}^2\mathbf{X}\right),\end{aligned} \tag{A.1}$$

where (a) is from the fact that $\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{BCA})$. Using the eigen-decomposition of $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$, we can rewrite (A.1) as

$$\begin{aligned}\underset{\mathbf{X}\in\mathcal{X}}{\operatorname{argmin}}\,\mathrm{MSE}\left(\mathbf{X}\right) &= \underset{\mathbf{X}\in\mathcal{X}}{\operatorname{argmax}}\,\mathrm{tr}\left(\left(\mathbf{I}_T+\mathbf{X}^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\mathbf{X}\right)^{-1}\mathbf{X}^H\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^H\mathbf{X}\right)\\ &\overset{(a)}{=} \underset{\widetilde{\mathbf{X}}\in\mathcal{X}}{\operatorname{argmax}}\,\mathrm{tr}\left(\left(\mathbf{I}_T+\widetilde{\mathbf{X}}^H\mathbf{\Lambda}\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^H\mathbf{\Lambda}^2\widetilde{\mathbf{X}}\right)\\ &\overset{(b)}{=} \underset{\widetilde{\mathbf{X}}\in\mathcal{X}}{\operatorname{argmax}}\,\mathrm{tr}\left(\left(\widetilde{\mathbf{X}}^H\left(\frac{1}{\rho}\mathbf{I}_{N_t}+\mathbf{\Lambda}\right)\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^H\mathbf{\Lambda}^2\widetilde{\mathbf{X}}\right)\end{aligned}$$

where (a) comes from the change of the variable $\widetilde{\mathbf{X}} = \mathbf{U}^H\mathbf{X}$, and (b) is from $\widetilde{\mathbf{X}}^H\widetilde{\mathbf{X}} = \rho\mathbf{I}_T$. Because $\rho^{-1}\mathbf{I}_{N_t}+\mathbf{\Lambda}$ and $\mathbf{\Lambda}^2$ are all real diagonal matrices with strictly positive entries in decreasing order, from the property of the block generalized Rayleigh quotient [116], the optimal solution for single-shot training is given as

$$\widetilde{\mathbf{X}}_{\mathrm{ss,opt}} = \sqrt{\rho}\mathbf{I}_{N_t[1:T]}.$$

Thus, $\mathbf{X}_{\mathrm{ss,opt}}$ becomes

$$\mathbf{X}_{\mathrm{ss,opt}} = \mathbf{U}\widetilde{\mathbf{X}}_{\mathrm{ss,opt}} = \sqrt{\rho}\mathbf{U}\mathbf{I}_{N_t[1:T]} = \sqrt{\rho}\mathbf{U}_{[1:T]},$$

which finishes the proof.

A.2 Proof of Lemma 2.2.2

The basic concept of majorization theory which is used to prove Lemma 2.2.2 is from [35, 117].

Let the real-valued function $f : \mathbb{R}^T \to \mathbb{R}$ as

$$f(\mathbf{x}) = \sum_{t=1}^{T} \frac{\rho x_t^2}{\rho x_t + 1}$$

with a vector $\mathbf{x} = [x_1, x_2, \cdots, x_T]^T$ and a constant $\rho > 0$. Note that $f(\mathbf{x})$ is the same as the second term in (2.8), which should be maximized to minimize the MSE. It is easy to show that $f(\mathbf{x})$ is Schur-convex because $f(\mathbf{x})$ is symmetric and $\frac{\rho x_t^2}{\rho x_t + 1}$ is convex. By majorization theory and the property of Schur-convexity, we have

$$\mathbf{x} \succ \mathbf{y} \Rightarrow f(\mathbf{x}) \geq f(\mathbf{y})$$

with arbitrary two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^T$. Because we assume $\lambda(\mathbf{R}_H) \succ \lambda(\mathbf{R}_L)$, we have $\mathrm{MSE}(\mathbf{X}_H) \leq \mathrm{MSE}(\mathbf{X}_L)$.

A.3 Proof of Lemma 2.2.3

First, we decompose $\mathbf{h} = \widehat{\mathbf{h}} + \mathbf{r}$ where $\widehat{\mathbf{h}}$ and $\mathbf{r}$ are independent because of the orthogonality of the MMSE estimator [81]. Note that the covariance of $\mathbf{r}$ is given as

$$\mathbf{R}_\mathbf{r} = \mathbf{R} - \mathbf{R}\mathbf{X}\left(\mathbf{I}_T + \mathbf{X}^H\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^H\mathbf{R}.$$

$$\begin{aligned}
\mathbf{R}_{1|0} &= \eta^2 \mathbf{R}_{0|0} + (1-\eta^2)\mathbf{R} \\
&= \mathbf{R} - \eta^2 \mathbf{R}\mathbf{X}_{0,\text{opt}} \left(\mathbf{I}_T + \mathbf{X}_{0,\text{opt}}^H \mathbf{R}\mathbf{X}_{0,\text{opt}}\right)^{-1} \mathbf{X}_{0,\text{opt}}^H \mathbf{R} \\
&\overset{(a)}{=} \mathbf{U}_0 \Lambda_0 \mathbf{U}_0^H - \eta^2 \left(\mathbf{U}_0 \mathbf{\Lambda}_{0[1:T]} \left(\mathbf{I}_T + \rho \,\text{diag}\left([\lambda_{0,1}, \cdots, \lambda_{0,T}]\right)\right)^{-1} \mathbf{\Lambda}_{0[1:T]}^H \mathbf{U}_0^H\right) \\
&= \mathbf{U}_0 \,\text{diag}\left(\left[\lambda_{0,1} - \eta^2 \frac{\rho\lambda_{0,1}^2}{\rho\lambda_{0,1}+1}, \cdots, \lambda_{0,T} - \eta^2 \frac{\rho\lambda_{0,T}^2}{\rho\lambda_{0,T}+1}, \lambda_{0,T+1}, \cdots, \lambda_{0,N_t}\right]\right) \mathbf{U}_0^H \\
&= \mathbf{U}_1 \mathbf{\Lambda}_1 \mathbf{U}_1^H
\end{aligned} \tag{A.2}$$

We let $\mathbf{R_r} = \mathbf{U_r}\mathbf{\Lambda_r}\mathbf{U_r}$ where $\mathbf{U_r}$ and $\mathbf{\Lambda_r} = \text{diag}\left([\lambda_{\mathbf{r},1}, \cdots, \lambda_{\mathbf{r},N_t}]\right)$ are the eigenvector matrix and the eigenvalue matrices (in decreasing order) of $\mathbf{R_r}$, respectively. Now, we expand $\Gamma_{\text{ss,opt}}$ as

$$\begin{aligned}
\Gamma_{\text{ss,opt}} &\overset{(a)}{=} E\left[E\left[|\mathbf{h}\mathbf{w}|^2 \,|\mathbf{h}\right]\right] \\
&= E\left[\mathbf{w}^H \left(\widehat{\mathbf{h}}\widehat{\mathbf{h}}^H + \mathbf{R_r}\right)\mathbf{w}\right] \\
&\overset{(b)}{=} \text{tr}\left(\mathbf{R}_{\widehat{\mathbf{h}}}\right) + E\left[\frac{\widehat{\mathbf{h}}^H \mathbf{R_r} \widehat{\mathbf{h}}}{\left\|\widehat{\mathbf{h}}\right\|^2}\right] \\
&\overset{(c)}{\leq} \text{tr}\left(\mathbf{R}_{\widehat{\mathbf{h}}}\right) + \lambda_{\mathbf{r},1} \\
&\leq \text{tr}\left(\mathbf{R}_{\widehat{\mathbf{h}}}\right) + \lambda_1 \\
&\overset{(d)}{=} \sum_{t=1}^{T} \frac{\rho\lambda_t^2}{\rho\lambda_t + 1} + \lambda_1
\end{aligned}$$

where the inner expectation is over $\mathbf{r}$ (or noise $\mathbf{n}$) and the outer expectation is over $\mathbf{h}$ in (a), (b) comes from $\mathbf{w} = \frac{\widehat{\mathbf{h}}}{\left\|\widehat{\mathbf{h}}\right\|}$, (c) is because $\|\mathbf{R_r}\mathbf{x}\|^2 \leq \lambda_{\mathbf{r},1}$ for any unit vector $\mathbf{x}$, and (d) can be easily derived similar to (2.8) with $\mathbf{X}_{\text{ss,opt}} = \sqrt{\rho}\mathbf{U}_{[1:T]}$.

A.4 Proof of Lemma 2.3.1

At $i = 1$, $\mathbf{R}_{1|0}$ is given as in (A.2) where (a) comes from $\mathbf{X}_{0,\text{opt}} = \sqrt{\rho}\mathbf{U}_{0[1:T]}$. Note that $\mathbf{U}_1$ and $\mathbf{U}_0$ have the same columns with a different order based on the

$$\begin{aligned}
\mathrm{MSE}\left(\mathbf{X}_{1,\mathrm{opt}}\right) &= \frac{1}{N_t}\mathrm{tr}\left(\mathbf{R}_{1|0} - \mathbf{R}_{1|0}\mathbf{X}_{1,\mathrm{opt}}\left(\mathbf{I}_T + \mathbf{X}_{1,\mathrm{opt}}^H\mathbf{R}_{1|0}\mathbf{X}_{1,\mathrm{opt}}\right)^{-1}\mathbf{X}_{1,\mathrm{opt}}^H\mathbf{R}_{1|0}\right) \\
&= \frac{1}{N_t}\mathrm{tr}\left(\mathbf{R}_{1|0} - \left(\mathbf{I}_T + \mathbf{X}_{1,\mathrm{opt}}^H\mathbf{R}_{1|0}\mathbf{X}_{1,\mathrm{opt}}\right)^{-1}\mathbf{X}_{1,\mathrm{opt}}^H\mathbf{R}_{1|0}^2\mathbf{X}_{1,\mathrm{opt}}\right) \\
&= \frac{1}{N_t}\left(\sum_{t=1}^{N_t}\lambda_{0,t} - \eta^2\sum_{t=1}^{T}\frac{\rho\lambda_{0,t}^2}{\rho\lambda_{0,t}+1} - \sum_{t=1}^{T}\frac{\rho\lambda_{1,t}^2}{\rho\lambda_{1,t}+1}\right) \\
&= 1 - \frac{1}{N_t}\left(\eta^2\sum_{t=1}^{T}\frac{\rho\lambda_{0,t}^2}{\rho\lambda_{0,t}+1} + \sum_{t=1}^{T}\frac{\rho\lambda_{1,t}^2}{\rho\lambda_{1,t}+1}\right) \qquad \text{(A.3)}
\end{aligned}$$

eigenvalues of $\mathbf{\Lambda}_1$. Because $\mathbf{X}_{1,\mathrm{opt}} = \mathbf{U}_{1[1:T]}$, the MSE of the block $i = 1$ is given as in (A.3). We can generalize (A.3) for $i > 1$ with recursive derivation, which finishes the proof.

A.5 Lemma to Prove Lemma 6.2.1

Lemma A.5.1 *For arbitrary s and c that satisfy $s > c > 0$, we have*

$$\left(\Phi(s)\right)^2 > \Phi(s+c)\Phi(s-c).$$

Proof With $s > c > 0$, we have the inequality

$$\Phi(s) - \Phi(s-c) > \Phi(s+c) - \Phi(s).$$

Then, we have

$$2\Phi(s) > \Phi(s+c) + \Phi(s-c)$$

which is equivalent to

$$\begin{aligned} 4\left(\Phi(s)\right)^2 &> \left(\Phi(s+c)+\Phi(s-c)\right)^2 \\ &= \left(\Phi(s+c)\right)^2 + \left(\Phi(s-c)\right)^2 + 2\Phi(s+c)\Phi(s-c) \\ &\overset{(a)}{\geq} 4\Phi(s+c)\Phi(s-c) \end{aligned}$$

where (a) is because

$$\left(\Phi(s+c)-\Phi(s-c)\right)^2 \geq 0,$$

which finishes the proof. ■

A.6 Proof of First-Order Stochastic Dominance in Lemma 6.2.2

We drop unnecessary subscripts to simplify notation. Recall that

$$\begin{aligned} y &= \sqrt{\frac{\rho}{N_t}}\mathbf{h}^T\mathbf{x} + n, \\ \widetilde{\mathbf{h}} &= \hat{y}\mathbf{h} \end{aligned}$$

where $\hat{y} = \mathrm{sgn}(y)$. Using the fact that $\mathbf{h}$ is rotationally invariant, we assume the transmitted vector is given as[1] $\mathbf{x} = \begin{bmatrix}\sqrt{N_t} & 0 & \cdots & 0\end{bmatrix}^T$. Then, we have

$$y = \sqrt{\rho}h_1 + n.$$

Because $y \sim \mathcal{N}(0, \frac{\rho+1}{2})$ and $n \sim \mathcal{N}(0, \frac{1}{2})$, the distribution of $\sqrt{\rho}h_1$ conditioned on y is $\mathcal{N}(\mu, \gamma^2)$ where

$$\mu = \frac{\rho}{\rho+1}y, \quad \gamma^2 = \frac{\rho}{2(\rho+1)}.$$

[1]In this proof, we do not have to restrict the elements of $\mathbf{x}$ from an M-ary constellation $\mathcal{S}$ because we consider the ML estimator not receiver.

Let $c = \frac{\rho}{\rho+1}$. Then, we can write $\sqrt{\rho}h_1 = cy + w$ where $w \sim \mathcal{N}(0, \gamma^2)$. Moreover, we have

$$\sqrt{\frac{\rho}{N_t}}\widetilde{\mathbf{h}}^T\mathbf{x} = \sqrt{\rho}\hat{y}h_1 = c|y| + \hat{y}w \stackrel{d}{=} |cy| + w$$

conditioned on y where the third equality comes from the independence of $\hat{y}$ and w. Note that $\stackrel{d}{=}$ denotes stochastic equivalence.

Now we want to compute the distribution of $\sqrt{\frac{\rho}{N_t}}\widetilde{\mathbf{h}}^T\mathbf{u}$ for a fixed $\mathbf{u}$ given y. Note that

$$\mathbf{h}^T\mathbf{u} = \sum_{i=1}^{2N_t} h_i u_i \stackrel{d}{=} u_1 h_1 + z\sqrt{N_t - u_1^2}$$

where $z \sim \mathcal{N}(0, \frac{1}{2})$. Then, we have

$$\begin{aligned}
\sqrt{\frac{\rho}{N_t}}\widetilde{\mathbf{h}}^T\mathbf{u} &\stackrel{d}{=} \frac{u_1}{\sqrt{N_t}}(c|y| + w) + \hat{y}z\sqrt{\rho\left(1 - \frac{u_1^2}{N_t}\right)} \\
&\stackrel{d}{=} \frac{u_1}{\sqrt{N_t}}(c|y| + w) + z\sqrt{\rho\left(1 - \frac{u_1^2}{N_t}\right)} \\
&= u(c|y| + w) + z\sqrt{\rho\left(1 - u^2\right)}
\end{aligned}$$

where the second equality is due to the independence of $\hat{y}$ and z and the third equality comes from the variable substitution $u = \frac{u_1}{\sqrt{N_t}}$. Note that $-1 \leq u < 1$. If $u = 1$, then $\mathbf{u}$ becomes $\mathbf{x}$, which violates our assumption.

We now break up $uw + z\sqrt{\rho\left(1 - u^2\right)}$ into two independent zero-mean Gaussian random variables v_1 and v_2 where

$$v_1 \sim \mathcal{N}\left(0, (1 - u^2)\frac{\rho^2}{2(\rho + 1)}\right), \quad v_2 \sim \mathcal{N}(0, \gamma^2).$$

Finally, for a given y, we have

$$\begin{aligned}
\sqrt{\frac{\rho}{N_t}}\widetilde{\mathbf{h}}^T\mathbf{u} &\stackrel{d}{=} u(c|y|+w)+z\sqrt{\rho\left(1-u^2\right)} \\
&= uc|y|+v_1+v_2 \\
&\stackrel{d}{<} |uc|y|+v_1|+v_2 \qquad \text{(A.4)} \\
&= |ucy+\hat{y}v_1|+v_2 \\
&\stackrel{d}{=} |ucy+v_1|+v_2 \qquad \text{(A.5)} \\
&\stackrel{d}{=} |cy|+w \qquad \text{(A.6)} \\
&\stackrel{d}{=} \sqrt{\frac{\rho}{N_t}}\widetilde{\mathbf{h}}^T\mathbf{x}.
\end{aligned}$$

To show the strict stochastic dominance in (A.4), recall that $uc|y|$ is a fixed number given y, and v_1 is a Gaussian random variable. Thus, the complementary cumulative distribution function of $|uc|y|+v_1|$ should be strictly greater than that of $uc|y|+v_1$. The stochastic equivalence in (A.5) is because $\hat{y}$ and v_1 are independent and (A.6) is due to the facts that

$$ucy+v_1 \sim \mathcal{N}\left(0, \frac{\rho^2}{2(\rho+1)}\right), \quad cy \sim \mathcal{N}\left(0, \frac{\rho^2}{2(\rho+1)}\right)$$

and $v_2 \stackrel{d}{=} w$. Thus, (6.16) holds, and we have the claim.

VITA

VITA

Junil Choi received the B.S. (with honors) and M.S. degrees from Seoul National University, Seoul, Korea, in 2005 and 2007, respectively. He is currently working toward the Ph.D. degree with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. From 2007 to 2011, he was a member of the technical staff at Samsung Electronics, Korea, where he contributed advanced codebook and feedback framework designs to 3GPP LTE-Advanced and IEEE 802.16m standards. His research interests are in the design and analysis of massive MIMO and distributed communication systems.

He was a co-recipient of the 2013 IEEE Globecom Signal Processing for Communications Symposium Best Paper Award and the 2008 Global Samsung Technical Conference Best Paper Award. He was a recipient of the Michael and Katherine Birck Fellowship from Purdue University in 2011; the Korean Government Scholarship Program for Study Overseas in 2011-2013; the Purdue ECE Graduate Student Association Outstanding Graduate Student Award in 2013; and the Purdue College of Engineering Outstanding Student Research Award in 2014. He was recognized as an Exemplary Reviewer of the IEEE WIRELESS COMMUNICATIONS LETTERS in 2013.